

---

# Causal Component Analysis

---

**Wendong Liang** <sup>1,2</sup>    **Armin Kekić** <sup>1</sup>    **Julius von Kügelgen** <sup>1,3</sup>    **Simon Buchholz** <sup>1</sup>

**Michel Besserve** <sup>1</sup>    **Luigi Gresele**<sup>\*1</sup>    **Bernhard Schölkopf**<sup>\*1</sup>

<sup>1</sup> Max Planck Institute for Intelligent Systems, Tübingen, Germany

<sup>2</sup> ENS Paris-Saclay, Gif-sur-Yvette, France    <sup>3</sup> University of Cambridge, United Kingdom

{wendong.liang, armin.kekic, jvk, simon.buchholz}@tue.mpg.de

{besserve, luigi.gresele, bs}@tue.mpg.de

## Abstract

Independent Component Analysis (ICA) aims to recover independent latent variables from observed mixtures thereof. Causal Representation Learning (CRL) aims instead to infer causally related (thus often statistically *dependent*) latent variables, together with the unknown graph encoding their causal relationships. We introduce an intermediate problem termed *Causal Component Analysis* (*CauCA*). CauCA can be viewed as a generalization of ICA, modelling the causal dependence among the latent components, and as a special case of CRL. In contrast to CRL, it presupposes knowledge of the causal graph, focusing solely on learning the unmixing function and the causal mechanisms. Any impossibility results regarding the recovery of the ground truth in CauCA also apply for CRL, while possibility results may serve as a stepping stone for extensions to CRL. We characterize CauCA identifiability from multiple datasets generated through different types of interventions on the latent causal variables. As a corollary, this interventional perspective also leads to new identifiability results for nonlinear ICA—a special case of CauCA with an empty graph—requiring strictly fewer datasets than previous results. We introduce a likelihood-based approach using normalizing flows to estimate both the unmixing function and the causal mechanisms, and demonstrate its effectiveness through extensive synthetic experiments in the CauCA and ICA setting.

## 1 Introduction

Independent Component Analysis (ICA) (Comon, 1994) is a principled approach to representation learning, which aims to recover independent latent variables, or sources, from observed mixtures thereof. Whether this is possible depends on the *identifiability* of the model (Wasserman, 2004; Lehmann and Casella, 2006): this characterizes assumptions under which a learned representation provably recovers (or *disentangles*) the latent variables, up to some well-specified ambiguities (Hyvärinen et al., 2023; Xi and Bloem-Reddy, 2023). A key result shows that, when nonlinear mixtures of the latent components are observed, the model is non-identifiable based on independent and identically distributed (i.i.d.) samples from the generative process (Hyvärinen and Pajunen, 1999; Darmois, 1951). Consequently, a learned model may explain the data as well as the ground truth, even if the corresponding representation is strongly entangled, rendering the recovery of the original latent variables fundamentally impossible.

Identifiability can be recovered under deviations from the i.i.d. assumption, e.g., in the form of temporal autocorrelation (Hyvärinen and Morioka, 2017; Hälvä and Hyvärinen, 2020) or spatial dependence (Hälvä et al., 2021) among the latent components; *auxiliary variables* which render the sources

---

<sup>\*</sup> Shared last author.

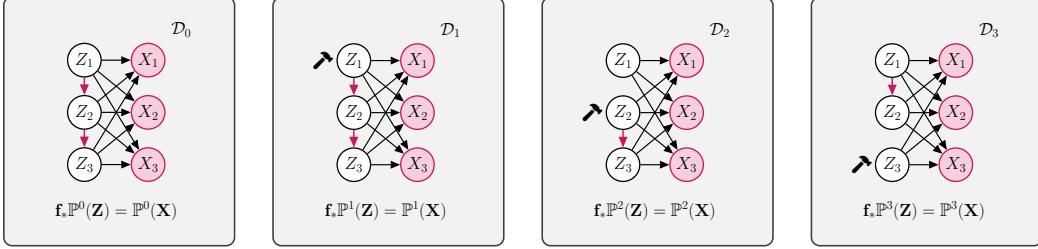


Figure 1: **Causal Component Analysis (CauCA).** We posit that observed variables  $\mathbf{X}$  are generated through a nonlinear mapping  $f$ , applied to unobserved latent variables  $\mathbf{Z}$  which are causally related. The causal structure  $G$  of the latent variables is assumed to be known, while the causal mechanisms  $\mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)})$  and the nonlinear mixing function are unknown. (Known or observed quantities are highlighted in red.) CauCA assumes access to multiple datasets  $\mathcal{D}_k$  that result from stochastic interventions on the latent variables. Its objective is to estimate both the unmixing function and the causal mechanisms.

*conditionally independent* (Hyvärinen and Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020a,b); or additional, noisy views (Gresele et al., 2019). An alternative path is to restrict the class of mixing functions (Zhang and Hyvärinen, 2009; Buchholz et al., 2022; Gresele et al., 2021).

Despite appealing identifiability guarantees for ICA, the independence assumption can be limiting, since interesting factors of variation in real-world data are often statistically, or causally, dependent (Träuble et al., 2021). This motivates Causal Representation Learning (CRL) (Schölkopf et al., 2021), which aims instead to infer causally related latent variables, together with a causal graph encoding their causal relationships. This is challenging if both the graph and the unmixing are unknown.

Identifiability results in causal representation learning therefore require strong assumptions such as counterfactual data (Locatello et al., 2020; von Kügelen et al., 2021; Brehmer et al., 2022), temporal structure (Lippe et al., 2022; Lachapelle et al., 2022), graph sparsity (Lachapelle et al., 2022; Lachapelle and Lacoste-Julien, 2022), a parametric family of latent distributions (Lachapelle et al., 2022; Squires et al., 2023), or strong restrictions on the mixing function class (Squires et al., 2023; Varici et al., 2023). Since knowing either the graph or the unmixing might help better recover the other, it has been argued that this gives rise to a *chicken-and-egg* problem in CRL (Brehmer et al., 2022).

We introduce an intermediate problem between ICA and CRL which we call *Causal Component Analysis (CauCA)*, see Fig. 1 for an overview. CauCA can be viewed as a generalization of ICA that models causal connections (and thus statistical dependence) among the latent components through a causal Bayesian network (Pearl, 2009). It can also be viewed as a special case of CRL that presupposes knowledge of the causal graph, and focuses on learning the unmixing function and causal mechanisms.

Since CauCA is solving the CRL problem with partial ground truth information, it is strictly easier than CRL. This implies that impossibility results for CauCA also apply for CRL. *Possibility* results for CauCA, on the other hand, while not automatically generalizing to CRL, can nevertheless serve as stepping stones, highlighting potential avenues for achieving corresponding results in CRL. Note also that there are only finitely many possible directed acyclic graphs for a fixed number of nodes, but the space of spurious solutions in representation learning (e.g., in nonlinear ICA) is typically infinite. By solving CauCA problems, we can therefore gain insights into the minimal assumptions required for addressing CRL problems. CauCA is applicable to scenarios in which domain knowledge can be used to specify a causal graph for the latent components. For instance, in computer vision applications, the image generation process can often be modelled based on a fixed graph (Sauer and Geiger, 2021; Tangemann et al., 2021).

**Structure and Contributions.** We start by recapitulating preliminaries on causal Bayesian networks and interventions in § 2. Next, we introduce Causal Component Analysis (CauCA) in § 3. Our primary focus lies in characterizing the identifiability of CauCA from multiple datasets generated through various types of interventions on the latent causal variables (§ 4). Importantly, all our results are applicable to the *nonlinear* and *nonparametric* case. The interventional perspective we take exploits the *modularity* of the causal relationships (i.e., the possibility to change one of them without affecting the others)—a concept that was not previously leveraged in works on nonlinear ICA. This leads extensions of existing results that require strictly fewer datasets to achieve the same level of identifiability. We introduce and investigate an estimation procedure for CauCA in § 5 and conclude with a discussion in the context of other related work in § 6. We highlight the following *main contributions*:

- We provide sufficient and necessary results in CauCA identifiability (Thm. 4.2, Prop. 4.3).
- We prove additional results for the special case with an empty graph, which corresponds to a novel ICA model with interventions on the latent variables (Prop. 4.4, Prop. 4.5, Thm. 4.6, Prop. 4.7).
- We show in synthetic experiments in both the CauCA and ICA settings that our normalizing flow-based estimation procedure effectively recovers the latent causal components (§ 5).

## 2 Preliminaries

**Notation.** We use  $\mathbb{P}$  to denote a probability distribution, with density function  $p$ . Uppercase letters  $X, Y, Z$  denote unidimensional and bold uppercase  $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$  denote multidimensional random variables. Lowercase letters  $x, y, z$  denote scalars in  $\mathbb{R}$  and  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  denote vectors in  $\mathbb{R}^d$ . We use  $\llbracket i, j \rrbracket$  to denote the integers from  $i$  to  $j$ , and  $[d]$  denotes the natural numbers from 1 to  $d$ . We use common graphical notation, see App. A for details. The *ancestors* of  $i$  in a graph are the nodes  $j$  in  $G$  such that there is a directed path from  $j$  to  $i$ , and they are denoted by  $\text{anc}(i)$ . The *closure* of the parents (resp. ancestors) of  $i$  is defined as  $\overline{\text{pa}}(i) := \text{pa}(i) \cup \{i\}$  (resp.  $\overline{\text{anc}}(i) := \text{anc}(i) \cup \{i\}$ ).

A key definition connecting directed acyclic graphs (DAGs) and probabilistic models is the following.

**Definition 2.1** (Distribution Markov relative to a DAG (Pearl, 2009)). *A joint probability distribution  $\mathbb{P}$  is Markov relative to a DAG  $G$  if it admits the factorization  $\mathbb{P}(Z_1, \dots, Z_d) = \prod_{i=1}^d \mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)})$ .*

Defn. 2.1 is a key assumption in directed graphical models, where a distribution being Markovian relative to a graph implies that the graph encodes specific independences within the distribution, which can be exploited for efficient computation or data storage (Peters et al., 2017, §6.5).

**Causal Bayesian networks and interventions.** Causal systems induce multiple distributions corresponding to different interventions. Causal Bayesian networks (CBNs; Pearl, 2009) can be used to represent how these interventional distributions are related. In a CBN with associated graph  $G$ , arrows signify causal links among variables, and the conditional probabilities  $\mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)})$  in the corresponding Markov factorization are called *causal mechanisms*.<sup>2</sup>

Interventions are modelled in CBNs by replacing a subset  $\tau_k \subseteq V(G)$  of the causal mechanisms by new, intervened mechanisms  $\{\tilde{\mathbb{P}}_j(Z_j | \mathbf{Z}_{\text{pa}^k(j)})\}_{j \in \tau_k}$  while all other causal mechanisms are left unchanged. Here,  $\text{pa}^k(j)$  denotes the parents of  $Z_j$  in the post-intervention graph in the interventional regime  $k$  and  $\tau_k$  the intervention targets. We will omit the superscript  $k$  when the parent set is unchanged and assume that interventions do not add new parents,  $\text{pa}^k(j) \subseteq \text{pa}(j)$ . Unless  $\tilde{\mathbb{P}}_j$  is a point mass, we call the intervention *stochastic* or soft. Further, we say that  $\tilde{\mathbb{P}}_j$  is a *perfect* intervention if the dependence of the  $j$ -th variable from its parents is removed ( $\text{pa}^k(j) = \emptyset$ ), corresponding to deleting all arrows pointing to  $j$ , sometimes also referred to as *graph surgery* (Spirtes et al., 2000).<sup>3</sup> An *imperfect* intervention is one for which  $\text{pa}^k(j) \neq \emptyset$ . We summarise this in the following definition.

**Definition 2.2** (CBN). *A causal Bayesian network (CBN) consists of a graph  $G$ , a collection of causal mechanisms  $\{\mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)})\}_{i \in [d]}$ , and a collection of interventions  $\{\{\tilde{\mathbb{P}}_j^k(Z_j | \mathbf{Z}_{\text{pa}^k(j)})\}_{j \in \tau_k}\}_{k \in [K]}$  across  $K$  interventional regimes. The joint probability for interventional regime  $k$  is given by:*

$$\mathbb{P}^k(\mathbf{Z}) := \begin{cases} \prod_{i=1}^d \mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)}) & k = 0 \\ \prod_{j \in \tau_k} \tilde{\mathbb{P}}_j^k(Z_j | \mathbf{Z}_{\text{pa}^k(j)}) \prod_{i \notin \tau_k} \mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)}) & \forall k \in [K] \end{cases} \quad (1)$$

where  $\mathbb{P}^0$  is the unintervened, or observational, distribution, and  $\mathbb{P}^k$  are interventional distributions.

*Remark 2.3.* The joint probabilities  $\mathbb{P}^k$  in (1) are uniquely factorized into causal mechanisms according to  $G$ . We therefore use the equivalent notation  $(G, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$ , where  $\mathbb{P}^k$  is defined as in (1).

<sup>2</sup>The term can also be used in structural causal models to denote deterministic functions of endogenous and exogenous variables in *assignments*, see (Peters et al., 2017, Def. 3.1). A central idea in causality (Pearl, 2009; Peters et al., 2017) is that causal mechanisms are *modular* or *independent*, i.e., it is possible to modify some without affecting the others: after an intervention, typically only a subset of the causal mechanisms change.

<sup>3</sup>A special case of perfect interventions are *hard* interventions, where  $\tilde{\mathbb{P}}_j$  corresponds to a Dirac distribution:  $\mathbb{P}(\mathbf{Z} | \text{do}(Z_j = z_j)) = \delta_{Z_j=z_j} \prod_{i \neq j} \mathbb{P}_i(Z_i | \mathbf{Z}_{\text{pa}(i)})$ .

### 3 Problem Setting

The main object of our study is a latent variable model termed *latent causal Bayesian network (CBN)*.

**Definition 3.1** (Latent CBN). A latent CBN is a tuple  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$ , where  $\mathbf{f} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  is a diffeomorphism (i.e. invertible with both  $\mathbf{f}$  and  $\mathbf{f}^{-1}$  differentiable).

**Data-generating process for Causal Component Analysis (CauCA).** In CauCA, we assume that we are given multiple datasets  $\{\mathcal{D}_k\}_{k \in \llbracket 0, K \rrbracket}$  generated by a latent CBN  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$ :

$$\mathcal{D}_k := \left( \tau_k, \left\{ \mathbf{x}^{(n,k)} \right\}_{n=1}^{N_k} \right), \quad \text{with} \quad \mathbf{x}^{(n,k)} = \mathbf{f}(\mathbf{z}^{(n,k)}) \quad \text{and} \quad \mathbf{z}^{(n,k)} \stackrel{\text{i.i.d.}}{\sim} \mathbb{P}^k, \quad (2)$$

where  $N_k$  denotes the sample size for interventional regime  $k$ , see Fig. 1 for an illustration. The graph  $G$  is assumed to be known. Further, we assume that the intervention targets  $\tau_k$  are observed, see § 6 for further discussion. Both the mixing function  $\mathbf{f}$  and the latent distributions  $\mathbb{P}^k$  in (2) are unknown.

The problem we aim to address is the following: given only the graph  $G$  and the datasets  $\mathcal{D}_k$  in (2), can we learn to *invert* the mixing function  $\mathbf{f}$  and thus recover the latent variables  $\mathbf{z}$ ? Whether this is possible, and up to what ambiguities, depends on the identifiability of CauCA.

**Definition 3.2** (Identifiability of CauCA). A model class for CauCA is a tuple  $(G, \mathcal{F}, \mathcal{P}_G)$ , where  $\mathcal{F}$  is a class of functions and  $\mathcal{P}_G$  is a class of joint distributions Markov relative to  $G$ . A latent CBN  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  is said to be in  $(G, \mathcal{F}, \mathcal{P}_G)$  if  $\mathbf{f} \in \mathcal{F}$  and  $\mathbb{P}^k \in \mathcal{P}_G$  for all  $k \in \llbracket 0, K \rrbracket$ . We say  $(G, \mathcal{F}, \mathcal{P}_G)$  has known intervention targets if all its elements share the same  $G$  and  $(\tau_k)_{k \in \llbracket 0, K \rrbracket}$ . We say that CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $\mathcal{S}$  (a set of functions called “indeterminacy set”) if for any two latent CBNs  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  and  $(G', \mathbf{f}', (\mathbb{Q}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$ , the equality of pushforward  $\mathbf{f}_* \mathbb{P}^k = \mathbf{f}'_* \mathbb{Q}^k \forall k \in \llbracket 0, K \rrbracket$  implies that  $\exists \mathbf{h} \in \mathcal{S}$  s.t.  $\mathbf{h} = \mathbf{f}'^{-1} \circ \mathbf{f}$  on the support of  $\mathbb{P}$ .

We justify the definition of known intervention targets and generalize them to a more flexible scenario in App. E. Defn. 3.2 is inspired by the identifiability definition of ICA in (Buchholz et al., 2022, Def. 1). Intuitively, it states that, if two models in  $(G, \mathcal{F}, \mathcal{P}_G)$  give rise to the same distribution, then they are equal up to ambiguities specified by  $\mathcal{S}$ . Consequently, when attempting to invert  $\mathbf{f}$  based on the data in (2), the inversion can only be achieved up to those ambiguities.

In the following, we choose  $\mathcal{F}$  to be the class of all  $\mathcal{C}^1$ -diffeomorphisms  $\mathbb{R}^d \rightarrow \mathbb{R}^d$ , denoted  $\mathcal{C}^1(\mathbb{R}^d)$ , and suppose the distributions in  $\mathcal{P}_G$  are absolutely continuous with full support in  $\mathbb{R}^d$ , with the density  $p^k$  differentiable.

A first question is what ambiguities are unavoidable by construction in CauCA, similar to scaling and permutation in ICA (Hyvärinen et al., 2023, § 3.1). The following Lemma characterizes this.

**Lemma 3.3.** For any  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  in  $(G, \mathcal{F}, \mathcal{P}_G)$ , and for any  $\mathbf{h} \in \mathcal{S}_{\text{scaling}}$  with

$$\mathcal{S}_{\text{scaling}} := \{ \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{h}(\mathbf{z}) = (h_1(z_1), \dots, h_d(z_d)), h_i \text{ is a diffeomorphism in } \mathbb{R} \} \quad (3)$$

there exists a  $(G, \mathbf{f} \circ \mathbf{h}, (\mathbb{Q}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  in  $(G, \mathcal{F}, \mathcal{P}_G)$  s.t.  $\mathbf{f}_* \mathbb{P}^k = (\mathbf{f} \circ \mathbf{h})_* \mathbb{Q}^k$  for all  $k \in \llbracket 0, K \rrbracket$ .

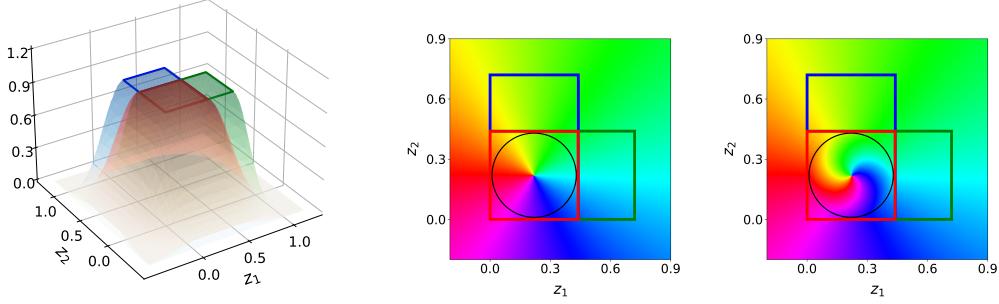
Lemma 3.3 states that, as in nonlinear ICA, the ambiguity up to element-wise nonlinear scaling is also unresolvable in CauCA. However, unlike in nonlinear ICA, there is no permutation ambiguity in CauCA: this is a consequence of the assumption of known intervention targets. The next question is under which conditions we can achieve identifiability up to (3), and when the ambiguity set is larger.

### 4 Theory

In this section, we investigate the identifiability of CauCA. We first study the general case (§ 4.1), and then consider the special case of ICA in which the graph is empty (§ 4.2).

#### 4.1 Identifiability of CauCA

**Single-node interventions.** We characterize the identifiability of CauCA based on *single-node* interventions. For datasets  $\mathcal{D}_k$  defined as in (2), every  $k > 0$  corresponds to interventions on a single variable: i.e.,  $\forall k > 0, |\tau_k| = 1$ . This is the setting depicted in Fig. 1, where each interventional dataset is generated by intervening on a single latent variable. The following assumption will play a key role in our proofs.



**Figure 2: Violation of the Interventional Discrepancy Assumption.** The shown distributions constitute a counterexample to identifiability that violates Asm. 4.1 and thus allows for spurious solutions, see App. C for technical details. (*Left*) Visualisation of the joint distributions of two independent latent components  $z_1$  and  $z_2$  after no intervention (red), and interventions on  $z_1$  (green) and  $z_2$  (blue). As can be seen, each distribution reaches the same plateau on some rectangular interval of the domain, coinciding within the red square. (*Center/Right*) Within the red square where all distributions agree, it is possible to apply a measure preserving automorphism which leaves all distributions unchanged, but non-trivially mixes the latents. The right plot shows a distance-dependent rotation around the centre of the black circle, whereas the middle plot show a reference identity transformation.

