Multivariable Causal Discovery for General Nonlinear Functions

Anonymous authors
Paper under double-blind review

Abstract

Today's methods for uncovering causal relationships from observational data either constrain functional assignments (linearity/additive noise assumptions) or the data generating process (e.g., non-i.i.d. assumptions). We assume non-i.i.d. data to develop a framework for causal discovery that works for general non-linear dependencies. We use nonlinear Independent Component Analysis (ICA) to infer the underlying sources from the observed variables. Unlike previous works, which use conditional independence tests, we rely on the Jacobian of the inference function to determine the causal relationships. In particular, we prove that, under strong identifiability, the inference function's Jacobian captures the sparsity structure of the causal graph; thus generalizing the classic LiNGAM method to the nonlinear case. Our approach avoids the cost of exponentially many independence tests and makes our method end-to-end differentiable. We demonstrate that the proposed method can infer the causal graph on multiple synthetic data sets, and in most scenarios outperforms previous work.

1 Introduction

Traditional statistical learning methods model correlations in data. Though they have achieved super-human performance in multiple fields they have limited value in understanding cause-effect relationships. A prevalent consequence of this shortcoming is the observed tendency for models to learn from spurious features or shortcuts (Geirhos et al., 2020) (e.g., classifying objects based on their backgrounds). In contrast, *causal models* (Pearl, 2009a) construct the world according to the Independent Causal Mechanisms (ICM) principle (Peters et al., 2017), where building blocks (mechanisms) neither influence nor inform each other. Modeling temperature T and altitude A is a classic example (Peters et al., 2017): changing A affects T, but not vice versa—this relationship is described by the Directed Acyclic Graph (DAG) $A \to T$. The ICM principle means that the same mechanism $p(T|A)$ describes how altitude affects temperature for different $p(A)$, but the same cannot be said about $p(A|T)$ and $p(T)$.

Causal Discovery (CD) describes the process of extracting causal structure from data in the form of a DAG. Having *interventional* data—such as in the form of Randomized Controlled Trials (RCTs)—is desirable as it enables answering questions of interventional nature, such as 'What will happen if variable X is changed?'. However, RCTs can be costly, infeasible (Eberhardt et al., 2012), or even unethical. Thus, developing effective CD methods reliant on *observational* data alone is of significant interest. In general, inferring the causal direction is provably impossible without additional constraints or assumptions (Zhang et al., 2015); therefore, existing methods constrain either the model class (i.e., the functions generating the observations) or the data distribution. On the model side, these constraints include linear (Shimizu et al., 2006; Tashiro et al., 2014; Shahbazinia et al., 2021; Zheng et al., 2018) or specific nonlinear relationships (e.g., with additive noise) (Hoyer et al., 2008; Peters et al., 2011; Schölkopf et al., 2021; Yu et al., 2019; Shen et al., 2020; Lachapelle et al., 2020; Ng et al., 2019). On the data side, assumptions include non-stationarity (Schölkopf et al., 2012; Monti et al., 2019) or exchangeability (Guo et al., 2022).

CD aims to infer the ground-truth cause-effect relationships, which connects it to the *identifiability* literature, where the goal is to learn a model equivalent to the ground truth (up to indeterminacies, such as permutations or element-wise nonlinearities).

An extensively studied method for learning identifiable representations is Independent Component Analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001), which requires that the inferred components *(sources)* are independent. Recent work has relied on nonlinear Independent Component Analysis (NLICA) (Zimmermann

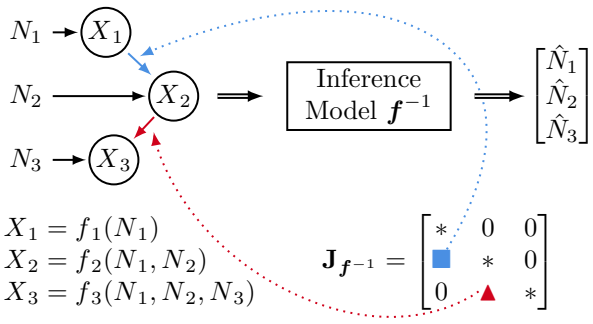

Figure 1: **The Jacobian of the inference network $\mathbf{J}_{\boldsymbol{f}^{-1}}$ informs about the DAG**. We show that if observations $\boldsymbol{X}$ are generated from noise variables $\boldsymbol{N}$ via a general nonlinear Structural Equation Model (SEM) $\boldsymbol{f}$, then the corresponding DAG can be inferred from the Jacobian of a model that identifies $\boldsymbol{N}$ under certain assumptions on $\boldsymbol{N}$

et al., 2021; Klindt et al., 2021; Hyvärinen & Morioka, 2016; Willetts & Paige, 2021; Khemakhem et al., 2020a; Hyvärinen et al., 2019; Morioka et al., 2021; Monti et al., 2019; Khemakhem et al., 2020b; Gresele et al., 2019; Hyvärinen & Morioka, 2017; Hyvärinen et al., 2010; Hälvä & Hyvärinen, 2020; Lachapelle et al., 2022) for identifiability.

Our work builds on the NonSENS method (Monti et al., 2019), which showed that NLICA can be used for CD with general nonlinear functions and observational data. Instead of using pairwise independence tests, we draw inspiration from the Linear Non-Gaussian Acyclic Model (LiNGAM) (Shimizu et al., 2006), which uses a weight matrix to infer the DAG of a linear causal model. We extend this approach to the nonlinear case by showing that the Jacobian of the inference function (mapping from observations $\boldsymbol{X}$ to noise variables $\boldsymbol{N}$)[1] captures the sparsity structure of the DAG, provided that strong identifiability is fulfilled (Khemakhem et al., 2020b, Def. 1). Relying on the Jacobian improves scalability, since it removes the cost of d^2 independence tests for a DAG with d nodes. We train our model with NLICA, and show in Lem. 1 that the assumption on the Data Generating Process (DGP) to be structured according to a DAG provides an inductive bias to resolve the permutation indeterminacy of NLICA.

Our **contributions** can be summarized as follows:

1. We show that causal models allow us to resolve the permutation indeterminacy of ICA (Lem. 1);
2. Our **main result** (Prop. 1) proves that we can infer the DAG from the Jacobian of the inference function and also improve scalability by removing the need for independence tests;
3. We propose an end-to-end multivariable CD method for general nonlinear functions from observational data;
4. We experimentally show that our proposed method can infer the DAG across multiple synthetic data sets.

2 Background

Here, we describe causal models and connect their estimation to ICA and defer the details to Appx. A.