**Assumption 4.1** (Interventional discrepancy). *For all  $\mathbf{z} \in \mathbb{R}^d$ , for all pairs of stochastic interventions  $\tilde{p}_i$  and corresponding causal mechanism for the  $i$ -th variable,  $p_i$ , we have  $\forall i \in [d]$*

$$\frac{\partial(\ln p_i)}{\partial z_i}(z_i | \mathbf{z}_{pa(i)}) \neq \frac{\partial(\ln \tilde{p}_i)}{\partial z_i}(z_i | \mathbf{z}_{pa^i(i)}) \quad \text{almost everywhere (a.e.)} \quad (4)$$

Asm. 4.1 can be applied to imperfect and perfect interventions alike (in the latter case the conditioning on the RHS disappears). Intuitively, Asm. 4.1 requires that the stochastic intervention is sufficiently different from the causal mechanism, formally expressed as the requirement that the partial derivative over  $z_i$  of the ratio between  $p_i$  and  $\tilde{p}_i$  is nonzero a.e. One case in which Asm. 4.1 is violated is when  $\partial p_i / \partial z_i$  and  $\partial \tilde{p}_i / \partial z_i$  are both zero on the same open subset of their support. In Fig. 2(*Left*), we provide an example of such a violation (see App. C for its construction), and apply a measure-preserving automorphism within the area where the two distributions agree (see Fig. 2(*Right*)).

We can now state our main result for CauCA with single-node interventions.

**Theorem 4.2.** *For CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$ ,*

- (i) *Suppose for each node in  $[d]$ , there is one (perfect or imperfect) stochastic intervention such that Asm. 4.1 is verified. Then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to*

$$\mathcal{S}_{\bar{G}} = \left\{ \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{h}(\mathbf{z}) = (h_i(\mathbf{z}_{\bar{anc}(i)}))_{i \in [d]}, \mathbf{h} \text{ is } \mathcal{C}^1\text{-diffeomorphism} \right\} \quad (5)$$

- (ii) *Suppose for each node  $i$  in  $[d]$ , there is one perfect stochastic intervention such that Asm. 4.1 is verified, then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $\mathcal{S}_{\text{scaling}}$ .*

Thm. 4.2 (i) states that for single-node stochastic interventions, perfect or imperfect, we can achieve identifiability up to an indeterminacy set where each reconstructed variable can at most be a mixture of ground truth variables corresponding to nodes in the closure of the ancestors of  $i$ . While this ambiguity set is larger than the one in eq. (3), it is still a nontrivial reduction in ambiguity with respect to the spurious solutions which could be generated without Asm. 4.1. A related result in (Squires et al., 2023, Thm. 1) shows that for *linear mixing, linear latent SCM and unknown graph*,  $d$  interventions are sufficient and necessary for recovering  $\bar{G}$  (the transitive closure of the ground truth graph  $G$ ) and the latent variables up to elementwise reparametrizations. Thm. 4.2 (i) instead proves that  $d$  interventions are sufficient for identifiability up to mixing of variables corresponding to the coordinates in  $\bar{G}$  for arbitrary  $\mathcal{C}^1$ -diffeomorphisms  $\mathcal{F}$ , non-parametric  $\mathcal{P}_G$  and known graph.

Thm. 4.2 (ii) shows that if we further constrain the set of interventions to *perfect* single-node, stochastic interventions only, then we can achieve a much stronger identifiability—i.e., identifiability up to scaling, which as discussed in § 3 is the best one we can hope to achieve in our problem setting without further assumptions. In short, the un-intervened distribution together with one single-node, stochastic perfect intervention per node is sufficient to give us the strongest achievable identifiability

in our considered setting. In App. D, we also discuss identifiability when only imperfect stochastic interventions are available.

While Thm. 4.2 (ii) provides sufficient conditions for identifiability up to  $\mathcal{S}_{\text{scaling}}$ , we prove below that the assumptions in Thm. 4.2 (ii) are *necessary*.

**Proposition 4.3.** *Given a DAG  $G$ , with  $d - 1$  perfect stochastic single node interventions on distinct targets, if the remaining unintervened node has any parent in  $G$ ,  $(G, \mathcal{F}, P_G)$  is not identifiable up to*

$$\mathcal{S}_{\text{reparam}} := \left\{ \mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{g} = \mathbf{P} \circ \mathbf{h}, \mathbf{P} \text{ is a permutation matrix, } \mathbf{h} \in \mathcal{S}_{\text{scaling}} \right\}. \quad (6)$$

A direct implication of Prop. 4.3 is that for any nonlinear causal representation learning problem,  $d - 1$  single-node interventions are not sufficient for identifiability.

## 4.2 Special Case: ICA with stochastic interventions on the latent components

An important special case of CauCA when the graph  $G$  is empty, corresponding to independent latent components. This defines a nonlinear ICA generative model where, additionally, we observe a variable  $\tau_k$  which indicates *which latent distributions change* in the interventional regime  $k$ , while *every other distribution is unchanged*. This nonlinear ICA generative model is closely related to similar models with *observed auxiliary variables* (Hyvärinen et al., 2019; Khemakhem et al., 2020a). It is natural to interpret  $\tau_k$  as an auxiliary variable: as we will see, our interventional interpretation allows us to derive novel results and re-interpret existing ones. In the following, we characterize identifiability for this setting.

**Single-node interventions.** We first focus on *single-node* stochastic interventions, where the following result proves that we can achieve the same level of identifiability as in Thm. 4.2 (ii) with one intervention less than in the case where the graph is non-trivial.

**Proposition 4.4.** *Suppose that  $G$  is the empty graph, and that there are  $d - 1$  variables intervened on, with one single target per dataset, such that Asm. 4.1 is satisfied. Then CauCA (in this case, ICA) in  $(G, \mathcal{F}, P_G)$  is identifiable up to  $\mathcal{S}_{\text{scaling}}$  defined as in eq. (3).*

The result above shows that identifiability can be achieved through single-node interventions on the latent variables using strictly fewer datasets (i.e., auxiliary variables) than previous results in the auxiliary variables setting ( $d - 1$  in our case,  $2d + 1$  in Hyvärinen et al. (2019)[Thm. 1]). One potentially confusing aspect of Prop. 4.4 is that the ambiguity set does not contain permutations—which is usually an unavoidable ambiguity in ICA. This is due to our considered setting with known targets, where a total ordering of the variables is assumed to be known. The result above can also be extended to the case of *unknown intervention targets*, where we only know that for each regime a distinct variable is intervened on, but we do not know which one: see App. E. For that case, we prove in Prop. E.6 that ICA in  $(G, \mathcal{F}, P_G)$  is in fact identifiable up to scaling and permutation.

We can additionally show that for nonlinear ICA,  $d - 1$  interventions are *necessary* for identifiability.

**Proposition 4.5.** *Given an empty graph  $G$ , with  $d - 2$  single-node interventions on distinct targets, with one single target per dataset, such that Asm. 4.1 is satisfied. Then CauCA (in this case, ICA) in  $(G, \mathcal{F}, P_G)$  is not identifiable up to  $\mathcal{S}_{\text{reparam}}$ .*

**Fat-hand interventions** A generalization of single-node interventions are *fat-hand interventions*: i.e., interventions where  $|\tau_k| > 1$ . In this section, we study this more general setting and focus on a weaker form of identification than for single-node intervention.

**Theorem 4.6.** *Suppose  $G$  is the empty graph. Suppose that our datasets encompass interventions over all variables in the latent graph, i.e.,  $\bigcup_{k \in [K]} \tau_k = [d]$ . Suppose for every  $k$ , the targets of interventions are a strict subset of all variables, i.e.,  $|\tau_k| = n_k$ ,  $n_k \in [d - 1]$ .*

(Block-interventional discrepancy) *Suppose that there are  $n_k$  interventions with target  $\tau_k$  such that  $\mathbf{v}_k(\mathbf{z}_{\tau_k}, 1) - \mathbf{v}_k(\mathbf{z}_{\tau_k}, 0), \dots, \mathbf{v}_k(\mathbf{z}_{\tau_k}, n_k) - \mathbf{v}_k(\mathbf{z}_{\tau_k}, 0)$  are linearly independent, where*

$$\mathbf{v}_k(\mathbf{z}_{\tau_k}, s) := ((\ln q_{k,1}^s)'(z_{\tau_k,1}), \dots, (\ln q_{k,n_k}^s)'(z_{\tau_k,n_k})) \quad (7)$$

*where  $q_k^s$  is the intervention of the  $s$ -th interventional regime that has the target  $\tau_k$ , and  $q_{k,j}^s$  is the  $j$ -th marginal of it.  $z_{\tau_k,j}$  is the  $j$ -th dimension of  $\mathbf{z}_{\tau_k}$ .  $s = 0$  denotes the unintervened regime.*

Then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is block-identifiable (following von Kügelgen et al. (2021)): namely, if  $\mathbf{f}_* \mathbb{P}^k = \mathbf{f}'_* \mathbb{Q}^k$  then for  $\varphi := \mathbf{f}'^{-1} \circ \mathbf{f}$ , for all  $k \in [K]$ ,

$$[\varphi(\mathbf{z})]_{\tau_k} = \varphi_{\tau_k}(\mathbf{z}_{\tau_k}). \quad (8)$$

The *block-interventional discrepancy* assumption is tightly connected to Asm. 4.1: for the case with an empty graph, and if  $\forall k : n_k = 1$ , linear independence of vectors block-interventional discrepancy is equal to Asm. 4.1. However, note that Thm. 4.6 is not a special case of Thm. 4.2, since the former only holds for the case with an empty graph, whereas the latter holds for general DAGs. Finally, we remark that Prop. 4.4 is not a special case of Thm. 4.6 in which  $n_k = 1 \forall k$ , since Prop. 4.4 only requires  $d - 1$  interventions instead of  $d$ .

Our interventional perspective also allows us to re-interpret and extend (Hyvarinen et al., 2019)[Thm.1]. In particular, the following Proposition holds.

**Proposition 4.7.** *Under the assumptions of Thm. 4.6, suppose there exist  $k \in [K]$  and there are  $2n_k$  interventions with targets  $\tau_k$  such that for any  $\mathbf{z}_{\tau_k} \in \mathbb{R}^{n_k}$ ,  $\mathbf{w}_k(\mathbf{z}_{\tau_k}, 1) - \mathbf{w}_k(\mathbf{z}_{\tau_k}, 0), \dots, \mathbf{w}_k(\mathbf{z}_{\tau_k}, 2n_k) - \mathbf{w}_k(\mathbf{z}_{\tau_k}, 0)$  are linearly independent, where*

$$\mathbf{w}_k(\mathbf{z}_{\tau_k}, s) := \left( \left( \frac{q_{k,1}^{s'}}{q_{k,1}^s} \right)' (z_{\tau_k,1}), \dots, \left( \frac{q_{k,n_k}^{s'}}{q_{k,n_k}^s} \right)' (z_{\tau_k,n_k}), \frac{q_{k,1}^{s'}}{q_{k,1}^s} (z_{\tau_k,1}), \dots, \frac{q_{k,n_k}^{s'}}{q_{k,n_k}^s} (z_{\tau_k,n_k}) \right)$$

where  $q_k^s$  is the intervention of the  $s$ -th interventional regime that has the target  $\tau_k$ , and  $q_{k,j}^s$  is the  $j$ -th marginal of it.  $z_{\tau_k,j}$  is the  $j$ -th dimension of  $\mathbf{z}_{\tau_k}$ .  $s = 0$  denotes the unintervened regime. Then

$$\varphi_{\tau_k} \in \mathcal{S}_{reparam} := \left\{ \mathbf{g} : \mathbb{R}^{n_k} \rightarrow \mathbb{R}^{n_k} \mid \mathbf{g} = \mathbf{P} \circ \mathbf{h} \text{ where } \mathbf{P} \text{ is a permutation matrix and } \mathbf{h} \in \mathcal{S}_{scaling} \right\}$$

**Remark 4.8.** The assumption of linear independence  $\mathbf{w}_k(\mathbf{z}_{\tau_k}, s), s \in [2n_k]$  is precisely the assumption of *variability* in (Hyvarinen et al., 2019)[Thm. 1]; however, we only assume it within a  $n_k$ -dimensional block (not over all  $d$  variables). We refer to it as *block-variability*.

Note that the block-variability assumption *implies* block-interventional discrepancy: i.e., block-variability is strictly stronger. Correspondingly, it leads to a stronger identification. Block-interventional discrepancy only allows *block-wise identifiability* within the  $n_k$ -dimensional intervened blocks based on  $n_k$  interventions. In contrast, the variability assumption can be interpreted as a sufficient assumption to achieve identification *up to permutation and scaling* within a  $n_k$ -dimensional block, based on  $2n_k$  fat-hand interventions (in both cases together with one un-intervened dataset). See also Fig. 3 for a visualization.

In the literature on nonlinear ICA with auxiliary variables, the variability assumption is assumed to hold over *all* variables, which in our setting can be interpreted as a requirement over  $2d$  fat-hand interventions over all latent variables simultaneously (plus one un-intervened distribution). In this sense, Prop. 4.7 and *block-variability* extend (Hyvarinen et al., 2019)[Thm. 1], which only considers the case where *all* variables are intervened, by exploiting variability to achieve a strong identification only *within a subset of the variables*.

## 5 Experiments

Our experiments aim to estimate a CauCA model based on a known graph and a collection of interventional datasets with known targets. We focus on the scenarios with single-node, perfect interventions described in § 4. For additional technical details, see App. G.

**Synthetic data-generating process.** We first sample DAGs  $G$  with an edge density of 0.5. To model the causal dependence among the latent variables, we use the family of CBNs induced by

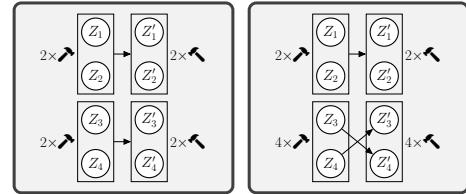
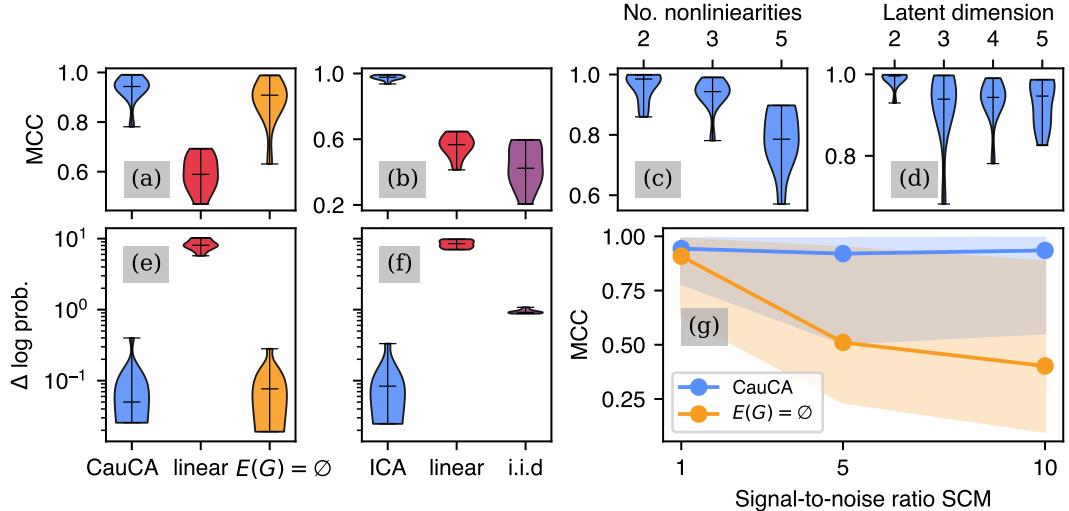


Figure 3: We use the hammer symbol together with a “times” symbol to represent how many interventional regimes are required by the two assumptions. (Left) (Thm. 4.6) For the *block-interventional discrepancy* assumption, we need  $n_k$  interventions to get to block-identification of  $\mathbf{z}_{\tau_k}$ . (Right) (Prop. 4.7) For the *block-variability* assumption, we need  $2n_k$  to get to elementwise identification up to scaling and permutation.



**Figure 4: Experimental results.** Figures (a) and (e) present the mean correlation coefficients (MCC) between true and learned latents and log-probability differences between the model and ground truth ( $\Delta \log \text{prob.}$ ) for Causal Component Analysis (CauCA) experiments. Misspecified models assuming a trivial graph ( $E(G) = \emptyset$ ) and a linear encoder function class are compared. All violin plots show the distribution of outcomes for 10 pairs of CBNs and mixing functions. Figures (c) and (d) display CauCA results with varying numbers of nonlinearities in the mixing function and latent dimension. For the ICA setting, MCC values and log probability differences are illustrated in (b) and (f). Baselines include a misspecified model (linear mixing) and a naive (single-environment) unidentifiable normalizing flow with an independent Gaussian base distribution (labelled *i.i.d.*). The naive baseline is trained on pooled data without using information about interventions and their targets. Figure (g) shows the median MCC for CauCA and the misspecified baseline ( $E(G) = \emptyset$ ) as the strength of the linear parameters relative to the exogenous noise in the structural causal model generating the CBN increases. The shaded areas show the range between minimum and maximum values. For detailed experimental information, refer to App. G.

linear Gaussian structural causal model (SCM) consistent with  $G$ . For the ground-truth mixing function, we use  $M$ -layer multilayer perceptrons  $\mathbf{f} = \sigma \circ \mathbf{A}_M \circ \dots \circ \sigma \circ \mathbf{A}_1$ , where  $\mathbf{A}_m \in \mathbb{R}^{d \times d}$  for  $m \in \llbracket 1, M \rrbracket$  denote invertible linear maps (sampled from a multivariate uniform distribution), and  $\sigma$  is an element-wise invertible nonlinear function. We then sample observed mixtures from these latent CBNs as described in eq. (2).

**Likelihood-based estimation procedure.** Our objective is to learn an encoder  $\mathbf{g}_{\theta} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  that approximates the inverse function  $\mathbf{f}^{-1}$  up to tolerable ambiguities, together with latent densities  $(p_{\theta}^k)_{k \in \llbracket 0, d \rrbracket}$  reproducing the ground truth up to corresponding ambiguities (cf. Lemma 3.3). We estimate the encoder parameters by maximizing the likelihood, which can be derived through a change of variables from eq. (1): for an observation in dataset  $k > 0$  taking a value  $\mathbf{x}$ , it is given by

$$\log p_{\theta}^k(\mathbf{x}) = \log |\det \mathbf{J}\mathbf{g}_{\theta}(\mathbf{x})| + \tilde{p}_{\tau_k}((\mathbf{g}_{\theta})_{\tau_k}(\mathbf{x})) + \sum_{i \neq \tau_k} \log p_i((\mathbf{g}_{\theta})_i(\mathbf{x}) | (\mathbf{g}_{\theta})_{\text{pa}(i)}(\mathbf{x})), \quad (9)$$

where  $\mathbf{J}\mathbf{g}_{\theta}(\mathbf{x})$  denotes the Jacobian of  $\mathbf{g}_{\theta}$  evaluated at  $\mathbf{x}$ . The learning objective can be expressed as  $\theta^* = \arg \max_{\theta} \sum_{k=0}^K \left( \frac{1}{N_k} \sum_{n=1}^{N_k} \log p_{\theta}^k(\mathbf{x}^{(n,k)}) \right)$ , with  $N_k$  representing the size of dataset  $k$ .

**Model architecture.** We employ normalizing flows (Papamakarios et al., 2021) to parameterize the encoder. Instead of the typically deployed base distribution with independent components, we use the collection of densities (one per interventional regime) induced by the CBN over the latents. Following the CauCA setting, the parameters of the causal mechanisms are learned while the causal graph is assumed known. For details on the model and training parameters, please refer to App. G.

**Settings.** We investigate two learning problems: (i) CauCA, corresponding to § 4.1, and (ii) ICA, where the sampled graphs in the true latent CBN contain no arrows, as discussed in § 4.2.

**Results.** (i) For a data-generating process with non-empty graph, experimental outcomes are depicted in Fig. 4 (a, e). We compare a well-specified CauCA model (blue) to misspecified baselines, including a model with correctly specified latent model but employing a linear encoder (red), and a model with a nonlinear encoder but assuming a causal graph with no arrows (orange). See caption of Fig. 4

for details on the metrics. The results demonstrate that the CauCA model accurately identifies the latent variables, benefiting from both the nonlinear encoder and the explicit modelling of causal dependence. We additionally test the effect of increasing a parameter influencing the magnitude of the sampled linear parameters in the SCM (we refer to this as *signal-to-noise ratio*, see App. G for details)—which increases the statistical *dependence* among the true latent components. The gap between the CauCA model and the baseline assuming a trivial graph widens (Fig. 4 (g)), indicating that *correctly modelling the causal relationships becomes increasingly important the more dependent the true latent variables are*. Finally, we verify that the model performs well for different number of layers  $M$  in the ground-truth nonlinear mixing (c) (performance degrades slightly for higher  $M$ ), and across various latent dimensionalities for the latent variable (d).

(ii) For data-generating processes where the graph contains no arrows (ICA), results are presented in Fig. 4 (b, f). Well-specified, nonlinear models (*blue*) are compared to misspecified linear baselines (*red*) and a naive normalizing flow baseline trained on pooled data (*purple*). The findings confirm that *interventional information provides useful learning signal even in the context of nonlinear ICA*.

## 6 Related Work and Discussion

**Causal representation learning.** In the present work, we focus on identifiability of *latent CBNs* with a *known graph*, based on *interventional data*, and investigate the *nonlinear and nonparametric* case. In CRL (*unknown graph*), many studies focus on identifiability of latent *SCMs* instead, which requires strong assumptions such as weak supervision (i.e., *counterfactual data*) (Locatello et al., 2020; von Kügelgen et al., 2021; Ahuja et al., 2022a; Brehmer et al., 2022). Alternatively, the setting where *temporal information* is available, i.e., dynamic Bayesian networks, has been studied extensively (Lachapelle et al., 2022; Lachapelle and Lacoste-Julien, 2022; Yao et al., 2021; Lippe et al., 2022, 2023). In a non-temporal setting, Squires et al. (2023); Varici et al. (2023) assume interventional data and *linear mixing functions*; and Liu et al. (2022) assume that the *latent distributions are linear Gaussian*. Ahuja et al. (2022b) identify latent representations by *deterministic hard interventions*, together with *parametric assumptions* on the mixing and an *independent support assumption*.