Structural Equation Models (SEMs). Given d-dimensional observed $\boldsymbol{X}=(X_1, \ldots, X_d)$ and noise (independent) variables $\boldsymbol{N}=(N_1, \ldots, N_d)$, their causal relationship is given by d *deterministic* functional assignments (Peters et al., 2017),

$$X_i := f_i\left(\boldsymbol{Pa}_i, N_i\right) \qquad \forall i, \tag{1}$$

where $\boldsymbol{Pa}_i \subset \boldsymbol{X}$ are the parents of X_i and f_i are the components of the vector-valued function $\boldsymbol{f}$. We describe the computation of $\boldsymbol{X}$ for a given $\boldsymbol{N}$ with an iterative process (denoting the iteration step with a superscript),

[1] In our paper, inference refers only to this process and not to amortized inference for direct graph discovery as proposed in Lorch et al. (2021)

which is a useful concept for justifying our proposal (§ 3). Initially, $\boldsymbol{N}$ is drawn from its density. To calculate $\boldsymbol{X}$ for $\boldsymbol{N}$, the functional assignment $\boldsymbol{f}$ needs to be applied d times. Namely, according to (1), each X_i requires that its parents $\boldsymbol{Pa}_i$ are calculated. After sampling $\boldsymbol{N}$, only the (empty) parent sets of root nodes are calculated. Thus, the first application of $\boldsymbol{f}$ yields the X_i values for such nodes. In the second iteration, the children of root nodes can be calculated (since we have all parents from the first iteration), and so on. This yields an iterative algorithmic formulation of the SEM, describing the computational graph given by the DAG as:

$$\boldsymbol{X} = \boldsymbol{X}^d := \boldsymbol{f}^{(d)}\left(\boldsymbol{X}^0, \boldsymbol{N}\right) = \boldsymbol{f}\left(\boldsymbol{X}^{(d-1)}, \boldsymbol{N}\right) = \boldsymbol{f}\left(\boldsymbol{f}\ldots\left(\boldsymbol{f}\left(\boldsymbol{X}^0, \boldsymbol{N}\right), \boldsymbol{N}\right), \boldsymbol{N}\right), \tag{2}$$

where $\boldsymbol{X}^0$ is the initial value (w.l.o.g., we assume $\boldsymbol{X}^0 = \boldsymbol{0}$, since calculating the functional assignments will overwrite every X_i). We will also denote $\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{N})$ to indicate that $\boldsymbol{X}$ is *deterministically* determined by a particular $\boldsymbol{N}$. As in most previous works (Vowels et al., 2022, Table 1), we assume *no confounders* (all variables are observed) and *faithfulness* (loosely speaking, the coefficients/functions will not cancel an edge, cf. Assum. A.2).

Causal Discovery (CD). In CD, the data is assumed to be generated by a causal process, and the aim is to infer the corresponding DAG, which enables reasoning about interventions (without the DAG, the joint distribution $p(\boldsymbol{N})$ only admits observational queries) (Peters et al., 2017; Pearl, 2009b). Algorithmic approaches include combinatoric search (Shimizu et al., 2006; Hoyer et al., 2008; Hyttinen et al., 2013; Mitrovic et al., 2018; Raskutti & Uhler, 2018; Spirtes et al., 2000; Vowels et al., 2022), continuous optimization (Zheng et al., 2018; Lee et al., 2019; Wei et al., 2020; Ng et al., 2020; Vowels et al., 2022), and neural networks (Yu et al., 2019; Ng et al., 2019; Khemakhem et al., 2021; Yang et al., 2020; Goudet et al., 2018; Kalainathan et al., 2018; Vowels et al., 2022; Kyono et al., 2020; Moraffah et al., 2020)—we focus on the latter. Zhang et al. (2015) proved that identifying the causal direction in a general SEM is impossible without constraints on the function class and/or data distribution.
Functional constraints can include linear (Shimizu et al., 2006; Zheng et al., 2018), additive nonlinear ($X_i = f_i(\boldsymbol{Pa}_i) + N_i$) (Hoyer et al., 2008; Ng et al., 2019; Lachapelle et al., 2020; Schölkopf et al., 2021), or affine nonlinear ($X_i = f_i(\boldsymbol{Pa}_i) + h_i(N_i)$) (Khemakhem et al., 2021; Shen et al., 2020) models. Regarding the data distribution, some models require access to interventions (Brouillard et al., 2020; Ke et al., 2020; Lippe et al., 2021); others assume that $\boldsymbol{N}$ is Gaussian (Kalainathan et al., 2018; Lachapelle et al., 2020) or non-Gaussian (Shimizu et al., 2006); or require non-stationarity (Monti et al., 2019), exchangeability (Guo et al., 2022), or discreteness (Ke et al., 2020) of $\boldsymbol{N}$. Our work was inspired by (Monti et al., 2019), which provides a bivariate CD method for general nonlinear functions and non-stationary data. The authors leverage recent results in NLICA (cf. next section for details) to identify the causal direction. Although they demonstrate applicability to multivariable problems, the use of pairwise independence tests constrains scalability. In our work, we extend these results with a more scalable end-to-end solution in § 3.

Identifiability and ICA. Independent Component Analysis (ICA) (Comon, 1994; Hyvärinen et al., 2001) models the observed variables $\boldsymbol{X}$ as a mixture of *independent* variables $\boldsymbol{N}$ via a deterministic function $\boldsymbol{f}$, and focuses on defining models that are *identifiable*—i.e., $\boldsymbol{N}$ can be recovered up to indeterminacies (e.g., scaling, permutation, sign flips, element-wise transformations). Since this is provably impossible in the nonlinear case without further assumptions (Darmois, 1951; Hyvärinen & Pajunen, 1999; Locatello et al., 2019), recent work has focused on incorporating *auxiliary* variables (Hyvärinen et al., 2019; Gresele et al., 2019; Khemakhem et al., 2020a; Gassiat et al., 2022), exploiting temporal structure in the data (Hyvärinen & Morioka, 2017; 2016; Hälvä & Hyvärinen, 2020; Morioka et al., 2021; Monti et al., 2019; Hyvärinen et al., 2010; Klindt et al., 2021; Zimmermann et al., 2021), or restricting the model class (Shimizu et al., 2006; Hoyer et al., 2008; Zhang & Hyvärinen, 2012; Gresele et al., 2021). Several works have related (nonlinear) ICA to SEM estimation (Gresele et al., 2021; Monti et al., 2019; Shimizu et al., 2006; von Kügelgen et al., 2021) by inverting the DGP—i.e., estimating $\boldsymbol{f}^{-1}$ with an *inference model*.

3 Proposed methods

3.1 Intuition

The method we propose can be intuitively understood as a nonlinear extension of LiNGAM (Shimizu et al., 2006; Hoyer et al., 2008; Peters et al., 2011). LiNGAM assumes a linear causal relationship between observations $\boldsymbol{X}$ and the noise variables $\boldsymbol{N}$, i.e., $\boldsymbol{X} = \mathbf{W}\boldsymbol{N}$. Since the noise variables are assumed to be

statistically independent, linear ICA can uncover the sources $\boldsymbol{N}$ from the observations $\boldsymbol{X}$, which in turn allows us to extract the DAG from $\mathbf{W}^{-1}$ as we show in the following example.

Example 1 (Motivating example for linear SEMs)**.** *Assume a linear causal model with three variables, the DAG $X_1 \to X_2 \to X_3$, and functional relationships: $X_1 := N_1$; $X_2 := aX_1 + N_2$; $X_3 := bX_2 + N_3 : a, b \in \mathbb{R} \setminus \{0\}$. The DGP generates samples according to the DAG and has the matrix form on the left—we focus on the elements below the main diagonal as for recovering the DAG, only the paths (i.e., series of directed edges) between X_i and X_j are required and the main diagonal expresses the $N_i \to X_i$ edges. Inverting the DGP with an inference model (i.e., expressing N_i as a function of X_j; LiNGAM uses ICA to estimate the DGP) yields the matrix on the right with elements below the main diagonal capturing the DAG's $X_i \to X_j$ edges (as shown by color coding):*

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ ab & b & 1 \end{bmatrix} \begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix}; \qquad \begin{bmatrix} N_1 \\ N_2 \\ N_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ 0 & -b & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

Our method extends LiNGAM to nonlinear DGPs. First, we learn an inference model to uncover $\boldsymbol{N}$ from the observations $\boldsymbol{X}$ using nonlinear ICA and then analyse the Jacobian structure of the inference model to extract the underlying DAG. Because NLICA only identifies the ground-truth DGP up to certain indeterminancies like scaling, permutation, and sign flips, we need to introduce the notion of structural equivalence in § 3.2 and a way to resolve permutation indeterminacies in § 3.3 before introducing our main theoretical result in § 3.4.