**Prior knowledge on the latent SCM.** Other prior works also leverage prior knowledge on the causal structure for representation learning. Yang et al. (2020) introduce the CausalVAE model, which aims to disentangle the endogenous and exogenous variables of an SCM, and prove identifiability up to affine transformations based on known intervention targets. Shen et al. (2022) also consider the setting in which the graph is (partially) known, but their approach requires additional supervision in the form of annotations of the ground truth latent. Leeb et al. (2023) embed an SCM into the latent space of an autoencoder, provided with a topological ordering allowing it to learn latent DAGs.

**Statistically dependent components.** Models with causal dependences among the latent variables are a special case of models where the latent variables are statistically dependent (Hyvärinen et al., 2023). Various extensions of the ICA setting allow for dependent variables: independent subspace analysis consider block-wise independent variables (Hyvärinen and Hoyer, 2000) (a similar notion is exploited by Lyu et al. (2022)), and topographic ICA models variable dependencies following a ‘topographic’ arrangement (Hyvärinen et al., 2001) (see also (Keller and Welling, 2021)). Independently modulated component analysis (Khemakhem et al., 2020b) extends previous work on identifiability in nonlinear ICA with auxiliary variables by allowing latent dependences which are assumed to be stationary, or independent of the auxiliary variable. Morioka and Hyvarinen (2023) introduce a multi-modal model where *within-modality dependence* among the latents is described by a modality-specific Bayesian Network, with *joint independence across the modalities*, and a mixing function for same-index variables across these networks. Unlike our work, it encodes no explicit notions of interventions.

**CauCA as a causal generalization of ICA.** As pointed out in § 4.2, the special case of CauCA with a trivial graph corresponds to a novel ICA model. Beyond the fact that CauCA allows statistical dependence described by general DAGs among the components, we argue that it can be viewed as a *causal* generalization of ICA. Firstly, we exploit the assumption of *localized and sparse* changes in the latent mechanisms (Schölkopf et al., 2021; Perry et al., 2022), in contrast to previous ICA works which exploit *non-stationarity* at the level of the entire joint distribution of the latent components (Hyvarinen and Morioka, 2016; Hyvarinen et al., 2019; Monti et al., 2020), leading to strong identifiability results (e.g., in Thm. 4.2 (ii)). Secondly, we exploit the modularity of causal mechanisms: i.e., it is possible to intervene on some of the mechanisms while leaving the others

*invariant* (Pearl, 2009; Peters et al., 2017). To the best of our knowledge, our work is the first ICA extension where latent dependence can actually be interpreted in a causal sense.

## Acknowledgements

The authors thank Vincent Stimper, Weiyang Liu, Siyuan Guo, Jinglin Wang, Corentin Correia, and Cian Eastwood for helpful comments and discussions.

## Funding Transparency Statement

This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 1IS18039B; L.G. was supported by the VideoPredict project, FKZ: 01IS21088.

## References

- Ahuja, K., Hartford, J. S., and Bengio, Y. (2022a). Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528. [Cited on page 9.]
- Ahuja, K., Wang, Y., Mahajan, D., and Bengio, Y. (2022b). Interventional causal representation learning. *arXiv preprint arXiv:2209.11924*. [Cited on page 9.]
- Brehmer, J., De Haan, P., Lippe, P., and Cohen, T. (2022). Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*. [Cited on pages 2, 9, and 16.]
- Buchholz, S., Besserve, M., and Schölkopf, B. (2022). Function classes for identifiable nonlinear independent component analysis. *arXiv preprint arXiv:2208.06406*. [Cited on pages 2 and 4.]
- Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314. [Cited on page 1.]
- Darmois, G. (1951). Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231. [Cited on page 1.]
- Durkan, C., Bekasov, A., Murray, I., and Papamakarios, G. (2019). Neural spline flows. *Advances in Neural Information Processing Systems*, 32. [Cited on page 34.]
- Greselle, L., Fissore, G., Javaloy, A., Schölkopf, B., and Hyvärinen, A. (2020). Relative gradient optimization of the jacobian term in unsupervised deep learning. *Advances in neural information processing systems*, 33:16567–16578. [Cited on page 34.]
- Greselle, L., Rubenstein, P. K., Mehrjou, A., Locatello, F., and Schölkopf, B. (2019). The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ICA. In *Uncertainty in Artificial Intelligence*, pages 217–227. PMLR. [Cited on page 2.]
- Greselle, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. (2021). Independent mechanism analysis, a new concept? In *Advances in neural information processing systems*, volume 34, pages 28233–28248. [Cited on page 2.]
- Hälvä, H. and Hyvärinen, A. (2020). Hidden markov nonlinear ICA: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR. [Cited on page 1.]
- Hälvä, H., Le Corff, S., Lehéricy, L., So, J., Zhu, Y., Gassiat, E., and Hyvärinen, A. (2021). Disentangling identifiable features from noisy data with structured nonlinear ICA. *Advances in Neural Information Processing Systems*, 34:1624–1633. [Cited on page 1.]
- Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720. [Cited on page 9.]

- Hyvärinen, A., Hoyer, P. O., and Inki, M. (2001). Topographic independent component analysis. *Neural computation*, 13(7):1527–1558. [Cited on page 9.]
- Hyvärinen, A., Khemakhem, I., and Monti, R. (2023). Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*. [Cited on pages 1, 4, and 9.]
- Hyvarinen, A. and Morioka, H. (2016). Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. *Advances in neural information processing systems*, 29. [Cited on pages 2 and 9.]
- Hyvarinen, A. and Morioka, H. (2017). Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR. [Cited on page 1.]
- Hyvärinen, A. and Pajunen, P. (1999). Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439. [Cited on pages 1, 22, and 25.]
- Hyvarinen, A., Sasaki, H., and Turner, R. (2019). Nonlinear ICA using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR. [Cited on pages 2, 6, 7, 9, 24, and 25.]
- Keller, T. A. and Welling, M. (2021). Topographic VAEs learn equivariant capsules. *Advances in Neural Information Processing Systems*, 34:28585–28597. [Cited on page 9.]
- Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020a). Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR. [Cited on pages 2 and 6.]
- Khemakhem, I., Monti, R., Kingma, D., and Hyvarinen, A. (2020b). Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ICA. *Advances in Neural Information Processing Systems*, 33:12768–12778. [Cited on pages 2 and 9.]
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. [Cited on page 34.]
- Lachapelle, S. and Lacoste-Julien, S. (2022). Partial disentanglement via mechanism sparsity. *arXiv preprint arXiv:2207.07732*. [Cited on pages 2 and 9.]
- Lachapelle, S., Rodriguez, P., Sharma, Y., Everett, K. E., Le Priol, R., Lacoste, A., and Lacoste-Julien, S. (2022). Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR. [Cited on pages 2 and 9.]
- Leeb, F., Lanzillotta, G., Annadani, Y., Besserve, M., Bauer, S., and Schölkopf, B. (2023). Structure by architecture: Structured representations without regularization. *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*. arXiv preprint 2006.07796. [Cited on page 9.]
- Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media. [Cited on page 1.]
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, E. (2023). Causal representation learning for instantaneous and temporal effects in interactive systems. In *The Eleventh International Conference on Learning Representations*. [Cited on page 9.]
- Lippe, P., Magliacane, S., Löwe, S., Asano, Y. M., Cohen, T., and Gavves, S. (2022). Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR. [Cited on pages 2 and 9.]
- Liu, Y., Zhang, Z., Gong, D., Gong, M., Huang, B., van den Hengel, A., Zhang, K., and Qinfeng Shi, J. (2022). Weight-variant latent causal models. *arXiv e-prints*, pages arXiv–2208. [Cited on page 9.]

- Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., and Tschannen, M. (2020). Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR. [Cited on pages 2 and 9.]
- Lyu, Q., Fu, X., Wang, W., and Lu, S. (2022). Understanding latent correlation-based multiview learning and self-supervision: An identifiability perspective. In *International Conference on Learning Representations*. [Cited on page 9.]
- Monti, R. P., Zhang, K., and Hyvärinen, A. (2020). Causal discovery with general non-linear relationships using non-linear ICA. In *Uncertainty in artificial intelligence*, pages 186–195. PMLR. [Cited on page 9.]
- Morioka, H. and Hyvarinen, A. (2023). Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International Conference on Artificial Intelligence and Statistics*, pages 3399–3426. PMLR. [Cited on page 9.]
- Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing flows for probabilistic modeling and inference. *The Journal of Machine Learning Research*, 22(1):2617–2680. [Cited on pages 8 and 34.]
- Pearl, J. (2009). *Causality*. Cambridge university press. [Cited on pages 2, 3, and 10.]
- Perry, R., Von Kügelgen, J., and Schölkopf, B. (2022). Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. In *Advances in Neural Information Processing Systems*. [Cited on page 9.]
- Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of causal inference: foundations and learning algorithms*. The MIT Press. [Cited on pages 3 and 10.]
- Sauer, A. and Geiger, A. (2021). Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*. [Cited on page 2.]
- Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634. [Cited on pages 2 and 9.]
- Shen, X., Liu, F., Dong, H., Lian, Q., Chen, Z., and Zhang, T. (2022). Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23:1–55. [Cited on page 9.]
- Spirites, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*. MIT press. [Cited on page 3.]
- Squires, C., Seigal, A., Bhate, S., and Uhler, C. (2023). Linear causal disentanglement via interventions. [Cited on pages 2, 5, and 9.]
- Tangemann, M., Schneider, S., von Kügelgen, J., Locatello, F., Gehler, P., Brox, T., Kümmerer, M., Bethge, M., and Schölkopf, B. (2021). Unsupervised object learning via common fate. In *2nd Conference on Causal Learning and Reasoning (CLEaR)*. arXiv:2110.06562. [Cited on page 2.]
- Träuble, F., Creager, E., Kilbertus, N., Locatello, F., Dittadi, A., Goyal, A., Schölkopf, B., and Bauer, S. (2021). On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, pages 10401–10412. PMLR. [Cited on page 2.]
- Varici, B., Acarturk, E., Shanmugam, K., Kumar, A., and Tajer, A. (2023). Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*. [Cited on pages 2 and 9.]
- von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. (2021). Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467. [Cited on pages 2, 7, 9, and 23.]
- Wasserman, L. (2004). *All of statistics: a concise course in statistical inference*, volume 26. Springer. [Cited on page 1.]

- Xi, Q. and Bloem-Reddy, B. (2023). Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, pages 6912–6939. PMLR. [Cited on page 1.]
- Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., and Wang, J. (2020). Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv e-prints*, pages arXiv–2004. [Cited on page 9.]
- Yao, W., Sun, Y., Ho, A., Sun, C., and Zhang, K. (2021). Learning temporally causal latent processes from general temporal data. *arXiv preprint arXiv:2110.05428*. [Cited on page 9.]
- Zhang, K. and Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *25th Conference on Uncertainty in Artificial Intelligence (UAI 2009)*, pages 647–655. AUAI Press. [Cited on page 2.]

# APPENDIX

## Overview

- App. A recapitulates the notation used in this paper.
- App. B contains the proofs of all theoretical statements presented in the paper.
- App. C contains a nontrivial counterexample of Thm. 4.2 (ii) when Asm. 4.1 is violated.
- App. D contains an additional result of identifiability based on Thm. 4.2 (i), when more imperfect stochastic interventions are available.
- App. E contains a general discussion on CauCA with unknown intervention targets, as well as a generalization of some of the identifiability results.
- App. F contains some theoretical results which were useful in the design of the experiments.
- App. G contains the details of the experiments.

## A Notations

Symbol	Description
$\bar{G}$	A directed acyclic graph with nodes $V(\bar{G}) = [d]$ and arrows $E(\bar{G})$
$(i, j)$	An ordered tuple representing an arrow in $E(\bar{G})$ , with $i, j \in V(\bar{G})$
$\llbracket i, j \rrbracket$	The integers $i, \dots, j$
$[d]$	The natural numbers $1, \dots, d$
$\text{pa}(i)$	Parents of $i$ , defined as $\{j \in V(\bar{G}) \mid (j, i) \in E(\bar{G})\}$
$\text{pa}^k(j)$	Parents of $j$ in the post-intervention graph in the intervention regime $k$
$\bar{\text{pa}}(i)$	Closure of the parents of $i$ , defined as $\text{pa}(i) \cup \{i\}$
$\text{anc}(i)$	Ancestors of $i$ , nodes $j$ in $\bar{G}$ such that there is a directed path from $j$ to $i$
$\bar{\text{anc}}(i)$	Closure of the ancestors of $i$ , defined as $\text{anc}(i) \cup \{i\}$
$\bar{G}$	Transitive closure of $G$ defined by $\text{pa}^{\bar{G}}(i) := \text{anc}^{\bar{G}}(i)$
$X, Y, Z$	Unidimensional random variables
$\mathbf{X}, \mathbf{Y}, \mathbf{Z}$	Multidimensional random variables
$x, y, z$	Scalars in $\mathbb{R}$
$\mathbf{x}, \mathbf{y}, \mathbf{z}$	Vectors in $\mathbb{R}^d$
$\mathbf{z}_{[i]}$	The $(1, \dots, i)$ dimensions of $\mathbf{z}$
$\varphi_i$	The function that outputs the $i$ -th dimension of the mapping $\varphi$
$\varphi_{[i]}$	The mapping that outputs the $(1, \dots, i)$ dimensions of the mapping $\varphi$
$\tau_k$	Intervention targets in interventional regime $k$
$\mathbb{P}, \mathbb{Q}$	Probability distributions
$p, q$	Density functions of $\mathbb{P}, \mathbb{Q}$
$\mathbb{P}_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)})$	Causal mechanism of variable $Z_i$
$\tilde{\mathbb{P}}_i^k(Z_i \mid \mathbf{Z}_{\text{pa}^k(i)})$	Intervened mechanism of variable $Z_i$ in interventional regime $k$
$\mathbb{P}^k(\mathbf{Z})$	$k \neq 0$ : interventional distribution in interventional regime $k$ (Defn. 2.2) $k = 0$ : unintervened distribution
$\mathcal{P}_G$	Class of latent joint probabilities that are Markov relative to $G$ , and absolutely continuous with full support in $\mathbb{R}^d$
$\mathcal{F}$	Function class of mixing function/decoders
$\mathcal{S}$	In this paper, it is assumed to be all $\mathcal{C}^1$ -diffeomorphisms
$\mathbf{f}$	Indeterminacy set, defined in Defn. 3.2
$\mathbf{f}_*\mathbb{P}$	Mixing function or decoder, a diffeomorphism $\mathbb{R}^d \rightarrow \mathbb{R}^d$
$G$	The pushforward measure of $\mathbb{P}$ by $\mathbf{f}$
$G \models G$	A directed acyclic graph without indices
$G \models G$	$G$ is an indexed graph of $G$ , i.e. $V(G) = [d]$ and there exists an isomorphism $G \rightarrow G$
$\text{Aut}_G$	Group of automorphisms of graph $G$
$\mathfrak{S}_d$	Group of permutations of $d$ elements

## B Proofs

### B.1 Lemmata

**Lemma 3.3.** For any  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  in  $(G, \mathcal{F}, \mathcal{P}_G)$ , and for any  $\mathbf{h} \in \mathcal{S}_{\text{scaling}}$  with

$$\mathcal{S}_{\text{scaling}} := \{ \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{h}(\mathbf{z}) = (h_1(z_1), \dots, h_d(z_d)), h_i \text{ is a diffeomorphism in } \mathbb{R} \} \quad (3)$$

there exists a  $(G, \mathbf{f} \circ \mathbf{h}, (\mathbb{Q}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  in  $(G, \mathcal{F}, \mathcal{P}_G)$  s.t.  $\mathbf{f}_*\mathbb{P}^k = (\mathbf{f} \circ \mathbf{h})_*\mathbb{Q}^k$  for all  $k \in \llbracket 0, K \rrbracket$ .

*Proof.* For any  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in \llbracket 0, K \rrbracket})$  in  $(G, \mathcal{F}, \mathcal{P}_G)$ , and for any  $\mathbf{h} \in \mathcal{S}_{\text{scaling}}$ , define  $\mathbf{g} := \mathbf{h}^{-1}$ , then  $\mathbf{g} \in \mathcal{S}_{\text{scaling}}$ . Define  $\mathbb{Q}^0 := \mathbf{g}_*\mathbb{P}^0$ . By Lemma B.3, for all  $i \in [d]$ ,  $\mathbb{Q}_i(\cdot | \mathbf{h}_{\text{pa}(i)}(\mathbf{z}_{\text{pa}(i)})) = (g_i)_*\mathbb{P}_i(\cdot | \mathbf{z}_{\text{pa}(i)})$ .

For  $k \in [K]$ , define

$$\tilde{\mathbb{Q}}_j^k(\cdot | \mathbf{g}_{\text{pa}^k(j)}(\mathbf{z}_{\text{pa}^k(j)})) := (g_j)_*\tilde{\mathbb{P}}_j^k(\cdot | \mathbf{z}_{\text{pa}^k(j)}) \quad (10)$$

Define  $\mathbb{Q}^k := \prod_{j \in \tau_k} \tilde{\mathbb{Q}}_j^k \prod_{j \notin \tau_k} \mathbb{Q}_j$ , by Lemma B.3 and (10),  $\mathbb{Q}^k = \mathbf{g}_* \mathbb{P}^k \forall k \in [K]$ . By definition of  $\mathbb{Q}^0$ ,  $\mathbb{Q}^k = \mathbf{g}_* \mathbb{P}^k \forall k \in [0, K]$ . i.e.,  $\mathbf{f}_* \mathbb{P}^k = (\mathbf{f} \circ \mathbf{h})_* \mathbb{Q}^k$ .

□

**Lemma B.1** (Lemma 2 of Brehmer et al. (2022)). *Let  $A = C = \mathbb{R}$  and  $B = \mathbb{R}^d$ . Let  $f : A \times B \rightarrow C$  be differentiable. Define differentiable measures  $\mathbb{P}_A$  on  $A$  and  $\mathbb{P}_C$  on  $C$ . Let  $\forall b \in B$ ,  $f(\cdot, b) : A \rightarrow C$  be measure-preserving, i.e.  $\mathbb{P}_C = f(\cdot, b)_* \mathbb{P}_A$ . Then  $f$  is constant in  $b$  over  $B$ .*

**Lemma B.2.** *For any distributions  $\mathbb{P}, \mathbb{Q}$  of full support on  $\mathbb{R}$ , with c.d.f  $F, G$ , there are only two diffeomorphisms  $T : \mathbb{R} \rightarrow \mathbb{R}$  such that  $T_* \mathbb{P} = \mathbb{Q}$ : they are  $G^{-1} \circ F$  and  $\bar{G}^{-1} \circ F$ , where  $\bar{G}(x) := 1 - G(x)$ .*

*Proof.*  $T$  is a diffeomorphism, then  $T'(x) \neq 0 \quad \forall x \in \mathbb{R}$ . Then the sign of  $T'(x)$  is either positive or negative everywhere.  $T$  is either strictly increasing or strictly decreasing on  $\mathbb{R}$ . Since  $\mathbb{P}$  and  $\mathbb{Q}$  are full support in  $\mathbb{R}$ ,  $F$  and  $G$  are strictly increasing in  $\mathbb{R}$ .

- If  $T$  is increasing:  $T_* \mathbb{P} = \mathbb{Q}$  implies that  $G(x) = \mathbb{Q}(X \leq x) = \mathbb{P}(T(X) \leq x)$ . Since  $T$  is strictly increasing,  $G(x) = \mathbb{P}(T(X) \leq x) = \mathbb{P}(X \leq T^{-1}(x)) = F \circ T^{-1}(x)$ . Thus  $x = G^{-1} \circ F \circ T^{-1}(x)$ .  $T = G^{-1} \circ F$ .
- If  $T$  is decreasing:  $G(x) = \mathbb{Q}(X \leq x) = \mathbb{P}(T(X) \leq x) = \mathbb{P}(X \geq T^{-1}(x)) = 1 - F(T^{-1}(x))$ . Thus  $T = \bar{G}^{-1} \circ F$ .

□

**Lemma B.3.** *Suppose  $\mathbb{P}, \mathbb{Q}$  Markov relative to  $G$ , absolutely continuous with full support in  $\mathbb{R}^d$ . Fix any functions  $\varphi_1, \dots, \varphi_d$  diffeomorphisms strictly monotonic in  $\mathbb{R}$ . Let  $\varphi := (\varphi_i)_{i \in [d]}$ . The following statements are equivalent:*

- (1)  $\mathbb{Q} = \varphi_* \mathbb{P}$
- (2)  $\forall i \in [d], \forall z_i \in \mathbb{R}, \forall \mathbf{z}_{pa(i)} \in \mathbb{R}^{\#pa(i)}, p_i(z_i | \mathbf{z}_{pa(i)}) = q_i(\varphi_i(z_i) | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) |\varphi'_i(z_i)|$ , or equivalently,  $\mathbb{Q}_i(\cdot | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) = (\varphi_i)_* \mathbb{P}_i(\cdot | \mathbf{z}_{pa(i)})$

*Proof.*  $\mathbb{Q}_i(\cdot | \varphi_{pa(i)}(\mathbf{z}_{pa(i)}))$  denotes the conditional probability of  $Z_i$ :  $\mathbb{Q}_i(Z_i | \varphi_{pa(i)}(\mathbf{z}_{pa(i)}))$ .

(2)  $\Rightarrow$  (1): Multiply the equations in (2) for  $n$  indices,

$$\prod_{i=1}^d p_i(z_i | \mathbf{z}_{pa(i)}) = \prod_{i=1}^d q_i(\varphi_i(z_i) | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) |\varphi'_i(z_i)|$$

Since  $\prod_{i=1}^d |\varphi'_i(z_i)| = |\det D\varphi(\mathbf{z})|$ , we obtain the equation in (1).