3.2 DAG equivalence

To justify using the Jacobian of the inference network $\boldsymbol{f}^{-1}$, akin to LiNGAM's use of a weight matrix, we first connect the DAG and $\mathbf{J}_{\boldsymbol{f}^{-1}}$ via fundamental concepts from graph theory. The *adjacency matrix* $\boldsymbol{\mathcal{A}}$ of a graph with d nodes is a binary $d \times d$ matrix where each matrix element indicates the presence, or absence, of an edge (i.e., a direct connection) between a pair of nodes X_i, X_j (Defn. A.4). The *connectivity matrix* $\boldsymbol{\mathcal{C}}$ of a graph with d nodes is a binary $d \times d$ matrix where each matrix element indicates the presence, or absence, of a *directed path* between two nodes X_i, X_j (Defn. A.5). For DAGs, both $\boldsymbol{\mathcal{A}}$ and $\boldsymbol{\mathcal{C}}$ are *strictly lower-triangular*—this is why we considered only the elements below the main diagonal in Ex. 1. Furthermore, the main diagonal of $\mathbf{J}_{\boldsymbol{f}^{-1}}$ has non-zero elements (Ex. 1). We describe the relationship between $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ and $(\mathbf{I}_d - \boldsymbol{\mathcal{A}})$ for a DAG via *structural equivalence*, and investigate its symmetries. Similar to the linear case (and shown more formally below), $\mathbf{J}_{\boldsymbol{f}^{-1}}$ and $(\mathbf{I}_d - \boldsymbol{\mathcal{A}})$ have the same sparsity structure, meaning $\forall i, j \ (\mathbf{J}_{\boldsymbol{f}^{-1}})_{ij} = 0 \Leftrightarrow (\mathbf{I}_d - \boldsymbol{\mathcal{A}})_{ij} = 0$. We denote this structural equivalence more formally as $\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$, with the full definition and its properties outlined in Defn. C.1. Ex. 1 intuits why our claim refers to $\boldsymbol{\mathcal{A}}$ and not $\boldsymbol{\mathcal{C}}$: in the matrix mapping from $\boldsymbol{X}$ to $\boldsymbol{N}$ only the edges (captured by $\boldsymbol{\mathcal{A}}$) are present.

3.3 Permutation indeterminacies

The inference model $\widehat{\boldsymbol{f}}^{-1}$ we learn from the observed data generally differs from the true inverse of $\boldsymbol{f}$ up to certain indeterminacies depending on the NLICA algorithm we use. Most commonly, this includes scaling, permutation, sign flips, and monotonic element-wise transformations (Hyvärinen et al., 2001; Khemakhem et al., 2020a; Zimmermann et al., 2021). While element-wise transformations such as scaling or sign-flips do not influence the sparsity structure of the Jacobian, permutations break our structural equivalence between the Jacobian and the ground-truth adjacency matrix. To this end, we note that with the right ordering(s)[2] the Jacobian $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ features a lower-triangular structure. The following lemma shows that this property determines the ordering of the noise variables such that they yield a lower-triangular Jacobian, i.e., all possible causal orderings that ensure structural equivalence to the ground-truth adjacency matrix (the proof is deferred to Appx. D.1):

Lemma 1. *[DAG DGPs resolve the permutation ambiguity of ICA] When $\boldsymbol{f}$ describes a DAG, then the permutation indeterminacy of ICA π_{ICA} can be accounted for such that the Jacobian of the inference network will have a lower-triangular Jacobian, even with unknown causal ordering π.*

Proof (Sketch). Given that the DGP is structured by a DAG, the adjacency matrix $\boldsymbol{\mathcal{A}}$ is lower triangular and Assum. A.2 ensures that diagonal elements are nonzero. The permutation indeterminacy of ICA (which

[2]The causal ordering does not need to be unique, e.g., in the DAG $X_i \leftarrow X_j \to X_k$ the nodes X_i and X_k are interchangeable.

is expressed as a left-multiplication, i.e., affects the rows) comprises matrices that do not violate lower-triangularity. This gives us a single permutation (for a unique causal ordering) or a set of permutations, each of which ensures a lower triangular $\boldsymbol{\mathcal{A}}$. □

We emphasize that Lem. 1 refers to *two permutations*: the permutation indeterminacy of ICA (Lem. 1 makes a claim about this) and the causal ordering of the SEM. These can be thought of as permuting the rows and columns of the inference model's Jacobian.

3.4 Main result

Relying on the properties of $\sim_{DAG}$ and Lem. 1, we prove that $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ can be used to extract the DAG for general nonlinear functions (akin to the linear case shown in Ex. 1; the proof is deferred to Appx. D.2):

Proposition 1. *[$\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$]The suitably permuted inference network Jacobian $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ is structurally equivalent to $(\mathbf{I}_D - \boldsymbol{\mathcal{A}})$ if $\boldsymbol{f}^{-1}$ is strongly identified (Khemakhem et al., 2020b, Def.1) up to scalings, sign flips, permutations, and zero-preserving transformations, when Assum. A.2 holds.*

Proof (Sketch). From the iterative formulation of the SEM in eq. (2), we note that $\boldsymbol{X}$ (or more precisely, $\boldsymbol{X}(\boldsymbol{N})$), is a fixpoint of $\boldsymbol{f}$. Thus, when we apply the chain rule to calculate $\mathbf{J}_{\boldsymbol{f}}$, we will only have two types of terms (on both sides), namely:

$$\mathbf{A} := \frac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{X}}\Big|_{\boldsymbol{X},\boldsymbol{N}};\ \mathbf{B} := \frac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{N}}\Big|_{\boldsymbol{X},\boldsymbol{N}}. \tag{3}$$

This expression leads us to a closed form of $\mathbf{J}_{\boldsymbol{f}}$. Then we apply the inverse function theorem at $(\boldsymbol{X}, \boldsymbol{N})$ to get $\mathbf{J}_{\boldsymbol{f}^{-1}}$. As the last step, we incorporate the indeterminacies—coming from strong identifiability—and show based on the properties of $\sim_{DAG}$ that the statement of the proposition holds. □

Prop. 1 implies that we can extract the DAG when $\boldsymbol{f}^{-1}$ can be strongly identified (Khemakhem et al., 2020b, Def.1)—i.e., we can reason about interventions (cf. § 2). We note that if $\mathbf{B} = \mathbf{I}_d$, then (19) describes Additive Noise Models (ANMs) (Hoyer et al., 2008), whereas when additionally $\mathbf{A}$ is constant, we recover LiNGAM (Shimizu et al., 2006).

3.5 Algorithm for CD and determining π

Based on Lem. 1 and Prop. 1, we propose a two-step approach for extracting the DAG from observational data for general nonlinear $\boldsymbol{f}$:

1. First, we use a suitable nonlinear ICA algorithm to estimate $\boldsymbol{f}^{-1}$ up to permutations and zero-preserving element-wise nonlinearities with an inference model $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$.
2. Second, we resolve the permutation indeterminacy by accounting for the causal graph structure.