(1)  $\Rightarrow$  (2): without loss of generality, choose a total order on  $V(G)$  that preserves the partial order of  $G$ :  $i > j$  if  $i \in pa(j)$ . Since  $\varphi$  is a diffeomorphism, by the change of variables formula,

$$p(\mathbf{z}) = q(\varphi(\mathbf{z})) |\det D\varphi(\mathbf{z})| \tag{11}$$

Since  $\mathbb{P}, \mathbb{Q}$  are Markov relative to  $G$ , write  $p, q$  as the factorization according to  $G$ :

$$\prod_{i=1}^d p_i(z_i | \mathbf{z}_{pa(i)}) = \prod_{i=1}^d q_i(\varphi_i(z_i) | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) |\varphi'_i(z_i)| \tag{12}$$

We will show by induction on the reverse order of  $[d]$  that for all  $i \in [d]$ ,

$$p_i(z_i | \mathbf{z}_{pa(i)}) = q_i(\varphi_i(z_i) | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) |\varphi'_i(z_i)|$$

Marginalize over  $z_d$ ,

$$\prod_{i=1}^{n-1} p_i(z_i | \mathbf{z}_{pa(i)}) = \int_{\mathbb{R}} \prod_{i=1}^d q_i(\varphi_i(z_i) | \varphi_{pa(i)}(\mathbf{z}_{pa(i)})) |\varphi'_i(z_i)| dz_d \tag{13}$$

(14)

Fix  $z_{[d-1]}$ , change of variable  $u = \varphi_d(z_d)$ ,  $du = \varphi'_d(z_d)$ ,

$$\prod_{i=1}^{d-1} p_i(z_i | \mathbf{z}_{\text{pa}(i)}) = \prod_{i=1}^{d-1} q_i(\varphi_i(z_i) | \varphi_{\text{pa}(i)}(\mathbf{z}_{\text{pa}(i)})) |\varphi'_i(z_i)| \quad (15)$$

Cancel the two sides of equation (12) by (15),

$$p_i(z_d | \mathbf{z}_{\text{pa}(d)}) = q_i(\varphi_d(z_d) | \varphi_{\text{pa}(d)}(\mathbf{z}_{\text{pa}(d)})) |\varphi'_d(z_d)|$$

Suppose the property is true for  $i+1, \dots, d$ . Then for  $i$ ,  $i$  is a leaf node in the first  $i$  nodes. We use the same proof as before, marginalize over  $z_i$  (same as (13) with  $d$  replaced by  $i$ ) on the joint distribution of  $i$  first variables (same as (12) with  $d$  replaced by  $i$ ), which is then divided by the obtained  $i-1$  marginal equation (same as (15) with  $d-1$  replaced by  $i-1$ ).  $\square$

## B.2 Proof of Thm. 4.2

**Theorem 4.2.** *For CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$ ,*

- (i) *Suppose for each node in  $[d]$ , there is one (perfect or imperfect) stochastic intervention such that Asm. 4.1 is verified. Then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to*

$$\mathcal{S}_{\overline{G}} = \left\{ \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{h}(\mathbf{z}) = (h_i(\mathbf{z}_{\text{anc}(i)}))_{i \in [d]}, \mathbf{h} \text{ is } C^1\text{-diffeomorphism} \right\} \quad (5)$$

- (ii) *Suppose for each node  $i$  in  $[d]$ , there is one perfect stochastic intervention such that Asm. 4.1 is verified, then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $\mathcal{S}_{\text{scaling}}$ .*

*Proof. Proof of (i):* Consider two latent CBNs achieving the same likelihood across all interventional regimes:  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d \rrbracket})$  and  $(G, \mathbf{f}', (\mathbb{Q}^i, \tau_i)_{i \in \llbracket 0, d \rrbracket})$ . Since the intervention targets  $(\tau_i)_{i \in \llbracket 0, d \rrbracket}$  are the same on both latent CBN, by rearranging the indices of  $G$  and correspondingly the indices in  $\mathbb{P}^i$ ,  $\mathbb{Q}^i$  and  $(\tau_i)_{i \in \llbracket 0, d \rrbracket}$ , we can suppose without loss of generality that the index of  $G$  preserves the partial order induced by  $E(G)$ :  $i < j$  if  $(i, j) \in E(G)$ . Since  $(\tau_i)_{i \in [d]}$  covers all  $d$  nodes in  $G$ , by rearranging  $(\tau_i)_{i \in [d]}$  we can suppose without loss of generality that  $\tau_i = i \forall i \in [d]$ .

In the  $i$ -th interventional regime,

$$\begin{aligned} \mathbb{P}^i(\mathbf{Z}) &= \widetilde{\mathbb{P}}_i(Z_i | \mathbf{z}_{\text{pa}^i(i)}) \prod_{j \in [d] \setminus i} \mathbb{P}(Z_j | \mathbf{z}_{\text{pa}(j)}) \\ \mathbb{Q}^i(\mathbf{Z}) &= \widetilde{\mathbb{Q}}_i(Z_i | \mathbf{z}_{\text{pa}^i(i)}) \prod_{j \in [d] \setminus i} \mathbb{Q}_j(Z_j | \mathbf{z}_{\text{pa}(j)}), \end{aligned}$$

where  $\text{pa}(j) = \text{pa}^i(j) \forall j \neq i$ , since intervening on  $i$  does not change the arrows towards  $j$ .

Define  $\varphi := \mathbf{f}'^{-1} \circ \mathbf{f}$ . Denote its  $i$ -th dimension output function as  $\varphi_i : \mathbb{R}^d \rightarrow \mathbb{R}$ . We will prove by induction that  $\forall i \in [d], \forall j \notin \text{anc}(i), \forall \mathbf{z} \in \mathbb{R}^d, \frac{\partial \varphi_i}{\partial z_j}(\mathbf{z}) = 0$ .

For any  $i \in \llbracket 0, d \rrbracket$ ,  $\mathbf{f}_* \mathbb{P}^i = \mathbf{f}'_* \mathbb{Q}^i$ . Since  $\varphi$  is a diffeomorphism, by the change of variable formula,

$$p^i(\mathbf{z}) = q^i(\varphi(\mathbf{z})) |\det D\varphi(\mathbf{z})| \quad (16)$$

For  $i = 0$ , factorize  $p^i$  and  $q^i$  according to  $G$ , then take the logarithm on both sides:

$$\sum_{j=1}^d \ln p_j(z_j | \mathbf{z}_{\text{pa}(j)}) = \sum_{j=1}^d \ln q_j(\varphi_j(\mathbf{z}) | \varphi_{\text{pa}(j)}(\mathbf{z})) + \ln |\det D\varphi(\mathbf{z})| \quad (17)$$

For  $i = 1$ ,  $\widetilde{\mathbb{Q}}_1$  has no conditionals, thus  $q^i$  is factorized as

$$q^1(\mathbf{z}) = \widetilde{q}_1(z_1) \prod_{j=2}^d q_j(z_j | \mathbf{z}_{\text{pa}(j)})$$

So the equation (16) for  $i = 1$  after taking logarithm is

$$\begin{aligned} \ln \tilde{p}_1(z_1) + \sum_{j=2}^d \ln p_j(z_j | \mathbf{z}_{\text{pa}(j)}) &= \ln \tilde{q}_1(\varphi_1(\mathbf{z})) + \sum_{j=2}^d q_j(\varphi_j(\mathbf{z}) | \varphi_{\text{pa}(j)}(\mathbf{z})) \\ &\quad + \ln |\det D\varphi(\mathbf{z})| \end{aligned} \quad (18)$$

subtract (18) by (17),

$$\ln \tilde{p}_1(z_1) - \ln p_1(z_1) = \ln \tilde{q}_1(\varphi_1(\mathbf{z})) - \ln q_1(\varphi_1(\mathbf{z})) \quad (19)$$

For any  $i \neq 1$ , take the  $i$ -th partial derivative of both sides:

$$0 = \left[ \frac{\tilde{q}'_1(\varphi_1(\mathbf{z}))}{\tilde{q}_1(\varphi_1(\mathbf{z}))} - \frac{q'_1(\varphi_1(\mathbf{z}))}{q_1(\varphi_1(\mathbf{z}))} \right] \frac{\partial \varphi_1}{\partial z_i}(\mathbf{z})$$

By Asm. 4.1, the term in the parenthesis is non-zero a.e. in  $\mathbb{R}^d$ . Thus  $\frac{\partial \varphi_1}{\partial z_i}(\mathbf{z}) = 0$  a.e. in  $\mathbb{R}^d$ . Since  $\varphi = f'^{-1} \circ f$  where  $f, f'$  are  $C^1$ -diffeomorphisms, so is  $\varphi$ .  $\frac{\partial \varphi_1}{\partial z_i}$  is continuous and thus equals zero everywhere.

Now suppose  $\forall k \in [i-1], \forall j \notin \overline{\text{anc}}(k), \forall \mathbf{z} \in \mathbb{R}^d, \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z}) = 0$ . Then for interventional regime  $i$ ,

$$\begin{aligned} \ln \tilde{p}_i(z_i | \mathbf{z}_{\text{pa}^i(i)}) + \sum_{j \neq i} \ln p_j(z_j | \mathbf{z}_{\text{pa}(j)}) &= \ln \tilde{q}_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z})) \\ &\quad + \sum_{j \neq i} q_j(\varphi_j(\mathbf{z}) | \varphi_{\text{pa}(j)}(\mathbf{z})) + \ln |\det D\varphi(\mathbf{z})| \end{aligned}$$

Subtracted by (17),

$$\ln \tilde{p}_i(z_i | \mathbf{z}_{\text{pa}^i(i)}) - \ln p_i(z_i | \mathbf{z}_{\text{pa}(i)}) = \ln \tilde{q}_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z})) - \ln q_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \quad (20)$$

For any  $\forall j \notin \overline{\text{anc}}(i), j \notin \text{pa}(i) \supset \text{pa}^i(i)$  by assumption. Take partial derivative over  $z_j$ :

$$0 = \frac{\sum_{k \in \overline{\text{pa}}^i(i)} \frac{\partial \tilde{q}_i}{\partial x_k}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{\tilde{q}_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z}))} - \frac{\sum_{k \in \overline{\text{pa}}(i)} \frac{\partial q_i}{\partial x_k}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{q_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))}$$

where  $x_k$  denotes the  $k$ -th dimension of the domain of  $\tilde{q}_i$  and  $q_i$ .

For all  $k \in \text{pa}(i)$ , since  $j \notin \overline{\text{anc}}(i)$ ,  $j$  is not in  $\overline{\text{anc}}(k)$  either. By the assumption of induction,  $\frac{\partial \varphi_k}{\partial z_j}(\mathbf{z}) = 0$ . Delete the partial derivatives that are zero:

$$0 = \left[ \frac{\frac{\partial \tilde{q}_i}{\partial x_i}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z}))}{\tilde{q}_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z}))} - \frac{\frac{\partial q_i}{\partial x_i}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))}{q_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))} \right] \frac{\partial \varphi_i}{\partial z_j}(\mathbf{z})$$

By Asm. 4.1, the term in parenthesis is nonzero a.e., thus  $\frac{\partial \varphi_i}{\partial z_j}(\mathbf{z}) = 0$  a.e. Since  $\frac{\partial \varphi_i}{\partial z_j}$  is continuous, it equals zero everywhere.

The induction is finished when  $i = d$ . We have proven that  $\forall i \in [d], \forall j \notin \overline{\text{anc}}(i), \forall \mathbf{z} \in \mathbb{R}^d, \frac{\partial \varphi_i}{\partial z_j}(\mathbf{z}) = 0$ . Namely,  $\varphi_i$  only depends on  $\mathbf{z}_{\overline{\text{anc}}(i)}$ .

**Proof of (ii):**

By the result proved in (i),  $\varphi := f'^{-1} \circ f \in \mathcal{S}_{\overline{G}}$ . Thus  $D\varphi(\mathbf{z})$  is lower triangular for all  $\mathbf{z} \in \mathbb{R}^d$ . Thus  $|\det D\varphi(\mathbf{z})| = \prod_{i=1}^d \left| \frac{\partial \varphi_i}{\partial z_i}(z_{[i]}) \right|$ , and for all  $i$ ,  $\varphi_i$  only depends on  $z_1, \dots, z_i$ . We will prove that  $\varphi_i$  only depends on  $z_i$ , i.e., it is constant on other variables.

To prove the conclusion in this item, we need the following lemma:

**Lemma B.4.** Given any  $\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n$  diffeomorphism such that for all  $i \in [n - 1]$ ,  $\frac{\partial \varphi_i}{\partial z_n}$  is zero everywhere, and given any two distributions  $\mathbb{P}, \mathbb{Q}$  that are absolutely continuous and have full support in  $\mathbb{R}^n$  such that  $\varphi_* \mathbb{P} = \mathbb{Q}$ , then the distributions of the first  $n - 1$  coordinates are preserved, i.e.,

$$(\varphi_{[n-1]})_* \mathbb{P}_{[n-1]}(\mathbf{z}_{[n-1]}) = \mathbb{Q}_{[n-1]}(\mathbf{z}_{[n-1]})$$

*Proof.* Fix  $\mathbf{z}_{[n-1]} \in \mathbb{R}^{n-1}$ . For all  $i \in [n - 1]$ ,  $\frac{\partial \varphi_i}{\partial z_n}$  is zero everywhere, and  $\varphi$  is a diffeomorphism, so  $\frac{\partial \varphi_n}{\partial z_n}$  is nonzero everywhere, otherwise there will exist  $\mathbf{z}$  such that  $\frac{\partial \varphi_n}{\partial z_n}(\mathbf{z})$  is singular. Therefore  $\frac{\partial \varphi_n}{\partial z_n}(\mathbf{z}_{[n-1]}, \cdot)$  is continuous and nonzero, and thus  $\varphi_n(\mathbf{z}_{[n-1]}, \cdot)$  is a diffeomorphism  $\mathbb{R} \rightarrow \mathbb{R}$ . So we can apply the change of variable  $u = \varphi_n(\mathbf{z}_{[n-1]}, z_n)$ ,  $du = \left| \frac{\partial \varphi_n}{\partial z_n}(\mathbf{z}_{[n-1]}, z_n) \right| dz_n$ .

$$\begin{aligned} \int_{\mathbb{R}} q(\varphi_{[n-1]}(\mathbf{z}_{[n-1]}), \varphi_n(z_n)) \left| \frac{\partial \varphi_n}{\partial z_n}(\mathbf{z}_{[n-1]}, z_n) \right| dz_n &= \int_{\mathbb{R}} q(\varphi_{[n-1]}(\mathbf{z}_{[n-1]}), u) du \\ &= q_{[n-1]}(\varphi_{[n-1]}(\mathbf{z}_{[n-1]})) \end{aligned} \quad (21)$$

In the equation  $p(\mathbf{z}) = q(\varphi(\mathbf{z})) |\det \varphi(\mathbf{z})|$ , marginalize over  $z_n$ :

$$\int_{\mathbb{R}} p(\mathbf{z}_{[n-1]}, z_n) dz_n = \int_{\mathbb{R}} q(\varphi_{[n-1]}(\mathbf{z}_{[n-1]}), \varphi_n(z_n)) \prod_{i=1}^n \left| \frac{\partial \varphi_i}{\partial z_i}(z_{[i]}) \right| dz_n \quad (22)$$

Using (21), we obtain

$$\begin{aligned} p_{[n-1]}(\mathbf{z}_{[n-1]}) &= q_{[n-1]}(\varphi_{[n-1]}(\mathbf{z}_{[n-1]})) \prod_{i=1}^{n-1} \left| \frac{\partial \varphi_i}{\partial z_i}(z_{[i]}) \right| \\ &= q_{[n-1]}(\varphi_{[n-1]}(\mathbf{z}_{[n-1]})) |\det D\varphi_{[n-1]}(\mathbf{z}_{[n-1]})| \end{aligned}$$

which is the density equation for push-forward measures that we want to prove.  $\square$

Back to the proof of (ii). For all  $i \in [d]$ ,  $(\varphi_1, \dots, \varphi_{i-1})$  is a diffeomorphism because  $|\det D\varphi_{[i-1]}(\mathbf{z}_{[i-1]})| = \prod_{j=1}^{i-1} \left| \frac{\partial \varphi_j}{\partial z_j}(\mathbf{z}) \right| \neq 0$ .

We will prove by induction on the reverse order of  $[d]$  that the  $i$ -th row off-diagonal entries of  $D\varphi(\mathbf{z})$  are zero for all  $\mathbf{z} \in \mathbb{R}^d$ .

In the interventional regime  $d$ , by the assumption on the indices of  $V(G) = [d]$  in the proof (i), the node  $d$  is not a parent of any node in  $[d - 1]$ . Thus the perfect stochastic intervention on  $z_d$  leads to the density  $p^d$  and  $q^d$  factorized as follows:

$$p_{[d-1]}(\mathbf{z}_{[d-1]}) \tilde{p}_d(z_d) = q_{[d-1]}(\varphi_{[d-1]}(\mathbf{z}_{[d-1]})) \tilde{q}_d(\varphi_d(\mathbf{z})) |\det D\varphi_{[d-1]}(\mathbf{z}_{[d-1]})| \left| \frac{\partial \varphi_d}{\partial z_d}(\mathbf{z}) \right|$$

Since  $D\varphi$  is lower triangular everywhere, cancel the terms of coordinate  $[d - 1]$  on both sides by Lemma B.4,

$$\tilde{p}_d(z_d) = \tilde{q}_d(\varphi_d(\mathbf{z}_{[d-1]}, z_d)) \left| \frac{\partial \varphi_d}{\partial z_d}(\mathbf{z}_{[d-1]}, z_d) \right| \quad (23)$$

which is equivalent to

$$\forall \mathbf{z}_{[d-1]} \in \mathbb{R}^{d-1}, \quad \tilde{\mathbb{Q}}_d = \varphi_d(\mathbf{z}_{[d-1]}, \cdot)_* \tilde{\mathbb{P}}_d \quad (24)$$

By Lemma B.1,  $\varphi_d$  is constant in the first  $d - 1$  variables.

Suppose the off-diagonal entries are zero for the  $i, i + 1, \dots, d$  rows of  $D\varphi(\mathbf{z})$ .

By the assumption on the indices of  $V(G)$  in the proof (i), the node  $i$  is not a parent of any node in  $[i - 1]$ . Thus the perfect stochastic intervention on  $z_i$  leads to the density  $p^i$  and  $q^i$  factorized as follows:

$$\begin{aligned} \int_{\mathbb{R}^{d-i}} p(\mathbf{z}) dz_{i+1} \cdots dz_d &= \int_{\mathbb{R}^{d-i}} q(\varphi(\mathbf{z})) \prod_{j=1}^d \left| \frac{\partial \varphi_j}{\partial z_j} (\mathbf{z}_{[j]}) \right| dz_{i+1} \cdots dz_d \\ &= \prod_{j=1}^i \left| \frac{\partial \varphi_j}{\partial z_j} (\mathbf{z}_{[j]}) \right| \int_{\mathbb{R}^{d-i}} q(\varphi_{[i]}(\mathbf{z}_{[i]}), \varphi_{i+1}(z_{i+1}), \dots, \varphi_d(z_d)) \prod_{k=i+1}^d \left| \frac{\partial \varphi_k}{\partial z_k} (z_k) \right| dz_{i+1} \cdots dz_d \end{aligned} \quad (25)$$

By a change of variables  $\begin{cases} u_{i+1} = \varphi_{i+1}(z_{i+1}) \\ \vdots \\ u_d = \varphi_d(z_d) \end{cases}$ , we get

$$\begin{aligned} p_{[i]}(\mathbf{z}_{[i]}) &= \prod_{j=1}^i \left| \frac{\partial \varphi_j}{\partial z_j} (\mathbf{z}_{[j]}) \right| \int_{\mathbb{R}^{d-i}} q(\varphi_{[i]}(\mathbf{z}_{[i]}), u_{i+1}, \dots, u_d) du_{i+1} \cdots du_d \\ &= q_{[i]}(\varphi_{[i]}(\mathbf{z}_{[i]})) |\det D\varphi_{[i]}(\mathbf{z}_{[i]})| \\ &= q_{[i-1]}(\varphi_{[i-1]}(\mathbf{z}_{[i-1]})) \tilde{q}_i(\varphi_i(\mathbf{z})) |\det D\varphi_{[i-1]}(\mathbf{z}_{[i-1]})| \left| \frac{\partial \varphi_i}{\partial z_i} (\mathbf{z}) \right| \end{aligned}$$

By Lemma B.4,  $p_{[i-1]}(\mathbf{z}_{[i-1]}) = q_{[i-1]}(\varphi_{[i-1]}(\mathbf{z}_{[i-1]})) |\det D\varphi_{[i-1]}(\mathbf{z}_{[i-1]})|$ . By Lemma B.1,  $\varphi_{[i]}$  is constant in the first  $i - 1$  variables.

In addition,  $D\varphi(\mathbf{z})$  is lower triangular for all  $z$ , so we have proven that  $\varphi \in \mathcal{S}_{\text{scaling}}$ .

□

### B.3 Proof of Prop. 4.3

**Proposition 4.3.** *Given a DAG  $G$ , with  $d - 1$  perfect stochastic single node interventions on distinct targets, if the remaining unintervened node has any parent in  $G$ ,  $(G, \mathcal{F}, P_G)$  is not identifiable up to*

$$\mathcal{S}_{\text{reparam}} := \left\{ \mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{g} = \mathbf{P} \circ \mathbf{h}, \mathbf{P} \text{ is a permutation matrix, } \mathbf{h} \in \mathcal{S}_{\text{scaling}} \right\}. \quad (6)$$

*Proof.* Without loss of generality by rearranging  $(\mathbb{P}^i, \tau_i)_{i \in [0, d-1]}$ , suppose that the unintervened variable is the node  $d$ . Fix any  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in [0, d-1]})$ , s.t.  $d - 1$  is a parent of  $d$ . Assume that the causal mechanism of  $Z_d$  only has one conditional variable  $Z_{d-1}$ . Let  $Z_d \sim \mathcal{N}(Z_{d-1}, 1)$ , Namely,

$$p_d(z_d \mid z_{d-1}) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(z_d - z_{d-1})^2}{2} \right)$$

We now construct  $(G, \mathbf{f}', (\mathbb{Q}^i, \tau_i)_{i \in [0, d-1]})$  such that  $\mathbf{f}_* \mathbb{P}^i = \mathbf{f}'_* \mathbb{Q}^i$  and  $\mathbf{f}'^{-1} \circ \mathbf{f} \notin \mathcal{S}_{\text{reparam}}$ .

Set  $\mathbb{Q}_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)}) := \mathbb{P}_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)})$ ,  $\widetilde{\mathbb{Q}}_i(Z_i) := \widetilde{\mathbb{P}}_i(Z_i \mid \mathbf{Z}_{\text{pa}(i)}) \quad \forall i \in [d-1]$ .

Set  $\mathbb{Q}_d(Z_d \mid Z_{d-1}) := \mathcal{N}(-Z_{d-1}, 1)$ ,  $\varphi(\mathbf{z}) := (z_1, \dots, -z_{d-1}, z_d - 2z_{d-1})$ , thus  $|\text{Det } D\varphi(\mathbf{z})| = 1 \quad \forall \mathbf{z} \in \mathbb{R}^d$ .

$$\begin{aligned} q_d(\varphi_d(\mathbf{z}) \mid \varphi_{d-1}(\mathbf{z})) &= \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(\varphi_d(\mathbf{z}) - \varphi_{d-1}(\mathbf{z}))^2}{2} \right) \\ &= \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{(z_d - 2z_{d-1} + z_{d-1})^2}{2} \right) \\ &= p_d(z_d \mid z_{d-1}) \end{aligned}$$

From the above equation, we infer that for the unintervened regime,

$$\prod_{j=1}^d p_j(z_j \mid \mathbf{z}_{\text{pa}(j)}) = \left[ \prod_{j=1}^d q_j(\varphi_j(\mathbf{z}) \mid \varphi_{\text{pa}(j)}(\mathbf{z})) \right] |\det D\varphi(\mathbf{z})|$$

and for the  $d - 1$  interventional regimes,

$$\forall i \in [d-1], \tilde{p}_i(z_i) \prod_{j \neq i} p_j(z_j \mid \mathbf{z}_{\text{pa}(j)}) = \tilde{q}_i(\varphi_i(z)) \left[ \prod_{j \neq i} q_j(\varphi_j(z) \mid \varphi_{\text{pa}(j)}(z)) \right] |\det D\varphi(z)|$$

i.e.,

$$\begin{aligned} \varphi_* \mathbb{P}^i &= \mathbb{Q}^i \quad \forall i \in \llbracket 0, d-1 \rrbracket. \\ \mathbf{f}_* \mathbb{P}^i &= (\mathbf{f} \circ \varphi^{-1})_* \mathbb{Q}^i \end{aligned}$$

However,  $(\mathbf{f} \circ \varphi^{-1})^{-1} \circ \mathbf{f} = \varphi \notin S_{\text{reparam}}$ .  $\square$

#### B.4 Proof of Prop. 4.4

**Proposition 4.4.** Suppose that  $G$  is the empty graph, and that there are  $d - 1$  variables intervened on, with one single target per dataset, such that Asm. 4.1 is satisfied. Then CauCA (in this case, ICA) in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $S_{\text{scaling}}$  defined as in eq. (3).

*Proof.* Without loss of generality by rearranging  $(\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d-1 \rrbracket}$ , suppose that the unintervened variable is the node  $d$ . We apply the induction in the proof of Thm. 4.2 (i). Since there are  $d - 1$  interventions, the induction stops at  $d - 1$ , and we can infer that for  $\varphi := \mathbf{f}'^{-1} \circ \mathbf{f}$ , for all  $i \in [d-1]$ ,  $\frac{\partial \varphi_i}{\partial z_j}(\mathbf{z}) = 0$  a.e.  $\forall j \neq i$ .

Similar to Thm. 4.2 (i), since  $\frac{\partial \varphi_i}{\partial z_j}$  is continuous, it equals zero everywhere. Thus  $D\varphi(z)$  is lower triangular for all  $z \in \mathbb{R}^d$ .

By Lemma B.4,  $(\varphi_{[d-1]})_* \mathbb{P}_{[d-1]}(\mathbf{Z}_{[d-1]}) = \mathbb{Q}_{[d-1]}(\mathbf{Z}_{[d-1]})$ . Namely,

$$\prod_{j=1}^{d-1} p_j(z_j) = \prod_{j=1}^{d-1} q_j(\varphi_j(\mathbf{z})) |\det D\varphi_{[d-1]}(\mathbf{z}_{[d-1]})| \quad (26)$$

Since for all  $i \in [d-1]$ ,  $\forall j \neq i$ ,  $\frac{\partial \varphi_i}{\partial z_j}(\mathbf{z}) = 0$ ,  $D\varphi(\mathbf{z})$  is lower triangular for all  $z \in \mathbb{R}^d$ . Thus  $|\det D\varphi(\mathbf{z})| = \prod_{j=1}^d |\partial_j \varphi_j(\mathbf{z})|$ . Moreover, in the unintervened dataset,

$$\prod_{j=1}^d p_j(z_j) = \prod_{j=1}^d q_j(\varphi_j(\mathbf{z})) |\partial_j \varphi_j(\mathbf{z})| \quad (27)$$

Divide (27) by (26),

$$p_d(z_d) = q_d(\varphi_d(\mathbf{z})) \left| \frac{\partial \varphi_d}{\partial z_d}(\mathbf{z}_{[d-1]}, z_d) \right|$$

which is equivalent to

$$\forall \mathbf{z}_{[d-1]} \in \mathbb{R}^{d-1}, \quad \widetilde{\mathbb{Q}}_d = \varphi_d(\mathbf{z}_{[d-1]}, \cdot)_* \widetilde{\mathbb{P}}_d. \quad (28)$$

By Lemma B.1,  $\varphi_d$  is constant in the first  $d - 1$  variables. We have proven that  $\varphi \in S_{\text{scaling}}$ .  $\square$

#### B.5 Proof of Prop. 4.5

**Proposition 4.5.** Given an empty graph  $G$ , with  $d - 2$  single-node interventions on distinct targets, with one single target per dataset, such that Asm. 4.1 is satisfied. Then CauCA (in this case, ICA) in  $(G, \mathcal{F}, \mathcal{P}_G)$  is not identifiable up to  $S_{\text{reparam}}$ .

*Proof.* Without loss of generality by rearranging  $(\mathbb{P}^i, \tau_i)_{i \in [0, d-2]}$ , suppose that the two unintervened variables are the nodes  $d-1, d$ . Fix any  $\mathbf{f}$  and  $(\mathbb{P}_i, \tilde{\mathbb{P}}_i)_{i \in [0, d-2]}$  such that for all  $i \in [d-2]$ ,  $\mathbb{P}_i, \tilde{\mathbb{P}}_i$  have any distribution that is absolutely continuous and full support in  $\mathbb{R}$  with a differentiable density, and such that Asm. 4.1 is satisfied. We will prove that whether we suppose independent Gaussian distributions are in the class of latent distributions or not, CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is not identifiable up to  $S_{\text{reparam}}$ .

**Case 1: Independent Gaussian distributions are in  $\mathcal{P}_G$ .**

By the famous result in linear ICA, if  $\mathbb{P}_{d-1}, \mathbb{P}_d$  form an isotropic Gaussian vector  $\mathcal{N}(\mathbf{0}, \Sigma)$ , i.e.,  $\Sigma$  is diagonal with the same variances on each dimension, then any rotation form a spurious solution. Namely, let  $\varphi(\mathbf{z}) = (z_1, \dots, z_{d-2}, z_{d-1} \cos(\theta) - z_d \sin(\theta), z_d \cos(\theta) + z_{d-1} \sin(\theta))$ ,  $\theta \neq k\pi$ , then

$$\begin{aligned} \forall i \in [0, d-2] \quad & \varphi_* \mathbb{P}^i = \mathbb{P}^i \\ & \mathbf{f}_* \mathbb{P}^i = (\mathbf{f} \circ \varphi^{-1})_* \mathbb{P}^i \end{aligned}$$

However,  $(\mathbf{f} \circ \varphi^{-1})^{-1} \circ \mathbf{f} = \varphi \notin S_{\text{reparam}}$ .

**Case 2: Independent Gaussian distributions are not in  $\mathcal{P}_G$ .**

Suppose that for all  $i \in \{d-1, d\}$ ,  $\mathbb{P}_i$  has the same density  $p_a$ :

$$p_a(z) = \begin{cases} \exp(-az^2) & z < 0 \\ 1 & 0 \leq z \leq 1 - \sqrt{\frac{\pi}{a}} \\ \exp\left(-a(z - (1 - \sqrt{\frac{\pi}{a}}))^2\right) & z > 1 - \sqrt{\frac{\pi}{a}} \end{cases}$$

where  $\sqrt{\frac{\pi}{a}} < 1$ . One can verify that  $p_a$  is a smooth p.d.f.

We construct a measure-preserving automorphism inspired by [Hyvärinen and Pajunen \(1999\)](#).

$$\varphi(\mathbf{Z}) = \begin{cases} \mathbf{Z} & ||\mathbf{Z}_{[d-1, d]}|| \geq R \\ \begin{pmatrix} \cos(\alpha(||\mathbf{Z}_{[d-1, d]} - \mathbf{C}|| - R)) Z_{d-1} \\ -\sin(\alpha(||\mathbf{Z}_{[d-1, d]} - \mathbf{C}|| - R)) Z_d \\ \cos(\alpha(||\mathbf{Z}_{[d-1, d]} - \mathbf{C}|| - R)) Z_d \\ +\sin(\alpha(||\mathbf{Z}_{[d-1, d]} - \mathbf{C}|| - R)) Z_{d-1} \end{pmatrix} & ||\mathbf{Z}_{[d-1, d]}|| < R \end{cases}$$

where  $\alpha \neq 0$ ,  $\mathbf{C} = (\frac{1}{2}(1 - \sqrt{\frac{\pi}{a}}), \frac{1}{2}(1 - \sqrt{\frac{\pi}{a}}))$ ,  $R \in (0, \frac{1}{2}(1 - \sqrt{\frac{\pi}{a}})]$ .

Now let us prove that  $\varphi$  preserves  $\mathbb{P}_{[d-1, d]}$ . By shifting the center of  $p_a$  to the origin, we only need to prove that the shifted  $\varphi_{[d-1, d]}$  preserves the uniform distribution over  $[-R, R]^2$ . One can verify that  $\varphi_{[d-1, d]}$  is a diffeomorphism over the 2-dimensional open disk  $D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\} \rightarrow D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\}$  and  $|\det(D\varphi_{[d-1, d]}(\mathbf{z}))| = 1$ . Thus  $p_a(\mathbf{z}) = p_a(\varphi_{[d-1, d]}(\mathbf{z})) |\det(D\varphi_{[d-1, d]}(\mathbf{z}))| \forall \mathbf{z} \in D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\}$ . Since  $\varphi = Id$  outside of the disk, this change of variables formula holds almost everywhere in  $\mathbb{R}^2$ , thus  $\mathbb{P}_{[d-1, d]} = (\varphi_{[d-1, d]})_* \mathbb{P}_{[d-1, d]}$ , namely,  $\varphi_{[d-1, d]}$  preserves  $\mathbb{P}_{[d-1, d]}$  on  $\mathbb{R}^2$ .

Moreover, since  $\varphi_{[d-2]}$  is identity,  $\tilde{\mathbb{P}}_i = (\varphi_i)_* \tilde{\mathbb{P}}_i$  and  $\mathbb{P}_i = (\varphi_i)_* \mathbb{P}_i$  for all  $i \in [d-2]$ . Thus

$$\begin{aligned} \forall i \in [0, d-2] \quad & \varphi_* \mathbb{P}^i = \mathbb{P}^i \\ & \mathbf{f}_* \mathbb{P}^i = (\mathbf{f} \circ \varphi^{-1})_* \mathbb{P}^i \end{aligned}$$

However,  $(\mathbf{f} \circ \varphi^{-1})^{-1} \circ \mathbf{f} = \varphi \notin S_{\text{reparam}}$ . □

## B.6 Further constraint on the indeterminacy set

**Corollary B.5.** *Based on the assumption of (2) of Thm. 4.2, if in every dataset we are given the set of possible intervention mechanisms:  $\mathcal{M} = (\mathcal{M}_i)_{i \in [d]}$ , then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $\mathcal{S}_{\mathcal{M}} := \{\varphi : \mathbb{R}^n \rightarrow \mathbb{R}^n | \varphi \in \mathcal{S}_{\text{scaling}}, \forall i \in [d], \varphi_i \in \mathcal{S}_{\mathcal{M}_i}\}$  where  $\mathcal{S}_{\mathcal{M}_i} := \{\bar{F}_{\mathbb{M}'_i}^{-1} \circ F_{\mathbb{M}_i} | \mathbb{M}_i, \mathbb{M}'_i \in \mathcal{M}_i\}$*

*In particular, if  $\mathcal{M}_i$  is singleton for all  $i$ , then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to  $\mathcal{S}_{\text{reflexion}} := \{\mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d | \forall i \in [d], h_i = \text{Id or } -\text{Id}\}$ .*

*Proof.* Based on the conclusion of (2) of Thm. 4.2, for any  $i \in [d]$ , choose any  $\mathbb{M}_i, \mathbb{M}'_i$  in  $\mathcal{M}_i$ ,

$$(\varphi_i)_* \mathbb{M}_i = \mathbb{M}'_i$$

By Lemma B.2, the only possible  $\varphi_i$  are  $F_{\mathbb{M}'_i}^{-1} \circ F_{\mathbb{M}_i}$  and  $\bar{F}_{\mathbb{M}'_i}^{-1} \circ F_{\mathbb{M}_i}$ . Thus  $\varphi_i \in \mathcal{S}_{\mathcal{M}_i}$ . In particular if  $\mathcal{M}_i$  is a singleton  $\{\mathbb{M}_i\}$ , then  $F_{\mathbb{M}_i}^{-1} \circ F_{\mathbb{M}_i} = \text{Id}$ , and  $\bar{F}_{\mathbb{M}_i}^{-1} \circ F_{\mathbb{M}_i} = -\text{Id}$ .  $\square$

## B.7 Proof of Thm. 4.6

**Theorem 4.6.** *Suppose  $G$  is the empty graph. Suppose that our datasets encompass interventions over all variables in the latent graph, i.e.,  $\bigcup_{k \in [K]} \tau_k = [d]$ . Suppose for every  $k$ , the targets of interventions are a strict subset of all variables, i.e.,  $|\tau_k| = n_k, n_k \in [d-1]$ .*

(Block-interventional discrepancy) *Suppose that there are  $n_k$  interventions with target  $\tau_k$  such that  $\mathbf{v}_k(\mathbf{z}_{\tau_k}, 1) - \mathbf{v}_k(\mathbf{z}_{\tau_k}, 0), \dots, \mathbf{v}_k(\mathbf{z}_{\tau_k}, n_k) - \mathbf{v}_k(\mathbf{z}_{\tau_k}, 0)$  are linearly independent, where*

$$\mathbf{v}_k(\mathbf{z}_{\tau_k}, s) := ((\ln q_{k,1}^s)'(z_{\tau_k,1}), \dots, (\ln q_{k,n_k}^s)'(z_{\tau_k,n_k})) \quad (7)$$

where  $q_k^s$  is the intervention of the  $s$ -th interventional regime that has the target  $\tau_k$ , and  $q_{k,j}^s$  is the  $j$ -th marginal of it.  $z_{\tau_k,j}$  is the  $j$ -th dimension of  $\mathbf{z}_{\tau_k}$ .  $s = 0$  denotes the unintervened regime. Then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is block-identifiable (following von Kriegelgen et al. (2021)): namely, if  $\mathbf{f}_* \mathbb{P}^k = \mathbf{f}'_* \mathbb{Q}^k$  then for  $\varphi := \mathbf{f}'^{-1} \circ \mathbf{f}$ , for all  $k \in [K]$ ,

$$[\varphi(\mathbf{z})]_{\tau_k} = \varphi_{\tau_k}(\mathbf{z}_{\tau_k}). \quad (8)$$

*Proof.* Fix  $k \in [K]$ . Then the equation of the  $k$ -th interventional regime is  $p^k(\mathbf{z}) = q^k(\varphi(\mathbf{z})) |\det \varphi(\mathbf{z})|$ . We write the equality of pushforward densities just as Prop. 4.4, and subtract the  $k$ -th interventional regime by the unintervened regime:

$$\sum_{j=1}^{n_k} [\ln p_{k,j}^s(z_{k,j}) - \ln p_{k,j}^0(z_{k,j})] = \sum_{j=1}^{n_k} [\ln q_{k,j}^s(\varphi_{\tau_k,j}(\mathbf{z})) - \ln q_{k,j}^0(\varphi_{\tau_k,j}(\mathbf{z}))]$$

Since  $|\tau_k| < d$ , there exists  $i \in [d] \setminus \tau_k$ . Take the partial derivative of  $z_i$ :

$$\begin{aligned} 0 &= \sum_{j=1}^{n_k} \left[ \frac{q_{k,j}^{s'}(\varphi_{\tau_k,j}(\mathbf{z}))}{q_{k,j}^s(\varphi_{\tau_k,j}(\mathbf{z}))} - \frac{q_{k,j}^{0'}(\varphi_{\tau_k,j}(\mathbf{z}))}{q_{k,j}^0(\varphi_{\tau_k,j}(\mathbf{z}))} \right] \frac{\partial \varphi_{\tau_k,j}}{\partial z_i}(\mathbf{z}) \\ &= (\mathbf{v}_k(\varphi_{\tau_k}(\mathbf{z}), s) - \mathbf{v}_k(\varphi_{\tau_k}(\mathbf{z}), 0))^\top \frac{\partial \varphi_{\tau_k}}{\partial z_i}(\mathbf{z}) \end{aligned}$$

By assumption, for all  $s \in [n_k]$ , there is one interventional regime in which the above equation holds. Those  $n_k$  equations form a linear system  $\mathbf{0} = \mathbf{M}_k(\mathbf{z}) \frac{\partial \varphi_{\tau_k}}{\partial z_i}(\mathbf{z})$ , where  $\mathbf{M}_k(\mathbf{z})$  is formed by rows  $(\mathbf{v}_k(\varphi_{\tau_k}(\mathbf{z}), s) - \mathbf{v}_k(\varphi_{\tau_k}(\mathbf{z}), 0))^\top$  for  $k \in [K]$ . Since  $\mathbf{M}_k(\mathbf{z})$  is invertible by assumption, the vector  $\frac{\partial \varphi_{\tau_k}}{\partial z_i}(\mathbf{z}) = \mathbf{0} \quad \forall \mathbf{z} \in \mathbb{R}^d$ . Such a result is valid for all  $i \in [d] \setminus \tau_k$ . Since  $\bigcup_{k \in I} \tau_k = [d]$ , all the non-diagonal entries of  $D\varphi(\mathbf{z})$  that are not in the blocks of  $\tau_k \times \tau_k$  are 0 a.e., and furthermore 0 everywhere since  $D\varphi$  is continuous by assumption. We conclude that  $\varphi_{\tau_k}$  only depends on  $\mathbf{z}_{\tau_k}$ , i.e.,  $[\varphi(\mathbf{z})]_{\tau_k} = \varphi_{\tau_k}(\mathbf{z}_{\tau_k})$ .

For all  $k \in [K]$ ,  $[d] \setminus \tau_k \neq \emptyset$ . If there exists  $\mathbf{z} \in \mathbb{R}^d$  such that  $\det(D\varphi_{\tau_k}(\mathbf{z})) = 0$ , since  $[\varphi(\mathbf{z})]_{\tau_i} = \varphi_{\tau_i}(\mathbf{z}_{\tau_i}) \forall i \in [n]$ , the vector  $\frac{\partial \varphi_{\tau_k}}{\partial z_i}(\mathbf{z}) = \mathbf{0}$  for all  $i \notin \tau_k$ . Thus the rows  $\tau_k$  of  $D\varphi(\mathbf{z})$  are linearly dependent, which implies  $\det(D\varphi(\mathbf{z})) = 0$ , which contradicts with  $\varphi$  invertible. Thus  $\varphi_{\tau_k}$  is a diffeomorphism.  $\square$

## B.8 Proof of Prop. 4.7

**Proposition 4.7.** *Under the assumptions of Thm. 4.6, suppose there exist  $k \in [K]$  and there are  $2n_k$  interventions with targets  $\tau_k$  such that for any  $\mathbf{z}_{\tau_k} \in \mathbb{R}^{n_k}$ ,  $\mathbf{w}_k(\mathbf{z}_{\tau_k}, 1) - \mathbf{w}_k(\mathbf{z}_{\tau_k}, 0), \dots, \mathbf{w}_k(\mathbf{z}_{\tau_k}, 2n_k) - \mathbf{w}_k(\mathbf{z}_{\tau_k}, 0)$  are linearly independent, where*

$$\mathbf{w}_k(\mathbf{z}_{\tau_k}, s) := \left( \left( \frac{q_{k,1}^{s'}}{q_{k,1}^s} \right)'(z_{\tau_k,1}), \dots, \left( \frac{q_{k,n_k}^{s'}}{q_{k,n_k}^s} \right)'(z_{\tau_k,n_k}), \frac{q_{k,1}^{s'}}{q_{k,1}^s}(z_{\tau_k,1}), \dots, \frac{q_{k,n_k}^{s'}}{q_{k,n_k}^s}(z_{\tau_k,n_k}) \right)$$

where  $q_k^s$  is the intervention of the  $s$ -th interventional regime that has the target  $\tau_k$ , and  $q_{k,j}^s$  is the  $j$ -th marginal of it.  $z_{\tau_k,j}$  is the  $j$ -th dimension of  $\mathbf{z}_{\tau_k}$ .  $s = 0$  denotes the unintervened regime. Then  $\varphi_{\tau_k} \in \mathcal{S}_{reparam} := \left\{ \mathbf{g} : \mathbb{R}^{n_k} \rightarrow \mathbb{R}^{n_k} \mid \mathbf{g} = \mathbf{P} \circ \mathbf{h} \text{ where } \mathbf{P} \text{ is a permutation matrix and } \mathbf{h} \in \mathcal{S}_{scaling} \right\}$

The proof is based on Theorem 1 of Hyvarinen et al. (2019).