Regarding the second step, we learn the permutations after training with an objective that enforces the estimated Jacobian to be lower-triangular. To this end, we need to learn both a permutation π for the causal ordering as well as a permutation π_{ICA} that resolves the indeterminacy in the noise variables introduced by ICA. We use the permuted absolute Jacobian $\mathbf{K}$ defined as

$$\mathbf{K} := \left|\mathbf{S}_{\text{ICA}} \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \mathbf{S}_{\pi}\right| \tag{4}$$

where $\mathbf{S}_{\text{ICA}}, \mathbf{S}_{\pi}$ are doubly-stochastic matrices that represent a soft permutation on both noise and observation variables, which we learn after ICA training. We then introduce a training loss inspired by LiNGAM (Shimizu et al., 2006) that encourages $\mathbf{K}$ to be lower-triangular by simultaneously maximizing i) the sum of the main diagonal and ii) the lower-triangular part, while also iii) minimizing the stricly-upper triangular part of $\mathbf{K}$,

$$\mathcal{L}_{\pi} = \sum_{i,j} \left[\alpha_d \left(\mathbf{K}\right)_{ii}^{-1} + \alpha_u \left(\mathbf{K}\right)_{i<j} - \alpha_l \left(\mathbf{K}\right)_{i\geq j}\right]. \tag{5}$$

The full learning algorithm is presented in Alg. 1.

Algorithm 1 Algorithm for multivariable CD and determining the causal order π

Input: dataset D, network parameters θ, Sinkhorn networks $\mathbf{S}_{\text{ICA}}, \mathbf{S}_\pi$, contrastive loss $\mathcal{L}_{CL}$
Initialize θ
while $\mathcal{L}_{CL}$ not converged **do**
  calculate $\mathcal{L}_{CL}$ for a batch from D
  update θ
end while
extract $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$
while $\mathcal{L}_\pi$ not converged **do**
  $\mathbf{K} = \left|\mathbf{S}_{\text{ICA}}\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}\mathbf{S}\right|$
  $\mathcal{L}_\pi = \sum_{i,j}\left[\alpha_d\left(\mathbf{K}\right)_{ii}^{-1} + \alpha_u\left(\mathbf{K}\right)_{i<j} - \alpha_l\left(\mathbf{K}\right)_{i\geq j}\right]$
  update $\mathbf{S}_{\text{ICA}}, \mathbf{S}_\pi$
end while

4 Experiments

4.1 Experimental setup.

Data Generating Process (DGP) We experiment with three DGPs: i) linear and ii) nonlinear SEMs (in the form of $\boldsymbol{X} = \boldsymbol{f}(\mathbf{W}\boldsymbol{N})$, as well as with iii) Multi-Layer Perceptrons (MLPs) with triangular weight matrices (as used in (Monti et al., 2019)). In all cases, the nonlinear activations (i.e., $\boldsymbol{f}$) are leaky ReLUs (with a slope of 0.25 for the SEMs and 0.1 for the triangular MLPs). For the SEM DGPs, we exlore three options: a) no permutation w.r.t. the causal ordering (i.e., only the ICA permutation remains); b) a sparse DGP (with each $X_i - X_j$ edge being nonzero with a 0.25 probability); and c) permuted causal ordering (with dense $\boldsymbol{\mathcal{A}}$). Additionally, we ensure that the ordering of N_i is unique (all cases), and that the DGP weights are $\gg 0$ (for the SEM DGPs) as otherwise we would be unable to distinguish weak connections from small elements in the Jacobian. That is, the estimate of a weak connection could be the same order of magnitude as the estimate of a zero element due to the stochasticity of training—we do not enforce this property for the triangular MLPs to compare to the results of (Monti et al., 2019), where such modification was not present. For the *permuted* SEM DGPs, we sample 6 different orderings and 5 seeds for each problem dimensionality $\{3; 5; 8; 10\}$—the number of seeds is 10 for non-permuted and sparse SEMs. For the triangular MLP, we use $d = 6$ and 5 seeds to compare to (Monti et al., 2019, Fig. 2) and vary the number of layers in the mixing. To use contrastive NLICA for training the inference model, the DGPs needs to satisfy the assumptions underlying the proof of identifiability (Zimmermann et al., 2021, Thm. 6)): the latent space is a hyperrectangle in $\mathbb{R}^d$, the marginal $p(\boldsymbol{N})$ is uniform, the conditional $p(\tilde{\boldsymbol{N}}|\boldsymbol{N})$ is Laplace, $\boldsymbol{X}$ is generated by a smooth and bijective mapping;

Inference model To (strongly) identify the SEM, we use contrastive NLICA (Zimmermann et al., 2021) to estimate $\widehat{\boldsymbol{f}}^{-1}$ with a hyperrectangle latent space in $\mathbb{R}^d$ and the contrastive loss uses the same metric as the conditional, which is L_1 for our case (Assum. E.1). Our architecture for the inference model is the same MLP as in (Zimmermann et al., 2021) (Tab. 5). To account for the permutation indeterminacies, we use two Sinkhorn networks (Mena et al., 2018), which are differentiable models for learning doubly-stochastic matrices. We observed that setting the lowest $d\left(d-1\right)/2$ elements (for dense DAGs) to zero and converting the resulting $\mathbf{K}$ matrix to binary often helped the convergence of the Sinkhorn networks. Moreover, instead using max to aggregate the different Jacobians over the batch, we found using the mean operator more stable in practice.

Metrics We measure learning the correct ordering by the ordering accuracy (Acc_π, only for the permuted case)—i.e., the ratio of causal variable pairs $\forall i < j : (N_i, N_j)$, such that the ranking (expressed by $\text{sign}\left(i-j\right)$) matches that in the inferred (permuted) ordering π, i.e., $\text{sign}\left(\pi\left(i\right) - \pi\left(j\right)\right)$. To normalize, we divide by the number of distinct edge pairs ($1/2(d-1)d$) We also report the accuracy (Acc) and the Structural Hamming Distance (SHD) (we use 1e−3 as the threshold in all scenarios) for inferring the edges of the DAG, as is

Table 1: Validation of Lem. 1 for linear and nonlinear SEMs with **unknown causal ordering** to measure how well our method recovers the causal ordering. Mean Correlation Coefficient (MCC) measures identifiability, $|\mathcal{E}^*|$ is the maximum number of edges in a DAG, Acc_π is the accuracy of recovering the pairwise causal ordering π, whereas π gives the ratio of learning a (any) paermutation in $\mathbf{S}_\pi$

		Linear				Nonlinear			
$\lvert\mathcal{E}^*\rvert$	d	MCC	Acc	Acc_π	π	MCC	Acc	Acc_π	π
6	3	1.	1.	1.	1.	1.	1.	1.	1.
15	5	$0.989_{\pm 0.039}$	$0.998_{\pm 0.009}$	$0.974_{\pm 0.078}$	0.76	$0.988_{\pm 0.039}$	$0.994_{\pm 0.021}$	$0.957_{\pm 0.129}$	0.583
36	8	$0.834_{\pm 0.238}$	$0.935_{\pm 0.081}$	$0.851_{\pm 0.183}$	0.414	$0.781_{\pm 0.219}$	$0.934_{\pm 0.051}$	$0.889_{\pm 0.15}$	0.345
55	10	$0.852_{\pm 0.251}$	$0.931_{\pm 0.051}$	$0.921_{\pm 0.147}$	0.233	$0.794_{\pm 0.255}$	$0.924_{\pm 0.073}$	$0.739_{\pm 0.252}$	0.276