*Proof.* Since the intervention targets are known, without loss of generality, suppose the interventions are on the first  $n_k$  variables. By the result of Thm. 4.6 we have  $p_k^s(\mathbf{z}_{\tau_k}) = q_k^s(\varphi_{\tau_k}(\mathbf{z}_{\tau_k})) |\det D\varphi_{\tau_k}(\mathbf{z}_{\tau_k})|$  where  $p_k^s$  denotes the joint distribution in the  $s$ -th interventional regime such that the intervention target is  $\tau_k$ . Factorize  $p_k^s$  and  $q_k^s$  in the change of variables formula, and take the logarithm:

$$\sum_{l=1}^{n_k} \ln p_{k,l}^s(z_l) = \sum_{l=1}^{n_k} \ln q_{k,l}^s(\varphi_l(\mathbf{z}_{\tau_k})) + \ln |\det D\varphi_{\tau_k}(\mathbf{z}_{\tau_k})| \quad (29)$$

where  $p_l^s$  is the  $l$ -th marginal of the intervention of the  $s$ -th interventional regime that has the target  $\tau_k$ .

For the unintervened regime, denote the density of the  $l$ -th marginal of  $\mathbb{P}$  as  $p_l$ . By the result of Thm. 4.6 we have

$$\sum_{l=1}^{n_k} \ln p_l^0(z_l) = \sum_{l=1}^{n_k} \ln q_l^0(\varphi_l(\mathbf{z}_{\tau_k})) + \ln |\det D\varphi_{\tau_k}(\mathbf{z}_{\tau_k})| \quad (30)$$

Subtract the equation (29) by (30):

$$\sum_{l=1}^{n_k} [\ln p_{k,l}^s(z_l) - \ln p_l^0(z_l)] = \sum_{l=1}^{n_k} [\ln q_{k,l}^s(\varphi_l(\mathbf{z}_{\tau_k})) - \ln q_l^0(\varphi_l(\mathbf{z}_{\tau_k}))]$$

For any  $j \in \tau_k = [n_k]$ , take the partial derivative over  $z_j$

$$\frac{p_{k,j}^{s'}(z_j)}{p_{k,j}^s(z_j)} - \frac{p_j^0(z_j)}{p_j^0(z_j)} = \sum_{l=1}^{n_k} \left[ \frac{q_{k,l}^{s'}(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^s(\varphi_l(\mathbf{z}_{\tau_k}))} - \frac{q_{k,l}^0(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^0(\varphi_l(\mathbf{z}_{\tau_k}))} \right] \frac{\partial \varphi_l}{\partial z_j}(\mathbf{z}_{\tau_k})$$

For any  $1 \leq k < j$ , take the partial derivative over  $z_k$ ,

$$\begin{aligned} 0 = \sum_{l=1}^{n_k} & \left[ \left( \frac{q_{k,l}^{s'}}{q_{k,l}^s} \right)'(\varphi_l(\mathbf{z}_{\tau_k})) - \left( \frac{q_{k,l}^0}{q_{k,l}^0} \right)'(\varphi_l(\mathbf{z}_{\tau_k})) \right] \frac{\partial \varphi_l}{\partial z_k}(\mathbf{z}_{\tau_k}) \frac{\partial \varphi_l}{\partial z_j}(\mathbf{z}_{\tau_k}) \\ & + \left[ \frac{q_{k,l}^{s'}(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^s(\varphi_l(\mathbf{z}_{\tau_k}))} - \frac{q_{k,l}^0(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^0(\varphi_l(\mathbf{z}_{\tau_k}))} \right] \frac{\partial^2 \varphi_l}{\partial z_k \partial z_j}(\mathbf{z}_{\tau_k}) \end{aligned}$$

For  $1 \leq k < j \leq n_k$  there are  $\frac{n_k(n_k-1)}{2}$  equations.

Define  $\mathbf{a}_l(\mathbf{z}_{\tau_k}) = \left( \frac{\partial \varphi_l}{\partial z_k}(\mathbf{z}_{\tau_k}) \frac{\partial \varphi_l}{\partial z_j}(\mathbf{z}_{\tau_k}) \right)_{1 \leq k \leq j \leq n_k}$ ,  $\mathbf{b}_l(\mathbf{z}_{\tau_k}) = \left( \frac{\partial^2 \varphi_l}{\partial z_k \partial z_j}(\mathbf{z}_{\tau_k}) \right)_{1 \leq k < j \leq n_k}$

Then the  $\frac{n_k(n_k-1)}{2}$  equations can be written as a linear system

$$0 = \sum_{l=1}^{n_k} \mathbf{a}_l(\mathbf{z}_{\tau_k}) \left[ \left( \frac{q_{k,l}^{s'}}{q_{k,l}^s} \right)' (\varphi_l(\mathbf{z}_{\tau_k})) - \left( \frac{q_{k,l}^{0'}}{q_{k,l}^0} \right)' (\varphi_l(\mathbf{z}_{\tau_k})) \right] \\ + \mathbf{b}_l(\mathbf{z}_{\tau_k}) \left[ \frac{q_{k,l}^{s'}(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^s(\varphi_l(\mathbf{z}_{\tau_k}))} - \frac{q_{k,l}^{0'}(\varphi_l(\mathbf{z}_{\tau_k}))}{q_{k,l}^0(\varphi_l(\mathbf{z}_{\tau_k}))} \right]$$

Define  $\mathbf{M}_k(\mathbf{z}_{\tau_k}) = (\mathbf{a}_1(\mathbf{z}_{\tau_k}), \dots, \mathbf{a}_{n_k}(\mathbf{z}_{\tau_k}), \mathbf{b}_1(\mathbf{z}_{\tau_k}), \dots, \mathbf{b}_{n_k}(\mathbf{z}_{\tau_k}))$ .

Collect the equations for  $s = 1, \dots, 2n_k$ ,

$$\mathbf{0} = \mathbf{M}_k(\mathbf{z}_{\tau_k})(\mathbf{w}_k(\varphi_{\tau_k}(\mathbf{z}_{\tau_k}), 1) - \mathbf{w}_k(\varphi_{\tau_k}(\mathbf{z}_{\tau_k}), 0), \dots, \mathbf{w}_k(\varphi_{\tau_k}(\mathbf{z}_{\tau_k}), 2n_k) - \mathbf{w}_k(\varphi_{\tau_k}(\mathbf{z}_{\tau_k}), 0))$$

By assumption, the matrix containing  $\mathbf{w}$  is invertible. Thus  $M_k(\mathbf{z}_{\tau_k}) = \mathbf{0}$ , which implies  $\mathbf{a}_\ell(\mathbf{z}_{\tau_k})$  are zero for all  $\mathbf{z}_{\tau_k}$ . By the same reasoning as [Hyvärinen et al. \(2019\)](#), each row in  $D\varphi_{\tau_k}(\mathbf{z}_{\tau_k})$  has only one non-zero term, and this does not change for different  $z$ , since otherwise by continuity there exists  $\mathbf{z}$  such that  $D\varphi_{\tau_k}(\mathbf{z}_{\tau_k})$  is singular, contradiction with the invertibility of  $\varphi_{\tau_k}$  ([Thm. 4.6](#)). Thus  $\forall i \in \tau_k$ ,  $\varphi_i$  is a function of one coordinate of  $\mathbf{z}_{\tau_k}$ . Since  $\varphi_{\tau_k}$  is invertible,  $\det D\varphi_{\tau_k}(\mathbf{z}_{\tau_k}) \neq 0$ , so  $\exists \sigma$  permutation of  $\tau_k$  s.t.  $\forall i \in \tau_k$ ,  $\frac{\partial \varphi_{\sigma(i)}}{\partial z_i}(\mathbf{z}_{\tau_k}) \neq 0$ . Thus  $\varphi_{\tau_k} \in \mathcal{S}_{\text{reparam}}$ .  $\square$

## C A Counterexample for Thm. 4.2 (ii) when Asm. 4.1 is violated

One trivial case of violation of Asm. 4.1 is when  $\mathbb{P}_i$  and  $\tilde{\mathbb{P}}_i$  are the same. In that case, the interventional regime  $i$  is useless, namely, it does not constrain at all the indeterminacy set. Our counterexample is in a non-trivial case of violation, where the intervened mechanisms are not deterministically related, and do not share symmetries with the causal mechanisms.

We construct a counterexample for Thm. 4.2 (ii) when Asm. 4.1 is violated. The visualization of it is in Fig. 2. The counterexample is similar to the non-Gaussian case in the proof of Prop. 4.5.

Suppose that  $d = 2$ ,  $E(G) = \emptyset$  and that for all  $i \in \{1, 2\}$ ,  $\mathbb{P}_i$  has the same density  $p_{a,b}$ :

$$p_{a,b}(z) = \begin{cases} \exp(-az^2) & z < 0 \\ 1 & 0 \leq z \leq 1 - \frac{1}{2}(\sqrt{\frac{\pi}{a}} + \sqrt{\frac{\pi}{b}}) \\ \exp(-b(z - (1 - \frac{1}{2}(\sqrt{\frac{\pi}{a}} + \sqrt{\frac{\pi}{b}}))^2)) & z > 1 - \frac{1}{2}(\sqrt{\frac{\pi}{a}} + \sqrt{\frac{\pi}{b}}) \end{cases} \quad (31)$$

where  $\sqrt{\frac{\pi}{a}} < 1$ ,  $\sqrt{\frac{\pi}{b}} < 1$ . One can verify that  $p_a, p_b$  are smooth p.d.f.

Suppose that for all  $i \in \{1, 2\}$ , the intervened mechanism  $\tilde{\mathbb{P}}_i$  has the same density  $p_{c,d}$ , defined in the same way as (31), such that  $\sqrt{\frac{\pi}{c}} < 1$ ,  $\sqrt{\frac{\pi}{d}} < 1$ , and  $c, d \notin \{a, b\}$ .

Set  $\lambda := \min_{(x,y) \in \{(a,b), (c,d)\}} \left( 1 - \frac{1}{2} \left( \sqrt{\frac{\pi}{x}} + \sqrt{\frac{\pi}{y}} \right) \right)$ , then over  $(0, \lambda)^2$  all the densities are constant, violating Asm. 4.1.

We construct a measure-preserving automorphism inspired by [Hyvärinen and Pajunen \(1999\)](#).

$$\varphi(\mathbf{z}) = \begin{cases} \mathbf{z} & \|\mathbf{z}\| \geq R \\ \left( \begin{array}{c} \cos(\alpha(\|\mathbf{z} - \mathbf{c}\| - R))z_1 - \sin(\alpha(\|\mathbf{z} - \mathbf{c}\| - R))z_2 \\ \cos(\alpha(\|\mathbf{z} - \mathbf{c}\| - R))z_2 + \sin(\alpha(\|\mathbf{z} - \mathbf{c}\| - R))z_1 \end{array} \right) & \|\mathbf{z}\| < R \end{cases}$$

where  $\mathbf{c} = (\frac{\lambda}{2}, \frac{\lambda}{2})$  denotes the center of rotation,  $R \in (0, \frac{\lambda}{2}]$  denotes the radius of the disk, and  $\alpha \neq 0$ .

Now let us prove that  $\varphi$  preserves  $\mathbb{P}^i$  for all  $i \in \llbracket 0, 2 \rrbracket$ . By shifting  $p_{a,b}$  by  $-\mathbf{c}$ , we only need to prove that the shifted  $\varphi$  preserves the uniform distribution over  $[-R, R]^2$ . One can verify that  $\varphi$  is a diffeomorphism over the 2-dimensional open disk  $D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\} \rightarrow D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\}$  and  $|\det(\varphi(\mathbf{z}))| = 1$ . Thus  $p_a(\mathbf{z}) = p_a(\varphi(\mathbf{z})) |\det(\varphi(\mathbf{z}))| \forall \mathbf{z} \in D^2(\mathbf{0}, R) \setminus \{\mathbf{0}\}$ . Since  $\varphi = Id$  outside of the disk, this change of variables formula holds almost everywhere in  $\mathbb{R}^2$ , thus  $\mathbb{P}_i = \varphi_* \widetilde{\mathbb{P}}_i$ ,  $\widetilde{\mathbb{P}}_i = \varphi_* \widetilde{\mathbb{P}}_i$ , which implies that  $\varphi_* \mathbb{P}^i = \mathbb{P}^i \forall i \in \llbracket 0, 2 \rrbracket$ .

Thus for all  $\mathbf{f} \in \mathcal{F}$ ,  $\mathbf{f}_* \mathbb{P}^i = (\mathbf{f} \circ \varphi^{-1})_* \mathbb{P}^i$ . However,  $(\mathbf{f} \circ \varphi^{-1})^{-1} \circ \mathbf{f} = \varphi$ , which is not in  $S_{\text{reparam}}$  or  $S_{\text{scaling}}$ .

*Remark C.1.* The above example can be easily generalized to any  $p_{a,b}$  such that the constants on the plateau are different between  $\mathbb{P}_i$  and  $\widetilde{\mathbb{P}}_i$  and the domains of the plateau intersects on a nonzero measure set. For  $d > 2$ , the above example can be generalized by constructing the same  $\mathbb{P}_i$  and  $\widetilde{\mathbb{P}}^i$  for  $i = 1, 2$ , and for  $i > 2$  we fix any  $\mathbb{P}_i$  and  $\widetilde{\mathbb{P}}_i$  verifying Asm. 4.1. Let  $\varphi_{\llbracket 1, 2 \rrbracket}$  be the same measure-preserving automorphism and  $\varphi_j = Id$  for  $j > 2$ .

## D Identifiability by structure-preserving stochastic interventions

In this section, we extend the result of Thm. 4.2 (i) to the case when we have access to more imperfect interventions. Here we focus on one special case of imperfect interventions, *structure-preserving interventions*, i.e., the interventions that do not change the parent set.

**Proposition D.1.** *For CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  assume the assumptions in Thm. 4.2 (i) hold. Fix any  $i \in [d]$  such that  $pa(i) \neq anc(i)$ ,  $pa(i) \neq \emptyset$ , and define  $n_i := |\overline{pa}(i)|$ . If there are  $n_i(n_i + 1)$  structure-preserving interventions on node  $i$  such that the variability assumption  $V^i$  holds, then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to*

$$\mathcal{S}_{\overline{G}_i} = \left\{ \mathbf{h} \in \mathcal{C}^1(\mathbb{R}^d) : \mathbf{h}(\mathbf{z}) = (h_j(\mathbf{z}_{\overline{anc}(j)}))_{j \in [d]} \mid h_j(\mathbf{z}_{\overline{anc}(j)}) = h_j(\mathbf{z}_{\overline{pa}(i)}) \forall j \in \overline{pa}(i) \right\}.$$

Namely, for all  $\varphi \in \mathcal{S}_{\overline{G}_i}$ , for all the nodes  $j \in \overline{pa}(i)$ , the reconstructed  $Z_j$  can at most be a mixture of variables corresponding to the nodes in the closure of parents of  $i$ , instead of the closure of ancestors of  $j$ .

The variability assumption  $V^i$  means

$$\begin{pmatrix} \mathbf{A}_i^1(\mathbf{z}) & \mathbf{B}_i^1(\mathbf{z}) \\ \vdots & \vdots \\ \mathbf{A}_i^{n_i(n_i+1)}(\mathbf{z}) & \mathbf{B}_i^{n_i(n_i+1)}(\mathbf{z}) \end{pmatrix} \in \mathbb{R}^{n_i(n_i+1) \times n_i(n_i+1)}$$

is invertible, where the symbols are defined as follows:

$$\mathbf{A}_i^t(\mathbf{z}) = \begin{pmatrix} \frac{\partial(g_i^{s_1,t} - h_i^{s_1})}{\partial x_{r_1}} (\varphi_i(\mathbf{z}) \mid \varphi_{pa(i)}(\mathbf{z})) \\ \vdots \\ \frac{\partial(g_i^{s_k,t} - h_i^{s_k})}{\partial x_{r_m}} (\varphi_i(\mathbf{z}) \mid \varphi_{pa(i)}(\mathbf{z})) \\ \vdots \\ \frac{\partial(g_i^{s_{n_i},t} - h_i^{s_{n_i}})}{\partial x_{r_{n_i}}} (\varphi_i(\mathbf{z}) \mid \varphi_{pa(i)}(\mathbf{z})) \end{pmatrix}^\top \in \mathbb{R}^{1 \times n_i^2},$$

$$\mathbf{B}_i^t(\mathbf{z}) = \begin{pmatrix} (g_i^{s_1,t} - h_i^{s_1}) (\varphi_i(\mathbf{z}) \mid \varphi_{pa(i)}(\mathbf{z})) \\ \vdots \\ (g_i^{s_{n_i},t} - h_i^{s_{n_i}}) (\varphi_i(\mathbf{z}) \mid \varphi_{pa(i)}(\mathbf{z})) \end{pmatrix}^\top \in \mathbb{R}^{1 \times n_i}$$

where  $r_k, s_k$  are the  $k$ -th variable in  $\overline{pa}(i)$ , and

$$g_i^{k,t} (z_i \mid \mathbf{z}_{pa(i)}) := \frac{\partial \tilde{q}_i^t(z_i \mid \mathbf{z}_{pa(i)})}{\partial z_k}, \quad h_i^k (z_i \mid \mathbf{z}_{pa(i)}) := \frac{\partial q_i(z_i \mid \mathbf{z}_{pa(i)})}{\partial z_k}$$

where  $\tilde{q}_i^t$  denotes the intervened mechanism in  $t$ -th interventional regime that has the interventional target  $i$ ,  $q_i$  denotes the causal mechanism on  $Z_i$ .

*Proof.* Based on the assumption of Thm. 4.2(i), reuse the proof of Thm. 4.2(i) from the equation (20):

$$\ln \tilde{p}_i^t(z_i | \mathbf{z}_{\text{pa}^i(i)}) - \ln p_i(z_i | \mathbf{z}_{\text{pa}(i)}) = \ln \tilde{q}_i^t(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}^i(i)}(\mathbf{z})) - \ln q_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))$$

where  $\tilde{p}_i^t, \tilde{q}_i^t$  denote the intervened mechanism in  $t$ -th interventional regime that has the interventional target  $i$ . Notice that by the assumption of structure-preserving interventions,  $\text{pa}^i(i) = \text{pa}(i)$ .

Thm. 4.2(i) has already concluded that  $\partial_j \varphi_i$  is constant 0 for all  $j \notin \text{anc}(i)$ . Now we are interested in  $\partial_j \varphi_i \forall j \in \text{anc}(i)$ . Take the partial derivative over  $z_j$  with  $j \in \text{anc}(i) \setminus \text{pa}(i)$  (non-empty by assumption):

$$0 = \frac{\sum_{k \in \overline{\text{pa}}(i)} \frac{\partial \tilde{q}_i^t}{\partial z_k}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{\tilde{q}_i^t(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))} - \frac{\sum_{k \in \overline{\text{pa}}(i)} \frac{\partial q_i}{\partial z_k}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{q_i(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z}))} \quad (32)$$

Recall that we define

$$g_i^{k,t}(z_i | \mathbf{z}_{\text{pa}(i)}) := \frac{\frac{\partial \tilde{q}_i^t}{\partial z_k}(z_i | \mathbf{z}_{\text{pa}(i)})}{\tilde{q}_i^t(z_i | \mathbf{z}_{\text{pa}(i)})}, \quad h_i^k(z_i | \mathbf{z}_{\text{pa}(i)}) := \frac{\frac{\partial q_i}{\partial z_k}(z_i | \mathbf{z}_{\text{pa}(i)})}{q_i(z_i | \mathbf{z}_{\text{pa}(i)})}$$

So (32) is rewritten as

$$0 = \sum_{k \in \overline{\text{pa}}(i)} \left[ g_i^{k,t}(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) - h_i^k(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \right] \partial_j \varphi_k(\mathbf{z})$$

Choose any  $l \in \text{pa}(i)$  (non-empty by assumption). Take the partial derivative of  $z_l$  on two sides:

$$0 = \sum_{k \in \overline{\text{pa}}(i)} \left[ \sum_{m \in \overline{\text{pa}}(i)} \partial_m (g_i^{k,t} - h_i^k)(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \partial_j \varphi_k(\mathbf{z}) \partial_l \varphi_m(\mathbf{z}) \right] \\ + (g_i^{k,t} - h_i^k)(\varphi_i(\mathbf{z}) | \varphi_{\text{pa}(i)}(\mathbf{z})) \partial_l \partial_j \varphi_k(\mathbf{z})$$

which can be rewritten as

$$0 = \mathbf{A}_i^t(\mathbf{z}) \mathbf{a}_{j,l}^i(\mathbf{z}) + \mathbf{B}_i^t(\mathbf{z}) \mathbf{b}_{j,l}^i(\mathbf{z}) \quad (33)$$

$$\text{where } \mathbf{a}_{j,l}^i(\mathbf{z}) = \begin{pmatrix} \partial_j \varphi_{p_1}(\mathbf{z}) \partial_l \varphi_{p_1}(\mathbf{z}) \\ \vdots \\ \partial_j \varphi_{p_k}(\mathbf{z}) \partial_l \varphi_{p_m}(\mathbf{z}) \\ \vdots \\ \partial_j \varphi_{p_{n_i}}(\mathbf{z}) \partial_l \varphi_{p_{n_i}}(\mathbf{z}) \end{pmatrix} \in \mathbb{R}^{n_i^2}, \quad \mathbf{b}_{j,l}^i(\mathbf{z}) = \begin{pmatrix} \partial_l \partial_j \varphi_{p_1}(\mathbf{z}) \\ \vdots \\ \partial_l \partial_j \varphi_{p_{n_i}}(\mathbf{z}) \end{pmatrix} \in \mathbb{R}^{n_i},$$

Collect  $n_i$  ( $n_i + 1$ ) equations, every one corresponding to one interventional regime, in the form of (33) for all  $t \in [n_i(n_i + 1)]$ :

$$\mathbf{0} = \mathbf{M}_i(\mathbf{z}) \begin{pmatrix} \mathbf{a}_{j,l}^i(\mathbf{z}) \\ \mathbf{b}_{j,l}^i(\mathbf{z}) \end{pmatrix} \quad (34)$$

where

$$\mathbf{M}_i(\mathbf{z}) := \begin{pmatrix} \mathbf{A}_i^1(\mathbf{z}) & \mathbf{B}_i^1(\mathbf{z}) \\ \vdots & \vdots \\ \mathbf{A}_i^{n_i(n_i+1)}(\mathbf{z}) & \mathbf{B}_i^{n_i(n_i+1)}(\mathbf{z}) \end{pmatrix} \in \mathbb{R}^{n_i(n_i+1) \times n_i(n_i+1)} \quad (35)$$

By assumption of variability  $V^i$ ,  $\mathbf{M}_i(\mathbf{z})$  is invertible for all  $\mathbf{z} \in \mathbb{R}^d$ . Thus (34) has a unique solution, which is  $\mathbf{a}_{j,l}^i = \mathbf{0}, \mathbf{b}_{j,l}^i = \mathbf{0}$ .