Table 2: Causal discovery performance for linear and nonlinear SEMs with **known causal ordering**. Mean Correlation Coefficient (MCC) measures identifiability, $|\mathcal{E}^*|$ is the maximum number of edges in a DAG, Acc is accuracy, Ours is our proposal, HSIC refers to using HSIC independence tests, and SHD is the Structural Hamming Distance

		Linear				Nonlinear			
$\lvert\mathcal{E}^*\rvert$	d	MCC	Acc(Ours)	Acc(HSIC)	SHD	MCC	Acc(Ours)	Acc(HSIC)	SHD
6	3	1.	1.	$0.7_{\pm 0.1}$	0.	1.	1.	$0.741_{\pm 0.105}$	$0.049_{\pm 0.14}$
15	5	$0.969_{\pm 0.066}$	$0.928_{\pm 0.131}$	$0.828_{\pm 0.116}$	$0.072_{\pm 0.131}$	$0.94_{\pm 0.09}$	$0.858_{\pm 0.172}$	$0.8_{\pm 0.102}$	$0.142_{\pm 0.171}$
36	8	1.	1.	$0.682_{\pm 0.17}$	0.	$0.982_{\pm 0.029}$	$0.872_{\pm 0.198}$	$0.823_{\pm 0.142}$	$0.128_{\pm 0.198}$
55	10	$0.965_{\pm 0.03}$	$0.832_{\pm 0.176}$	$0.551_{\pm 0.003}$	$0.168_{\pm 0.176}$	$0.962_{\pm 0.025}$	$0.636_{\pm 0.239}$	$0.638_{\pm 0.134}$	$0.364_{\pm 0.239}$

standard practice in the literature (Lachapelle et al., 2020; Monti et al., 2019; Ke et al., 2020; Vowels et al., 2022).

Comparison We use the linear and nonlinear SEM DGPs to showcase that our method can infer the DAG while also learning the correct ordering. Then, we compare to NonSENS (Monti et al., 2019), which, unlike our proposal, does CD on an edge-by-edge basis. Thus, the causal ordering π does not affect how NonSENS operates. We use the HSIC independence test (Gretton et al., 2005) on top of contrastive NLICA (Zimmermann et al., 2021) to provide a close comparison with NonSENS (Monti et al., 2019). Notably, since our assumptions provide identifiability up to generalized permutations, there is no need to perform linear ICA on top of contrastive NLICA. Thus, we test independence between the observations X_i and the *inferred* noise variables $\hat{N}_j$—although the number of tests is d^2, we use a Bonferroni correction factor of 4, since each edge is determined based on four tests (Monti et al., 2019).

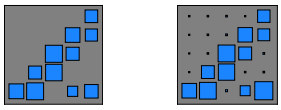

Figure 2: Hinton diagrams ($d = 5$): ground truth (left), estimate (right). Size equals magnitude.

4.2 Results.

In all experiments except those in Tab. 1, we used the output of the matching problem as an oracle (solved via the Hungarian algorithm (Kuhn, 1955)) to correct for the permutation indeterminacy of ICA.

The permutation indeterminacies can be resolved (verifying Lem. 1). Tab. 1 corroborates the result of Lem. 1: it is possible to resolve the permutation indeterminacy by assuming a DAG DGP. However, Acc_π strongly depends on the performance of NLICA, measured by Mean Correlation Coefficient (MCC). As MCC deteriorates, the correct causal ordering cannot be recovered. Nonetheless, erroneous solutions resulting from training stochasticity (the most frequent problem according to our observations) can be simply filtered out: in this case the doubly stochastic matrices usually do not converge to a permutation matrix. Inspecting their elements or automatically rejecting such solutions is straightforward. Thus, we report two quantities in Tab. 1: Acc_π is the ratio of inferring the order of causal variable pairs *when the Sinkhorn networks converged*

to permutation matrices; π (with a slight abuse of notation), on the other hand, reports the ratio of the successful attempts to recover permutation matrices. Clearly, failing to converge to a permutation matrix is the bottleneck of this step, since despite failing to scalably recover π, in case of converging to a permutation matrix the captured graph reflects most of the edges. This is reported by the Acc_π column that is calculated after applying the learned (not necessarily correct) permutations.

Competitive performance on linear and nonlinear SEMs. Tab. 2 demonstrates (with π being known) that our method outperforms HSIC in the linear case and is at least comparable to HSIC in the nonlinear case—note that the entries in $\mathbf{J}_{\widehat{f}^{-1}}$ were ordered by absolute value and the smallest ones were zeroed out—namely, these are the elements of the Jacobian that most probably correspond to the zeros in the true Jacobian. However, this modification might require additional knowledge about the sparsity of the DAG. Fig. 3 describes how precision and recall changes for threshold values from [1e−7; 1e0] for *sparse* DAGs. Notably, the nonlinear curves are better than the linear ones. For additional results on sparse SEMs (Tab. 6) and SEMs with unknown causal ordering (Tab. 7, evaluation is done after accounting for the causal ordering), cf. Appx. E. For sparse SEMs, HSIC is slightly better for larger graphs, whereas in the case of unknown causal ordering, our proposal has better accuracy in most cases. To visualize the inferred graph structure, we plot a Hinton diagram of the true and estimated Jacobians in Fig. 2, showing that $\mathbf{J}_{\widehat{f}^{-1}}$ can capture the edges of an underlying *sparse* DAG.

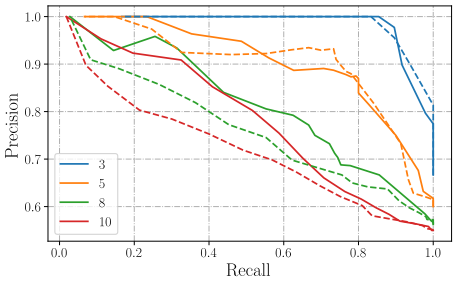

Figure 3: Precision vs recall for thresholds in [1e−7; 1e0] for linear (dashed) and nonlinear (solid) *sparse* SEMs

Competitive performance on triangular MLPs from (Monti et al., 2019). Tab. 3 summarizes our results with the triangular MLP of (Monti et al., 2019). Despite having small weights (appr. 1e−2) in the ground truth Jacobian $\mathbf{J}_{f^{-1}}$, our method was able to infer most edges in the DAG. Importantly, the resulting accuracies are larger than those of our adapted version of NonSENS (Monti et al., 2019). Moreover, our method has the advantage of simultaneously inferring all edges based on the structure of $\mathbf{J}_{\widehat{f}^{-1}}$—thus, it does not require d^2 pairwise independence test for a DAG with d nodes. Our application of HSIC independence tests resulted in surprisingly low accuracy, despite utilizing an NLICA method with identifiability guarantees up to generalized permutations. Interestingly, HSIC resulted in (close-to) chance-level performance in our repeated experiments—by careful inspection of the DGP, we found that the weights are in the order of 1e−4, which might explain such bad performance. As noted above, since Monti et al. (2019) did not constrain the weights, we used a uniform initialization scheme, which might led to mismatching experimental conditions. Though the use of HSIC was inspired by NonSENS (Monti et al., 2019), since we made different assumptions on the DGP, the results only represent HSIC's (but not NonSENS's) performance.

Table 3: Causal discovery performance for the triangular MLP from (Monti et al., 2019) with $d = 6$. $|l|$ denotes the number of MLP layers, Acc the accuracy, Ours is our proposal, HSIC refers to using HSIC independence tests. Chance level is (for the dense MLP) $21/36 = 0.583$

$\|l\|$	MCC	Acc (Ours)	Acc (HSIC)
1	1.	$0.933_{\pm 0.042}$	0.583
2	1.	0.944	0.583
3	$0.997_{\pm 0.003}$	1.	0.583
4	$0.978_{\pm 0.016}$	$0.922_{\pm 0.097}$	$0.6_{\pm 0.033}$
5	$0.603_{\pm 0.062}$	$0.711_{\pm 0.054}$	$0.589_{\pm 0.011}$

5 Related work

Independence tests for CD. Traditional CD methods (Pearl, 2009b; Spirtes & Zhang, 2016; Spirtes et al., 2000; Peters et al., 2017) rely on statistical (conditional) independence tests to infer the graph structure. Recent works also leverage such tests (Shimizu et al., 2006; Monti et al., 2019; Guo et al., 2022; Karlsson & Krijthe, 2022) to uncover hidden confounders (Karlsson & Krijthe, 2022), for bivariate (Janzing et al., 2009; Monti et al., 2019) or multivariable CD (Guo et al., 2022) for nonlinear SEMs. LiNGAM (Shimizu et al., 2006), which inspired our work, also relies on independence tests to prune edges. Although independence

tests provide additional information via significance values, they are not differentiable and can be costly, as d latents require d^2 tests.

Optimization-based CD. Zheng et al. (2018) introduced the continuous optimization-based NOTEARS algorithm for linear SEMs, which has inspired further research (Khemakhem et al., 2021; Lorch et al., 2021; Ng et al., 2019; Schölkopf et al., 2021; Yu et al., 2019; Lachapelle et al., 2020; Kalainathan et al., 2018) to provide differentiable methods for CD in neural networks. Most of the differentiable solutions (Khemakhem et al., 2021; Ng et al., 2019; Schölkopf et al., 2021; Yu et al., 2019) constrain the function class, some of them (Lachapelle et al., 2020; Kalainathan et al., 2018) both the function class and the data distribution.

Using the adjacency matrix $\boldsymbol{\mathcal{A}}$. Several methods leverage the adjacency matrix for CD: Shimizu et al. (2006) use a weight matrix to infer the DAG in the linear case, Zheng et al. (2018) use $\boldsymbol{\mathcal{A}}$ as a regularizer in NOTEARS, Ng et al. (2019) reformulates the SEM with an adjacency matrix for additive models, Schölkopf et al. (2021) models $\boldsymbol{\mathcal{A}}$ with an LSTM in a variational framework, whereas Lachapelle et al. (2020) calculates the Jacobian of the inference network to enforce acyclicity, generalizing to nonlinear additive models. Rolland et al. (2022) consider the same model class as Lachapelle et al. (2020), but they rely on the Jacobian of the score function. In (Brouillard et al., 2020), $\boldsymbol{\mathcal{A}}$ appears for the interventional case. Our work reasons about the Jacobian of the inference network, but it extends to general nonlinear functions for observational data.

CD from interventions. Many algorithms can incorporate interventions (Brouillard et al., 2020; Ke et al., 2020; Lippe et al., 2021; 2022; Lorch et al., 2021). Interestingly, (Ke et al., 2020) provide an extension of (Yu et al., 2019; Zheng et al., 2018) to interventional data, and of the bivariate method of (Bengio et al., 2020) to a multivariable one. It is remarkably similar to our proposal, as both make assumptions only on the data (i.e., admitting general nonlinear functional relationships), but as (Ke et al., 2020) requires interventions, its path is orthogonal to ours. So is the work of (Lippe et al., 2021), which removes any requirement on the data, scales to multiple variables, but requires interventions.
We provide a detailed comparison the properties of related CD methods in Tab. 4 in Appx. B.

6 Discussion

Limitations. Our theory requires the guarantees of strong identifiability but not the use of a specific (NLICA) algorithm. Though our experiments demonstrate that fulfilling strong identifiability is sufficient for CD, we do not vary the NLICA algorithm. Our method's applicability is limited for inferring weak edges, similar to (Shimizu et al., 2006; Tashiro et al., 2014; Shahbazinia et al., 2021; Lachapelle et al., 2020). As demonstrated in § 4, despite its competitive performance, the success of our proposed method highly relies on the performance of NLICA, which can be limited for higher-dimensional problems. Nonetheless, based on our comparisons, this seems to be an issue for the HSIC independence test as well. A possible explanation could be that the class of SEMs is harder to learn with specific NLICA algorithms; indeed, we observed that deploying contrastive NLICA (Zimmermann et al., 2021) achieves much better MCC on general (non-triangular) invertible MLPs. To ensure that particular entries in the Jacobian are non-zero everywhere, our assumptions require that the underlying DAG for the DGP is the same for all data points, which might be restrictive . For instance, if the DAG models the interaction of physical objects, then cause-effect relationships are only present when, e.g., the objects are touching each other or their magnetic/electric fields affect each other—in the literature, this setting is considered in (Sontakke et al., 2021; Seitzer et al., 2021).

Unknown causal ordering. Accounting for the causal ordering is, to the best of our knowledge, only found in (Shimizu et al., 2006). Binary CD methods such as (Monti et al., 2019) alleviate this step as they work on an edge-by-edge basis. Other non-ICA-based methods can also avoid this step since the DAG is *invariant* to changes in the causal ordering—meaning that reordering X_i in the observation vector $\boldsymbol{X}$ (cf. Defn. A.7) does not affect the edges of the graph, only their representation in form of an adjacency matrix. However, to resolve the permutation indeterminacy of ICA, we need to account for the causal ordering, since only then can the Jacobian be lower-triangular. Although extracting a lower-triangular Jacobian is easier to interpret and potentially better suited, e.g., as a building block of causal representation learning (since the causal ordering of N_i is always the same), our method extracts the DAG even without resolving these indeterminacies. That is, our demonstration that the permutation indeterminacies can be resolved should mostly be considered as corroboration of Lem. 1.

Extensions to related work. Using neural networks for CD is discussed in several papers (Monti et al., 2019; Khemakhem et al., 2021; Lachapelle et al., 2020; Lippe et al., 2021; 2022; Brouillard et al., 2020), many of them uses the adjacency matrix, the Jacobian of the inference network (Shimizu et al., 2006; Schölkopf et al., 2021; Lachapelle et al., 2020) or that of the score function Rolland et al. (2022). Furthermore, methods that can handle general nonlinear relationships either require interventions (Brouillard et al., 2020; Lippe et al., 2021; 2022) or rely on independence tests (Guo et al., 2022; Monti et al., 2019). Our method was inspired by LiNGAM (Shimizu et al., 2006) to use the Jacobian of the inference network for inferring the DAG and utilizes NLICA (similar to (Monti et al., 2019)) to provide theoretical guarantees (Prop. 1) for multivariable CD with arbitrary nonlinear functions. Furthermore, we also prove (Lem. 1) and demonstrate (Tab. 1) that the permutation indeterminacy of ICA—and that of an unknown causal ordering—can be resolved in the nonlinear case.