$\mathbf{a}_{j,l}^i(\mathbf{z}) = \mathbf{0}$  implies  $\forall k, m \in \overline{\text{pa}}(i), \partial_j \varphi_k(\mathbf{z}) \partial_l \varphi_m(\mathbf{z}) = 0$ .

Since  $l \in pa(i)$ ,  $\partial_l \varphi_l(\mathbf{z}) \neq 0$  is in  $\mathbf{a}_{j,l}^i(\mathbf{z})$ , so  $\forall k \in \overline{pa}(i), \partial_j \varphi_k(\mathbf{z}) \partial_l \varphi_l(\mathbf{z}) = 0$ , which implies  $\partial_j \varphi_k(\mathbf{z}) = 0$ . We have proven that for all  $j \in anc(i) \setminus pa(i)$ , for all  $k \in \overline{pa}(i)$ , for all  $\mathbf{z} \in \mathbb{R}^d$ ,  $\partial_j \varphi_k(\mathbf{z}) = 0$ . Namely,  $\varphi_k$  only depends on  $\overline{pa}(i)$ . Combining with the result in Thm. 4.2(i), we obtain the conclusion.  $\square$

## E Known vs. unknown intervention targets

In the main paper, for simplicity, we only provided the minimal set of notation required for describing the problem of CauCA with *known intervention targets*. However, we believe that a more general version of CauCA should also be considered for cases where the targets are unknown. In fact, in the following, we will distinguish many problem settings, ranging from *totally known targets* to *totally unknown targets*. Each setting may be more or less suited to model a collection of datasets, depending on the amount and kind of prior knowledge available.

In the following, we provide a general framework in which CauCA with unknown intervention targets can be rigorously formulated. Note that  $G$  denotes a DAG such that  $V(G) = [d]$ , indexed by natural numbers, which correspond to the indices of targets  $\tau_i$  of interventions.  $\mathsf{G}$  denotes instead a DAG equipped only with a *partial order* induced by the arrows, namely,  $V(\mathsf{G})$  is a set not necessarily indexed by natural numbers. For example, consider  $V(\mathsf{G}) = \{\text{"cloudy"}, \text{"sprinkle"}, \text{"raining"}, \text{"wet grass"}\}$ . In this case, the probability distribution  $\mathbb{P}$  that is Markov relative to this graph is not defined because of the lack of indices:  $\mathbb{P}_1$  might denote the marginal of “cloudy”, “sprinkle”, “raining” or “wet grass”. However,  $\mathbb{P}$  is Markov relative to an *indexed DAG* of  $\mathsf{G}$ , denoted  $G \models \mathsf{G}$ , which denotes that there exists a bijection  $\sigma$  s.t.  $(u, v) \in E(G)$  iff  $(\sigma(u), \sigma(v)) \in E(\mathsf{G})$ .

**Definition E.1.** Given two DAGs  $G \models \mathsf{G}$  and  $G' \models \mathsf{G}$ , an isomorphism from  $G$  to  $G'$  is a bijection  $\sigma$  of  $V(G) = [d]$  such that  $(i, j) \in E(G)$ ,  $(\sigma(i), \sigma(j)) \in E(G')$ . An automorphism of  $G$  is an isomorphism  $G \rightarrow G$ .

**Definition E.2** (Identifiabilities of Causal component analysis, general setting). Given  $\mathsf{G}$  a partially ordered DAG,  $\mathcal{F}$  a class of diffeomorphisms,  $\mathcal{P}_{\mathsf{G}}$  a set of distributions such that for every  $\mathbb{P} \in \mathcal{P}_{\mathsf{G}}$  there exists  $G \models \mathsf{G}$  such that  $\mathbb{P} \in \mathcal{P}_G$ , we define  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$  as a class of latent CBN  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  such that  $G \models \mathsf{G}$ ,  $f \in \mathcal{F}$  and  $\mathbb{P} \in \mathcal{P}_G$ .

(i) We define CauCA with known intervention targets in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$  as a class of latent CBN models such that all latent CBN models have the same  $G$  and  $(\tau_i)_{i \in \llbracket 0, K \rrbracket}$ .

(ii) We define CauCA with known intervention targets up to graph automorphisms in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$  as a class of latent CBN models such that for any two latent CBN models  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  and  $(G', \mathbf{f}', (\mathbb{P}^i, T'_i)_{i \in \llbracket 0, K \rrbracket})$ , there exists  $\sigma$  isomorphism  $G \rightarrow G'$  such that  $T'_i = \sigma(\tau_i)$  for all  $i \in [K]$ .

(iii) We define CauCA with matched intervention targets in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$  as a class of latent CBN models such that for any two latent CBN models  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  and  $(G', \mathbf{f}', (\mathbb{P}^i, \tau'_i)_{i \in \llbracket 0, K \rrbracket})$ , there exists  $\sigma \in \mathfrak{S}_d$  such that  $G' = \sigma(G)$ ,  $\tau'_i = \sigma(\tau_i)$  for all  $i \in [K]$ .

(iv) We define CauCA with unknown intervention targets in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$  as a class of latent CBN models such that for any two latent CBN models  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  and  $(G', \mathbf{f}', (\mathbb{P}^i, \tau'_i)_{i \in \llbracket 0, K \rrbracket})$ , there exists  $\sigma$  isomorphism  $G \rightarrow G'$ .

We say that the CauCA in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$ <sup>4</sup> is identifiable up to  $\mathcal{S}$  if for any  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  and  $(G', \mathbf{f}', (\mathbb{Q}^i, \tau'_i)_{i \in \llbracket 0, K \rrbracket})$  in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$ , the relation  $\mathbf{f}_* \mathbb{P}^i = \mathbf{f}'_* \mathbb{Q}^i \forall i \in \llbracket 0, K \rrbracket$  implies that there is  $\mathbf{h} \in \mathcal{S}$  such that  $\mathbf{h} = \mathbf{f}'^{-1} \circ \mathbf{f}$  on the support of  $\mathbb{P}$ .

**Remark E.3.** In this general framework, the dataset  $\mathcal{D}_i := \left( \{\mathbf{x}^{(j)}\}_{j=1}^{N_i} \right)$ ,  $T_i$  denotes the nodes in  $V(\mathsf{G})$  (“cloudy”, “sprinkler” etc) instead of nodes  $\tau_i \subset [d]$  in  $V(G)$  (2).

If all  $d$  variables are included in the known targets, then there exists a unique bijection  $\sigma : V(G) \rightarrow V(\mathsf{G})$ . This implies that for any latent CBN  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, K \rrbracket})$  in  $(\mathsf{G}, \mathcal{F}, \mathcal{P}_{\mathsf{G}})$ ,  $\tau_i = \sigma^{-1}(T_i)$  i.e. the targets  $\tau_i$  are uniquely defined by  $T_i$  in each interventional regime. In this case, without loss of generality, we can suppose that  $\forall i \in V(G)$ , if  $i \in pa(j)$ , then  $i < j$ . This can be achieved by rearranging the nodes in the graph and, correspondingly, the coordinates of

<sup>4</sup>Or  $(G, \mathcal{F}, \mathcal{P}_G)$  for CauCA with known intervention targets.

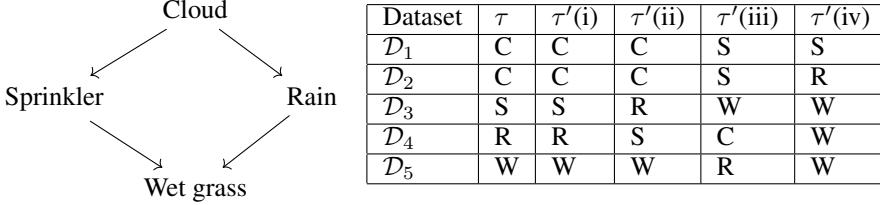


Figure 5: Typical cases in the four definitions of CauCA in Defn. E.2. Each row corresponds to a dataset in which there is one perfect intervention target: Cloud (C), Sprinkler (S), Rain (R), and Wet grass (W). We consider two latent CBN models with their nodes corresponding to the same nodes in  $V(G)$ , with intervention targets  $(\tau_i)_{i \in [5]}$  and  $(\tau'_i)_{i \in [5]}$ . Suppose we are given  $G$  and  $(\tau_i)_{i \in [5]}$ . The mapping  $V(G') \rightarrow V(G)$  are unknown in all settings except with known intervention targets. We give one example of mapping  $V(G') \rightarrow V(G)$  for each of the 4 settings in Defn. E.2.

$(\mathbb{P}^k, \tau_k) \forall k \in \llbracket 0, K \rrbracket$ . When the intervention targets are totally unknown,  $(G, \mathcal{F}, \mathcal{P}_G)$  only gives us the information of an unordered graph  $G$ , and in every interventional regime the candidate latent CBN models that achieve the same likelihood might intervene on totally different variables. In this case, we cannot rearrange the nodes without loss of generality.

Our main paper has shown identifiability results about Defn. E.2 (i). In the following, we generalize Thm. 4.2 to CauCA with known intervention targets up to graph automorphisms. We also generalize Thm. 4.6 to matched intervention targets. A generalization for CauCA with totally unknown intervention targets is an open question.

**Assumption E.4** (Interventional discrepancy, general version). *For all  $\mathbf{z} \in \mathbb{R}^d$ , for all pairs of stochastic interventions  $\tilde{p}_{\tau_i}$  and corresponding causal mechanism for the  $\tau_i$ -th variable,  $p_{\tau_i}$ , we have  $\forall i \in [d]$*

$$\frac{\partial(\ln p_{\tau_i})}{\partial z_{\tau_i}}(z_{\tau_i} | \mathbf{z}_{pa(\tau_i)}) \neq \frac{\partial(\ln \tilde{p}_{\tau_i})}{\partial z_{\tau_i}}(z_{\tau_i} | \mathbf{z}_{pa^i(\tau_i)}) \quad a.e. \quad (36)$$

**Theorem E.5.** *For CauCA with known intervention targets up to graph automorphisms in  $(G, \mathcal{F}, \mathcal{P}_G)$ , suppose that Asm. E.4 holds.*

(i) *Suppose for each node in  $[d]$ , there is one (perfect or structure-preserving) stochastic intervention such that Asm. E.4 is verified. Then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to*

$$\begin{aligned} \mathcal{S}_{Aut\bar{G}} = & \left\{ \mathbf{P}_\sigma \circ \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d | \sigma \in Aut_G, \mathbf{P}_\sigma \text{ permutation matrix of } \sigma, \right. \\ & \left. \mathbf{h}(\mathbf{z}) = (h_i(\mathbf{z}_{\overline{anc}(i)}))_{i \in [d]}, \mathbf{h} \text{ is } \mathcal{C}^1\text{-diffeomorphism} \right\} \end{aligned}$$

(ii) *Suppose for each node  $i$  in  $[d]$ , there is one perfect stochastic intervention such that Asm. E.4 is verified, then CauCA in  $(G, \mathcal{F}, \mathcal{P}_G)$  is identifiable up to*

$$\begin{aligned} \mathcal{S}_{G\text{-scaling}} := & \left\{ \mathbf{P}_\sigma \circ \mathbf{h} | \sigma \in Aut_G, \mathbf{P}_\sigma \text{ permutation matrix of } \sigma, \mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^d, \right. \\ & \left. \mathbf{h}(\mathbf{z}) = (h_1(z_1), \dots, h_d(z_d)) \text{ for some } h_i \in \mathcal{C}^1(\mathbb{R}, \mathbb{R}) \text{ with } |h'_i(\mathbf{z})| > 0 \quad \forall \mathbf{z} \in \mathbb{R}^d \right\} \end{aligned}$$

*Proof.* **Proof of (i):** The proof is based on the proof of Thm. 4.2(i). Consider two latent models achieving the same likelihood across all interventional regimes:  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d \rrbracket})$  and  $(G, \mathbf{f}', (\mathbb{Q}^i, \tau'_i)_{i \in \llbracket 0, d \rrbracket})$ . By the definition of latent CBN models with known targets up to graph automorphisms, there exists  $\sigma$  automorphism of  $G$  s.t. the targets  $(\tau'_i)_{i \in [d]} = (\sigma(1), \dots, \sigma(d))$ . Since  $(\tau_i)_{i \in [d]}$  covers all  $d$  nodes in  $G$ , by rearranging  $(\tau_i)_{i \in [d]}$  we can suppose without loss of generality that  $\tau_i = i \forall i \in [d]$ . Namely, in the  $i$ -th interventional regime,

$$\mathbb{P}^i = \widetilde{\mathbb{P}}_i(Z_j | \mathbf{Z}_{pa^i(j)}) \prod_{j \neq i} \mathbb{P}(Z_j | \mathbf{Z}_{pa(j)}), \quad \mathbb{Q}^i = \widetilde{\mathbb{Q}}_{\sigma(i)}(Z_{\sigma(i)} | \mathbf{Z}_{pa^{\sigma(i)}(\sigma(i))}) \prod_{j \neq i} \mathbb{Q}(Z_{\sigma(j)} | \mathbf{Z}_{pa(\sigma(j))}).$$

Define  $\varphi := \mathbf{f}'^{-1} \circ \mathbf{f}$ . Denote its  $i$ -th dimension output function as  $\varphi_i : \mathbb{R}^d \rightarrow \mathbb{R}$ . We will prove by induction that  $\forall i \in [d], \forall j \notin \overline{anc}(i), \forall \mathbf{z} \in \mathbb{R}^d, \frac{\partial \varphi_{\sigma(i)}}{\partial z_j}(\mathbf{z}) = 0$ .

For any  $i \in \llbracket 0, d \rrbracket$ ,  $\mathbf{f}_*\mathbb{P}^i = \mathbf{f}'_*\mathbb{Q}^i$ . Since  $\varphi$  is a diffeomorphism, by the change of variables formula,

$$p^i(\mathbf{z}) = q^i(\varphi(\mathbf{z})) |\det \varphi(\mathbf{z})| \quad (37)$$

For  $i = 0$ , factorize  $p_i$  and  $q^i$  according to  $G$ , then take the logarithm on both sides:

$$\sum_{j=1}^d \ln p_j(z_j | \mathbf{z}_{\text{pa}(j)}) = \sum_{j=1}^d \ln q_j(\varphi_j(\mathbf{z}) | \varphi_{\text{pa}(j)}(\mathbf{z})) + \ln |\det D\varphi(\mathbf{z})| \quad (38)$$

For  $i = 1$ ,  $\tilde{\mathbb{Q}}_1$  has no conditionals, and so does  $Z_{\sigma(1)}$ . Thus  $q^i$  is factorized as

$$q^1(\mathbf{z}) = \tilde{q}_{\sigma(1)}(z_{\sigma(1)}) \prod_{j \neq \sigma(1)} q_j(z_j | \mathbf{z}_{\text{pa}(j)})$$

So the equation (37) for  $i = 1$  after taking logarithm is

$$\begin{aligned} \ln \tilde{p}_1(z_1) + \sum_{j \neq 1} \ln p_j(z_j | \mathbf{z}_{\text{pa}(j)}) &= \ln \tilde{q}_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z}) | \varphi_{\text{pa}^{\sigma(1)}(\sigma(1))}(\mathbf{z})) \\ &\quad + \sum_{j \neq \sigma(1)} q_j(\varphi_j(\mathbf{z}) | \varphi_{\text{pa}(j)}(\mathbf{z})) + \ln |\det D\varphi(\mathbf{z})| \end{aligned} \quad (39)$$

subtract (39) by (38),

$$\ln \tilde{p}_1(z_1) - \ln p_1(z_1) = \ln \tilde{q}_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z})) - \ln q_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z})) \quad (40)$$

For any  $i \neq 1$ , take the  $i$ -th partial derivative of both sides:

$$0 = \left[ \frac{\tilde{q}'_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z}))}{\tilde{q}_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z}))} - \frac{q'_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z}))}{q_{\sigma(1)}(\varphi_{\sigma(1)}(\mathbf{z}))} \right] \frac{\partial \varphi_{\sigma(1)}}{\partial z_i}(\mathbf{z})$$

By Asm. E.4, the term in the parenthesis is non-zero a.e. in  $\mathbb{R}^d$ . Thus  $\frac{\partial \varphi_{\sigma(1)}}{\partial z_i}(\mathbf{z}) = 0$  a.e. in  $\mathbb{R}^d$ . Since  $\varphi = \mathbf{f}'^{-1} \circ \mathbf{f}$  where  $\mathbf{f}, \mathbf{f}'$  are  $C^1$ -diffeomorphisms, so is  $\varphi$ .  $\frac{\partial \varphi_{\sigma(1)}}{\partial z_i}$  is continuous and thus equals zero everywhere.

Now suppose  $\forall k \in [i-1], \forall j \notin \overline{\text{anc}}(k), \frac{\partial \varphi_{\sigma(k)}}{\partial z_j}(\mathbf{z}) = 0, \forall \mathbf{z} \in \mathbb{R}^d$ . Then for interventional regime  $i$ ,

$$\begin{aligned} \ln \tilde{p}_i(z_i | \mathbf{z}_{\text{pa}^i(i)}) + \sum_{j \neq i} \ln p_j(z_j | \mathbf{z}_{\text{pa}^j(j)}) &= \ln \tilde{q}_{\sigma(i)}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z})) \\ &\quad + \sum_{j \neq i} q_{\sigma(j)}(\varphi_{\sigma(j)}(\mathbf{z}) | \varphi_{\text{pa}(\sigma(i))}(\mathbf{z})) + \ln |\det D\varphi(\mathbf{z})| \end{aligned}$$

Subtracted by (38),

$$\begin{aligned} \ln \tilde{p}_i(z_i | \mathbf{z}_{\text{pa}^i(i)}) - \ln p_i(z_i | \mathbf{z}_{\text{pa}(i)}) &= \ln \tilde{q}_{\sigma(i)}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z})) \\ &\quad - \ln q_{\sigma(i)}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}(\sigma(i))}(\mathbf{z})) \end{aligned}$$

For any  $j \notin \overline{\text{anc}}(i), j \notin \text{pa}(i) \supset \text{pa}^i(i)$  by assumption. Take partial derivative over  $z_j$ :

$$\begin{aligned} 0 &= \frac{\sum_{k \in \overline{\text{pa}}^{\sigma(i)}(\sigma(i))} \frac{\partial \tilde{q}_{\sigma(i)}}{\partial \mathbf{x}_k}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{\tilde{q}_{\sigma(i)}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z}))} \\ &\quad - \frac{\sum_{k \in \overline{\text{pa}}(\sigma(i))} \frac{\partial q_{\sigma(i)}}{\partial \mathbf{x}_k}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}(\sigma(i))}(\mathbf{z})) \frac{\partial \varphi_k}{\partial z_j}(\mathbf{z})}{q_{\sigma(i)}(\varphi_{\sigma(i)}(\mathbf{z}) | \varphi_{\text{pa}(\sigma(i))}(\mathbf{z}))} \end{aligned} \quad (41)$$

Now prove that  $G_{\sigma(i)} = \sigma(G_i)$ :

$G_{\sigma(i)}$  is obtained by either a perfect stochastic or a structure-preserving stochastic intervention. For a structure-preserving stochastic intervention,  $G = G_{\sigma(i)} = \sigma(G_i)$ . For a perfect stochastic intervention, since  $\tau'_i = \sigma(\tau_i)$ , in  $G_{\sigma(i)}$  only the arrows towards  $\sigma(\tau_i)$  are deleted, which correspond to deleting arrows towards  $\tau_i$  in  $G_i$ .

Thus  $G_{\sigma(i)} = \sigma(G_i)$ . Thus  $\text{pa}^{\sigma(i)}(\sigma(i)) = \text{pa}^{\sigma(i)}(\sigma(i)) = \sigma(\text{pa}^i(i))$ , the last equality by the definition of automorphism  $\sigma$ ,  $\sigma(\text{pa}(i)) = \text{pa}(\sigma(i))$ .

The assumption of induction says  $\forall k \in [i-1] \quad \forall j \notin \overline{\text{anc}}(k) \quad \frac{\partial \varphi_{\sigma(k)}}{\partial z_j}(\mathbf{z}) = 0 \quad \forall \mathbf{z} \in \mathbb{R}^d$ . For all  $k \in \text{pa}(i) \supset \text{pa}^i(i)$ , since  $j \notin \overline{\text{anc}}(i)$ ,  $j \notin \overline{\text{anc}}(k)$  as well. By induction,  $\frac{\partial \varphi_{\sigma(k)}}{\partial z_j}(\mathbf{z}) = 0$ . So the second sum of the right-hand side of (41) can be canceled except for  $k = \sigma(i)$ . Also,  $\text{pa}(\sigma(i)) = \sigma(\text{pa}(i))$ ,  $\text{pa}^{\sigma(i)}(\sigma(i)) = \sigma(\text{pa}^i(i))$ . By the same assumption in the current induction, for all  $l \in \text{pa}^{\sigma(i)}(\sigma(i)) = \sigma(\text{pa}^i(i))$ ,  $\frac{\partial \varphi_l}{\partial z_j}(\mathbf{z}) = 0$ . So the first sum of the right-hand side of (41) can be canceled except for  $k = \sigma(i)$ .