Conclusion. We introduced a two-step process to leverage strong identifiability for inferring the DAG of multivariable causal models with general nonlinear functions. Our method uses the Jacobian of the inference function (mapping from observables to independent variables) and can be thought as a generalization of LiNGAM to the nonlinear case. We prove that this Jacobian captures the sparsity structure of the DAG, and show that by working with causal models, we can resolve the permutation indeterminacy of ICA under certain assumptions. Since we do not use conditional independence tests, but learn the causal ordering with Sinkhorn networks, our method provides an end-to-end solution for CD and avoids the cost of exponentially many independence tests. We experimentally demonstrate that our proposal can infer the DAG in multiple synthetic data sets.

References

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Nan Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher J. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 9

Philippe Brouillard, Sébastien Lachapelle, Alexandre Lacoste, Simon Lacoste-Julien, and Alexandre Drouin. Differentiable causal discovery from interventional data. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 3, 9, 10, 16

Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. 1, 3

George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, pp. 231, 1951. 3

Frederick Eberhardt, Clark Glymour, and Richard Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv:1207.1389 [cs, stat]*, 2012. arXiv: 1207.1389. 1

Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. Deconvolution with unknown noise distribution is possible for multivariate signals. *Ann. Statist.*, 50(1), February 2022. ISSN 0090-5364. doi: 10.1214/21-aos2106. URL `https://doi.org/10.1214/21-aos2106`. 3

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020. 1

Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and interpretable models in computer vision and machine learning*, pp. 39–80. Springer, 2018. 3

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. PMLR, July 2019. URL `https://proceedings.mlr.press/v115/gresele20a.html`. 2, 3

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanisms analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 28233–28248. Curran Associates, Inc., December 2021. URL https://proceedings.neurips.cc/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf. 3

Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In Sanjay Jain, Hans Ulrich Simon, and Etsuji Tomita (eds.), *Algorithmic Learning Theory*, Lecture Notes in Computer Science, pp. 63–77, Berlin, Heidelberg, 2005. Springer. ISBN 978-3-540-31696-1. doi: 10.1007/11564089_7. 7

Siyuan Guo, Viktor Tóth, Bernhard Schölkopf, and Ferenc Huszár. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *ArXiv preprint*, abs/2203.15756, 2022. 1, 3, 8, 10, 16

Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pp. 939–948. AUAI Press, 2020. 2, 3

Patrik O. Hoyer, Dominik Janzing, Joris M. Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou (eds.), *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008*, pp. 689–696. Curran Associates, Inc., 2008. 1, 3, 5

Antti Hyttinen, Patrik O Hoyer, Frederick Eberhardt, and Matti Järvisalo. Discovering cyclic causal models with latent variables: a general sat-based procedure. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pp. 301–310, 2013. 3

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3765–3773, 2016. 2, 3

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In Aarti Singh and Xiaojin (Jerry) Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 2017. 2, 3

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/s0893-6080(98)00140-3. URL https://doi.org/10.1016/s0893-6080(98)00140-3. 3

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., New York, May 2001. ISBN 047140540X, 0471221317. doi: 10.1002/0471221317. URL https://doi.org/10.1002/0471221317. 1, 3, 4

Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019. 2, 3

Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. pp. 23, 2010. 2, 3

Dominik Janzing, Jonas Peters, Joris Mooij, and Bernhard Schölkopf. Identifying confounders using additive noise models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 249–257, 2009. 8

Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *ArXiv preprint*, abs/1803.04929, 2018. 3, 9, 16

Rickard K. A. Karlsson and Jesse H. Krijthe. Combining observational datasets from multiple environments to detect hidden confounding, May 2022. URL `http://arxiv.org/abs/2205.13935`. arXiv:2205.13935 [cs, stat]. 8

Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C. Mozer, Chris Pal, and Yoshua Bengio. Learning Neural Causal Models from Unknown Interventions. *arXiv:1910.01075 [cs, stat]*, August 2020. URL `http://arxiv.org/abs/1910.01075`. arXiv: 1910.01075. 3, 7, 9, 16

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Silvia Chiappa and Roberto Calandra (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020a. 2, 3, 4

Ilyes Khemakhem, Ricardo Pio Monti, Diederik P. Kingma, and Aapo Hyvärinen. ICE-BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. 2, 5, 17

Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3520–3528. PMLR, 2021. 3, 9, 10, 16

David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 3

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2 (1-2):83–97, 1955. 7

Trent Kyono, Yao Zhang, and Mihaela van der Schaar. Castle: Regularization via auxiliary causal graph discovery. *Advances in Neural Information Processing Systems*, 33:1501–1512, 2020. 3

Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural DAG learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 1, 3, 7, 9, 10, 16

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pp. 428–484. PMLR, 2022. 2

Hao-Chih Lee, Matteo Danieletto, Riccardo Miotto, Sarah T Cherng, and Joel T Dudley. Scaling structural learning with no-bears to infer causal transcriptome networks. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pp. 391–402. World Scientific, 2019. 3

Phillip Lippe, Taco Cohen, and Efstratios Gavves. Efficient neural causal discovery without acyclicity constraints. *ArXiv preprint*, abs/2107.10483, 2021. 3, 9, 10, 16

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences, June 2022. URL `http://arxiv.org/abs/2202.03169`. Number: arXiv:2202.03169 arXiv:2202.03169 [cs, stat]. 9, 10

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. 3

Lars Lorch, Jonas Rothfuss, Bernhard Schölkopf, and Andreas Krause. DiBS: Differentiable Bayesian structure learning. *ArXiv preprint*, abs/2105.11839, 2021. 2, 9

Gonzalo E. Mena, David Belanger, Scott W. Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 6

Jovana Mitrovic, Dino Sejdinovic, and Yee Whye Teh. Causal inference via kernel deviance measures. *Advances in neural information processing systems*, 31, 2018. 3

Ricardo Pio Monti, Kun Zhang, and Aapo Hyvärinen. Causal discovery with general non-linear relationships using non-linear ICA. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pp. 186–195. AUAI Press, 2019. 1, 2, 3, 6, 7, 8, 9, 10, 16

Raha Moraffah, Bahman Moraffah, Mansooreh Karami, Adrienne Raglin, and Huan Liu. Causal adversarial network for learning conditional and interventional distributions. *ArXiv preprint*, abs/2008.11376, 2020. 3

Hiroshi Morioka, Hermanni Hälvä, and Aapo Hyvärinen. Independent innovation analysis for nonlinear vector autoregressive process. In Arindam Banerjee and Kenji Fukumizu (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1549–1557. PMLR, 2021. 2, 3

Ignavier Ng, Zhuangyan Fang, Shengyu Zhu, Zhitang Chen, and Jun Wang. Masked gradient-based causal structure learning. *ArXiv preprint*, abs/1910.08527, 2019. 1, 3, 9, 16

Ignavier Ng, AmirEmad Ghassami, and Kun Zhang. On the role of sparsity and dag constraints for learning linear dags. *Advances in Neural Information Processing Systems*, 33:17943–17954, 2020. 3

Judea Pearl. Causal inference in statistics: An overview. *Stat. Surv.*, 3(none), January 2009a. ISSN 1935-7516. doi: 10.1214/09-ss057. URL `https://doi.org/10.1214/09-ss057`. 1