The equation rewrites after deleting the partial derivatives that are zero:

$$0 = \left[ \frac{\frac{\partial \tilde{q}_{\sigma(i)}}{\partial \varphi_{\sigma(i)}} (\varphi_{\sigma(i)}(\mathbf{z}) \mid \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z}))}{\tilde{q}_{\sigma(i)} (\varphi_{\sigma(i)}(\mathbf{z}) \mid \varphi_{\text{pa}^{\sigma(i)}(\sigma(i))}(\mathbf{z}))} - \frac{\frac{\partial q_{\sigma(i)}}{\partial \varphi_{\sigma(i)}} (\varphi_{\sigma(i)}(\mathbf{z}) \mid \varphi_{\text{pa}(\sigma(i))}(\mathbf{z}))}{q_{\sigma(i)} (\varphi_{\sigma(i)}(\mathbf{z}) \mid \varphi_{\text{pa}(\sigma(i))}(\mathbf{z}))} \right] \frac{\partial \varphi_{\sigma(i)}}{\partial z_j}(\mathbf{z})$$

By Asm. E.4, the term in parenthesis is nonzero a.e., thus  $\frac{\partial \varphi_{\sigma(i)}}{\partial z_j}(\mathbf{z}) = 0$  a.e. Since  $\frac{\partial \varphi_{\sigma(i)}}{\partial z_j}$  is continuous, it equals zero everywhere.

The induction is finished when  $i = d$ . We have proven that  $\forall i \in [d], \forall j \notin \overline{\text{anc}}(i), \forall \mathbf{z} \in \mathbb{R}^d, \frac{\partial \varphi_{\sigma(i)}}{\partial z_j}(\mathbf{z}) = 0$ , i.e.  $(\mathbf{P}_\sigma^{-1} D\varphi(\mathbf{z}))_{ij} = (D\varphi(\mathbf{z}))_{\sigma(i)j} = \frac{\partial \varphi_{\sigma(i)}}{\partial z_j}(\mathbf{z}) = 0$

Thus  $\mathbf{P}_\sigma^{-1} \varphi \in \mathcal{S}_{\bar{G}}$ .  $\varphi \in \mathcal{S}_{\text{Aut}\bar{G}}$ .

**Proof of (ii):**

By the result of (i),  $\psi := \mathbf{f}'^{-1} \circ \mathbf{f} \in \mathcal{S}_{\text{Aut}\bar{G}}$ , thus there exists a permutation matrix  $\mathbf{P}_\sigma$  s.t.  $\mathbf{P}_\sigma^{-1} \psi \in \mathcal{S}_{\bar{G}}$ . Denote  $\varphi = \mathbf{P}_\sigma^{-1} \psi$ . Apply Thm. 4.2(ii), we can prove that  $\varphi \in \mathcal{S}_{\text{scaling}}$ . Thus  $\psi = \mathbf{P}_\sigma \varphi \in \mathcal{S}_{G\text{-scaling}}$ .  $\square$

**Proposition E.6.** Suppose that  $G$  is the empty graph, and that there are  $d-1$  variables intervened on, with one single target per dataset, such that Asm. E.4 holds. Then ICA with matched intervention targets in  $(G, \mathcal{F}, \mathcal{P}_G)$  with single-node interventions (Defn. E.2) is identifiable up to

$$\mathcal{S}_{\text{reparam}} := \left\{ \mathbf{g} : \mathbb{R}^d \rightarrow \mathbb{R}^d \mid \mathbf{g} = \mathbf{P} \circ \mathbf{h} \text{ where } \mathbf{P} \text{ is a permutation matrix and } \mathbf{h} \in \mathcal{S}_{\text{scaling}} \right\}$$

*Proof.* For an empty graph  $G$ ,  $\text{Aut}_G = \mathfrak{S}_d$ . Thus ICA with matched intervention targets is the same as known intervention targets up to graph automorphisms. Consider two latent models achieving the same likelihood across all interventional regimes:  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d \rrbracket})$  and  $(G, \mathbf{f}', (\mathbb{Q}^i, \tau'_i)_{i \in \llbracket 0, d \rrbracket})$ . By the definition of latent CBN models with known targets up to graph automorphisms, there exists  $\sigma$  automorphism of  $G$  s.t. the targets  $(\tau'_i)_{i \in [d]} = (\sigma(\tau_1), \dots, \sigma(\tau_d))$ .

Apply Thm. 4.6 to  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d-1 \rrbracket})$  and  $(G, \mathbf{f}' \circ \mathbf{P}_\sigma, (\mathbb{Q}^{\sigma^{-1}(i)}, \tau_i)_{i \in \llbracket 0, d-1 \rrbracket})$ , then  $\varphi := (\mathbf{f}' \circ \mathbf{P}_\sigma)^{-1} \circ \mathbf{f} \in \mathcal{S}_{\text{scaling}}$ . Then for  $(G, \mathbf{f}, (\mathbb{P}^i, \tau_i)_{i \in \llbracket 0, d-1 \rrbracket})$  and  $(G, \mathbf{f}', (\mathbb{Q}^i, \tau'_i)_{i \in \llbracket 0, d-1 \rrbracket})$ ,  $\mathbf{f}'^{-1} \circ \mathbf{f} = \mathbf{P}_\sigma \circ \varphi \in \mathcal{S}_{\text{reparam}}$ .  $\square$

## F Additional theoretical results used in the design of experiments

### F.1 Multi-objective and pooled objective

Our identifiability theory implies that the ground-truth latent CBN could be learned by maximizing likelihood across all interventional regimes, i.e.,

$$\theta^* = \bigcap_{k=0}^d \arg \max_{\theta} \frac{1}{N_k} \sum_{n=1}^{N_k} \log p_{\theta}^k(\mathbf{x}^{(n,k)})$$

where  $\log p_{\theta}^k(\mathbf{x}^{(n,k)})$  is defined in (9). This is a multi-objective optimization problem.

In general, multi-objective optimization is hard; however, we show that, because of our assumptions, we can equivalently optimize a pooled objective. In fact, suppose

$$f_k(\theta) = \log p_{\theta}^k(\mathbf{x}) \quad f(\theta) = \sum_{k=0}^d f_k(\theta)$$

Define  $\Theta_k := \arg \max_{\theta} f_k(\theta)$ . By our problem setting we know  $\bigcap_{k=0}^d \Theta_k \neq \emptyset$ .

On the one hand, for all  $\hat{\theta} \in \bigcap_{k=0}^d \Theta_k$ ,  $\hat{\theta} \in \arg \max_{\theta} f(\theta)$ . On the other hand, we will prove by contradiction that  $\forall \hat{\theta} \in \arg \max_{\theta} f(\theta)$ ,  $\hat{\theta} \in \bigcap_{k=0}^d \Theta_k$ . Suppose there exists  $i \in [0, d]$  such that  $\hat{\theta} \notin \Theta_i$ . Then for all  $\theta^* \in \bigcap_{k=0}^d \Theta_k$ ,  $f_i(\hat{\theta}) < f_i(\theta^*)$ . Thus  $f(\hat{\theta}) = \sum_{k=0}^d f_k(\hat{\theta}) < \sum_{k=0}^d f_k(\theta^*) = f(\theta^*)$ . This yields a contradiction with  $\hat{\theta} \in \arg \max_{\theta} f(\theta)$ .

We thus conclude that  $\bigcap_{k=0}^d \arg \max_{\theta} f_k(\theta) = \arg \max_{\theta} f(\theta)$ .

### F.2 Expressivity in multi-intervention learning

To learn the latent CBN, we need to learn both the latent distributions (both the unintervened causal mechanisms and the intervened ones) and mixing functions. For the latent distributions, a natural question is whether any of them could be fixed without loss of generality. This is for example possible in the context of nonlinear ICA, where due to the indeterminacy up to element-wise nonlinear scaling of the latent variables, we can arbitrarily fix their univariate distributions w.l.o.g. The following proposition elucidates this matter for CauCA.

**Proposition F.1.** *For a ground truth latent CBN,  $(G, \mathbf{f}, (\mathbb{P}^k, \tau_k)_{k \in [0, d]})$  where  $(\mathbb{P}^k, \tau_k)_{k \in [1, d]}$  are obtained by perfect stochastic intervention on every variable respectively, fix at most one element for each  $k$  in  $\{\mathbb{Q}_k | \text{pa}(k) = \emptyset\} \cup \{\widetilde{\mathbb{Q}}_k | k \in [d]\}$ , there exists  $(G, \mathbf{f}', (\mathbb{Q}^k, \tau'_k)_{k \in [0, d]})$  s.t.  $\mathbf{f}_* \mathbb{P}^k = \mathbf{f}'_* \mathbb{Q}^k, \mathbf{f}'^{-1} \circ \mathbf{f} \in \mathcal{S}_{\text{scaling}}$ .*

In practice, this means that in order to learn  $\mathbf{f}'$  and  $(\mathbb{Q}^k)_{k \in [d]}$  with  $d$  perfect stochastic interventions, even if we fix all the intervention mechanisms  $(\widetilde{\mathbb{Q}}_k)_{k \in [d]}$ , we can still learn a true latent model up to scaling functions. Equivalently, we could also fix instead  $\{\mathbb{Q}_k | \text{pa}(k) = \emptyset\} \cup \{\widetilde{\mathbb{Q}}_k | \text{pa}(k) \neq \emptyset\}$ .

*Proof.* For any  $k \in [d]$ ,  $\mathbb{P}^k = \widetilde{\mathbb{P}}_k \prod_{j \neq k} \mathbb{P}_j$ . Without loss of generality by exchanging  $\mathbb{Q}_k$  with  $\widetilde{\mathbb{Q}}_k$ , suppose we are given  $\widetilde{\mathbb{Q}}_k$ , by Lemma B.2, there are two possible diffeomorphisms for  $T_k$  st.  $(T_k)_* \mathbb{P}_k = \widetilde{\mathbb{Q}}_k$ . Choose one of them arbitrarily. Set  $\mathbf{T} := (T_k)_{k \in [d]}$ .

Define  $\mathbb{Q}^0 := \mathbf{T}_* \mathbb{P}^0$ . By Lemma B.3,  $\forall i \in [d], \forall z_i \in \mathbb{R}, \forall \mathbf{z}_{\text{pa}(i)} \in \mathbb{R}^{\#\text{pa}(i)}$ ,

$$p_i(z_i | \mathbf{z}_{\text{pa}(i)}) = q_i(T_i(z_i) | \mathbf{T}_{\text{pa}(i)}(\mathbf{z}_{\text{pa}(i)})) |T'_i(z_i)|$$

Multiply the causal mechanisms of all  $i \in [d] \setminus \{k\}$  and the intervened mechanism of  $k$ , we get

$$\widetilde{p}_k(z_k) \prod_{i \neq k} p_i(z_i | \mathbf{z}_{\text{pa}(i)}) = \widetilde{q}_k(T_k(z_k)) |T'_k(z_k)| \prod_{i \neq k} q_i(T_i(z_i) | \mathbf{T}_{\text{pa}(i)}(\mathbf{z}_{\text{pa}(i)})) |T'_i(z_i)|$$

which is equivalent to  $\tilde{p}^k(\mathbf{z}) = \tilde{q}^k(\mathbf{T}(\mathbf{z})) |\det \mathbf{T}(\mathbf{z})|$ . This is the change of variable formula of the diffeomorphism  $\mathbf{T}$  in  $\mathbb{R}^d$ , and implies  $\mathbb{Q}^k = \mathbf{T}_* \mathbb{P}^k$ .

Define  $\mathbf{f}' := \mathbf{T} \circ \mathbf{f}^{-1}$ , then  $\mathbf{f}'^{-1} \circ \mathbf{f} = \mathbf{T} \in \mathcal{S}_{\text{scaling}}$ .  $\square$

**Remark** If  $G$  is non trivial, given  $(G, \mathbf{f}, \mathbb{P}^0)$ , in general there does not exist  $\mathbb{Q}^0 \in P_G$ ,  $\mathbf{f}' \in \mathcal{F}$  s.t.  $\mathbf{f}_* \mathbb{P}^0 = \mathbf{f}'_* \mathbb{Q}^0$ ,  $\mathbf{f}'^{-1} \circ \mathbf{f} \in \mathcal{S}_{\text{scaling}}$ .

To see why, consider the graph  $z_1 \rightarrow z_2$ . Fix  $\mathbb{Q}_1$ , there are only two possible diffeomorphisms for  $T_1$  s.t.  $\mathbb{Q}_1 = (T_1)_* \mathbb{P}_1$  by Lemma B.2.

If  $\mathbb{Q}_2 (Z_2 | T_1(z_1))$  is fixed, to find  $T_2$  s.t.

$$(T_2)_* \mathbb{P}_2 (Z_2 | z_1) = \mathbb{Q}_2 (Z_2 | T_1(z_1)) \quad \forall z_1 \in \mathbb{R}$$

by Lemma B.2, the only possible  $T_2$  are  $G(\cdot | T_1(z_1))^{-1} \circ F(\cdot | z_1)$  and  $\bar{G}(\cdot | T_1(z_1))^{-1} \circ F(\cdot | z_1)$ . In general, these two functions depend on  $z_1$ . For example if  $\mathbb{P}(Z_1) \perp\!\!\!\perp \mathbb{P}(Z_2)$ ,  $\mathbb{Q}(Z_2 | z_1) = \mathcal{N}(z_1, 1)$  then  $T_2$  depend on  $z_1$ . Namely, There is no  $\mathbf{T} \in \mathcal{S}_{\text{scaling}}$  s.t.  $\mathbb{Q}^0 = \mathbf{T}_* \mathbb{P}^0$ .

## G Details Experiments

### G.1 Synthetic Data Generation

**Directed Acyclic Graph (DAG).** In order to generate data, we begin by sampling a random DAG  $G \sim \mathbb{Q}_G$ , where  $\mathbb{Q}_G$  is a distribution over DAGs. The edge density of the DAG is set to 0.5 in topological order, meaning that the edges in the DAG are constrained to follow the variable index order: an edge  $Z_i \rightarrow Z_j$  can only exist if  $j > i$ . To construct the DAG, we individually sample each potential edge  $Z_i \rightarrow Z_j$  with  $j > i$  with a probability of 0.5, while all other edges are assigned a probability of 0. For experiments conducted in the CauCA setting with non-trivial graphs (i.e., not empty), we reject and redraw any sampled DAGs that do not contain any edges.

**Causal Bayesian network (CBN).** To sample data from a CBN, we start by drawing the parameters of a linear Gaussian Structural Causal Model (SCM) with additive noise:

$$Z_i := \sum_{j \in \text{pa}(i)} \alpha_{i,j} Z_j + \varepsilon_i, \quad (42)$$

where the linear parameters are drawn from a uniform distribution  $\alpha_{i,j} \sim \text{Uniform}(-a, a)$  and the noise variable is Gaussian  $\varepsilon_i \sim \mathcal{N}(0, 1)$ . The *signal-to-noise ratio* for the SCM, denoted as  $a/\text{std}(\varepsilon_i) = a$ , describes the strength of the dependence of the causal variables relative to the exogenous noise. For most experiments, the signal-to-noise ratio is set to 1. In Fig. 4 (g), we explore values ranging from 1 to 10. To specify the latent CBNs, we can define the conditional distributions entailed by the SCMs defined as in eq. (42), see also App. G.2 and eq. (46).

**Generating related datasets.** For a given CBN, we generate  $d+1$  related datasets: one observational dataset and  $d$  datasets where the CBN was modified by a perfect stochastic intervention on one variable (i.e., one dataset for each variable in the CBN). W.l.o.g. we assume that the  $k^{\text{th}}$  variable was intervened on in dataset  $k$ ; with  $k=0$  being the observational dataset. The intervention is applied in the SCM by removing the influence of the parent variables and changing the exogenous noise by shifting its mean up or down. Hence, for dataset  $k$  we have

$$Z_k := \tilde{\varepsilon}_k, \quad \text{with } \tilde{\varepsilon}_k \sim \mathcal{N}(\mu, 1), \quad (43)$$

where the mean of the noise  $\mu$  is fixed per dataset and uniformly sampled from  $\{\pm 2\}$ . Each dataset comprises a total of 200,000 data points, resulting in  $(d+1) \times 200,000$  data points in total for each CBN.

**Mixing function.** The mixing function takes the form of a multilayer perceptron  $\mathbf{f} = \sigma \circ \mathbf{A}_M \circ \dots \circ \sigma \circ \mathbf{A}_1$ , where  $\mathbf{A}_m \in \mathbb{R}^{d \times d}$  for  $m \in [\![1, M]\!]$  denote invertible linear maps, and  $\sigma$  is an element-wise invertible nonlinear function. The elements of the linear maps are sampled independently  $(\mathbf{A}_m)_{i,j} \sim \text{Uniform}(0, 1)$  for  $i, j \in [\![1, d]\!]$ . A sampled matrix  $\mathbf{A}_m$  is rejected and re-drawn if

$|\det \mathbf{A}_m| < 0.1$  to rule out linear maps that are (close to) singular. The invertible element-wise nonlinearity is a leaky-tanh activation function:

$$\sigma(x) = \tanh(x) + 0.1x, \quad (44)$$

as used in (Gresele et al., 2020).

## G.2 Model architecture

**Normalizing flows.** We use normalizing flows (Papamakarios et al., 2021) to learn an encoder  $\mathbf{g}_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ . Normalizing flows model observations  $\mathbf{x}$  as the result of an invertible, differentiable transformation  $\mathbf{g}_\theta$  on some base variables  $\mathbf{z}$ ,

$$\mathbf{x} = \mathbf{g}_\theta(\mathbf{z}). \quad (45)$$

We apply a series of  $L = 12$  such transformations  $\mathbf{g}_{\theta^l}^l : \mathbb{R}^d \rightarrow \mathbb{R}^d$ , which we refer to as *flow layers*, such that the resulting transformation is given by  $\mathbf{g}_\theta = \mathbf{g}_{\theta^L}^L \circ \dots \circ \mathbf{g}_{\theta^1}^1$ . We use Neural Spline Flows (Durkan et al., 2019) for the invertible transformation, with a 3-layer feedforward neural network with hidden dimension 128 and a permutation in each flow layer.

**Base distribution.** We extend the typically used simple base distributions to encode information about the CBN. We have one distribution per dataset  $(\hat{p}_{\theta_{\text{CBN}}^k}^k)_{k \in \llbracket 0, d \rrbracket}$  over the learned base noise variables  $\mathbf{z}$ . The conditional density of latent variable  $i$  in dataset  $k$  is given by

$$\hat{p}_{\theta_{\text{CBN}}^k}^k(z_i \mid \mathbf{z}_{\text{pa}(i)}) = \mathcal{N}\left(\sum_{j \in \text{pa}(i)} \hat{\alpha}_{i,j} z_j, \hat{\sigma}_i\right), \quad (46)$$

when  $i \neq k$ , i.e., when variable  $i$  is not intervened on. For  $i = k$ , we have

$$\hat{p}_{\theta_{\text{CBN}}^k}^k(z_i) = \mathcal{N}(\hat{\mu}_i^k, \hat{\sigma}_i^k). \quad (47)$$

In summary, the base distribution parameters are the parameters of the linear relationships between parents and children in the CBN  $(\hat{\alpha}_{i,j})_{i,j \in \llbracket 1, d \rrbracket}$ , the standard deviations for each variable in the observational setting  $(\hat{\sigma}_i)_{i \in \llbracket 1, d \rrbracket}$ , and mean and standard deviation for the intervened variable in each dataset  $(\hat{\mu}_i^k, \hat{\sigma}_i^k)_{i,k \in \llbracket 1, d \rrbracket}$ .

**Linear baseline model.** For the linear baseline models shown in Fig. 4 (a, b, e, f) we replace the nonlinear transformations  $(\mathbf{g}_{\theta^l}^l)_{l \in \llbracket 1, L \rrbracket}$  by a single linear transformation. The base distribution stays the same.

**Graph-misspecified model.** In order to test the impact of providing knowledge about the causal structure, we compare the CauCA model to one that assumes the latents are independent. This is achieved by setting  $\hat{\alpha}_{i,j} = 0 \forall i, j \in \llbracket 1, d \rrbracket$  in the base distribution (46).

## G.3 Training and model selection

**Training parameters.** We use the ADAM optimizer (Kingma and Ba, 2014) with cosine annealing learning rate scheduling, starting with a learning rate of  $5 \times 10^{-3}$  and ending with  $1 \times 10^{-7}$ . We train the model for 50–200 epochs with a batch size of 4096. The number of epochs was tuned manually for each type of experiment to ensure reliable convergence of the validation log probability.

**Pooled objective.** The learning objective described in § 5 is using the pooled rather than the multi-objective formulation. In App. F, we prove that for our problem the two are equivalent.

**Fixing CBN parameters.** As explained in Prop. F.1, we can fix some of the CBN parameters w.l.o.g. In our case, we fix the noise parameters for intervened mechanisms of non-root variables and observational mechanisms of root variables.

**Model selection.** For each drawn latent CBN, we train three models with different initializations and select the model with the highest validation log probability at the end of training.

**Compute.** Each training run takes 2–8 hours on NVIDIA RTX-6000 gpus. For the experiments shown in the main paper, we performed 450 training runs (30 per violin plot / point in Fig. 4 (g)) which sums up to around 2250 compute hours (assuming an average run time of 5 hours).