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, 2 edition, September 2009b. ISBN 9780511803161, 9780521895606, 9780521749190. doi: 10.1017/cbo9780511803161. URL `https://doi.org/10.1017/cbo9780511803161`. 3, 8

J Peters, J Mooij, D Janzing, and B Schölkopf. Identifiability of causal graphs using functional models. In FG Cozman and A Pfeffer (eds.), *27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*, pp. 589–598, Corvallis, OR, USA, 2011. AUAI Press. 1, 3

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. 1, 2, 3, 8, 15

Garvesh Raskutti and Caroline Uhler. Learning directed acyclic graph models based on sparsest permutations. *Stat*, 7(1):e183, 2018. 3

Paul Rolland, Volkan Cevher, Matthäus Kleindessner, Chris Russel, Bernhard Schölkopf, Dominik Janzing, and Francesco Locatello. Score matching enables causal discovery of nonlinear additive noise models. *ArXiv preprint*, abs/2203.04413, 2022. 9, 10, 16

Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress, 2012. 1

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proc. IEEE*, 109(5):612–634, May 2021. ISSN 0018-9219, 1558-2256. doi: 10.1109/jproc.2021.3058954. URL `https://doi.org/10.1109/jproc.2021.3058954`. 1, 3, 9, 10, 16

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. *Advances in Neural Information Processing Systems*, 34:22905–22918, 2021. 9

Amirhossein Shahbazinia, Saber Salehkaleybar, and Matin Hashemi. ParaLiNGAM: Parallel causal structure learning for linear non-gaussian acyclic models. *arXiv: Distributed, Parallel, and Cluster Computing*, 2021. DOI:, MAG ID: 3202634766. 1, 9

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning. *ArXiv preprint*, abs/2010.02637, 2020. 1, 3, 16

Shohei Shimizu, Patrik O Hoyer, Aapo Hyvärinen, and Antti Kerminen. A linear non-gaussian acyclic model for causal discovery. pp. 28, 2006. 1, 2, 3, 5, 8, 9, 10, 16

Sumedh A. Sontakke, Arash Mehrjou, Laurent Itti, and Bernhard Schölkopf. Causal curiosity: RL agents discovering self-supervised experiments for causal representation learning. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9848–9858. PMLR, 2021. 9

Peter Spirtes and Kun Zhang. Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, 3(1):3, February 2016. ISSN 2196-0089. doi: 10.1186/s40535-016-0018-x. URL `https://doi.org/10.1186/s40535-016-0018-x`. 8

Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 3, 8

Mikhail Fedorovich Subbotin. On the law of frequency of error. *Mat. Sb.*, 31(2):296–301, 1923. 19

Tatsuya Tashiro, Shohei Shimizu, Aapo Hyvärinen, and Takashi Washio. ParceLiNGAM: A causal ordering method robust against latent confounders. *Neural Comput.*, 26(1):57–83, January 2014. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco_a_00533. URL `https://doi.org/10.1162/neco_a_00533`. 1, 9

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *ArXiv preprint*, abs/2106.04619, 2021. 3

Matthew J. Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like DAGs? a survey on structure learning and causal discovery. *ACM Comput. Surv.*, abs/2103.02582, April 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3527154. URL `https://doi.org/10.1145/3527154`. 3, 7

Dennis Wei, Tian Gao, and Yue Yu. Dags with no fears: A closer look at continuous optimization for learning bayesian networks. *Advances in Neural Information Processing Systems*, 33:3895–3906, 2020. 3

Matthew Willetts and Brooks Paige. I Don't Need $\mathbf{u}$: Identifiable Non-Linear ICA Without Side Information. *ArXiv preprint*, abs/2106.05238, 2021. 2

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. *ArXiv preprint*, abs/2004.08697, 2020. 3

Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG}-{GNN: Dag structure learning with graph neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7154–7163. PMLR, 2019. 1, 3, 9, 16

Kun Zhang and Aapo Hyvärinen. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv:1205.2599 [cs, stat]*, 2012. arXiv: 1205.2599. 3

Kun Zhang, Jiji Zhang, and Bernhard Schölkopf. Distinguishing cause from effect based on exogeneity. *ArXiv preprint*, abs/1504.05651, 2015. 1, 3

Xun Zheng, Bryon Aragam, Pradeep Ravikumar, and Eric P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 9492–9503, 2018. 1, 3, 9, 16

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12979–12990. PMLR, 2021. 1, 3, 4, 6, 7, 9, 18, 19

A SEMs

Definition A.1 (SEM)**.** *A SEM describes causal relationships via a set of structural assignments (Peters et al., 2017):*

$$X_i := f_i\left(\boldsymbol{Pa}_i, N_i\right), \qquad \forall i \in \mathcal{I} = \{, \ldots, d\}, \tag{6}$$

where X_i are the endogenous, N_i the exogenous/noise variables, $\boldsymbol{Pa}_i \subseteq \boldsymbol{X} \setminus \{X_i\}$ denotes the parent set of X_i, $\mathcal{I}$ the set of indices, and f_i the mappings.

Definition A.2 (Reduced form of SEM)**.** *The reduced form of the SEM expresses all X_i as a function of only the N_i variables, i.e.:*

$$X_i := f_i\left(\boldsymbol{N}^i\right), \qquad \forall i \in \mathcal{I} = \{0, \ldots, d-1\}, \tag{7}$$

with the same notation as in Defn. A.1, slightly abusing f_i and denoting a subset of $\boldsymbol{N}$ by $\boldsymbol{N}^i \subseteq \boldsymbol{N}$.

Definition A.3 (Causal ordering)**.** *The causal ordering π is a bijective automorphism on the index set $\mathcal{I}$. Namely, $\pi : \mathcal{I} \to \mathcal{I}$ so that $\forall X_i \neq X_j$, it holds that if $\pi(i) < \pi(j) \implies X_j \notin \boldsymbol{Pa}_i$.*

The definition means that only a node with a smaller index in π can be a parent of a node with a larger index. Note that though X_i *can* be a parent of X_j, it is not necessary, but X_j *cannot* be a parent of X_i. Multiple orderings may exist, e.g. if there are multiple X_i so that they only have a single parent. π helps to have a unique description of the edges in the graph. Namely, if the edges are organized in the adjacency matrix $\boldsymbol{\mathcal{A}}$ according to π, then $\boldsymbol{\mathcal{A}}$ will be strictly lower triangular.

Definition A.4 (Adjacency matrix)**.** *The adjacency matrix $\boldsymbol{\mathcal{A}}$ is a binary $d \times d$ matrix, where $\boldsymbol{\mathcal{A}}_{ij} = 1 \iff X_j \in \boldsymbol{Pa}_i$. The rows of $\boldsymbol{\mathcal{A}}$ are ordered by π; thus, $\boldsymbol{\mathcal{A}}$ is strictly lower-triangular.*

$\boldsymbol{\mathcal{A}}$ only describes the edges of the DAG, which gives the direct cause-effect relationships. Nodes can be influence each other via paths (i.e., a set of directed edges that can be traversed between the two nodes), which can be described by the connectivity matrix $\boldsymbol{\mathcal{C}}$

Definition A.5 (Connectivity matrix)**.** *The connectivity matrix $\boldsymbol{\mathcal{C}}$ is a binary $d \times d$ matrix, where $\boldsymbol{\mathcal{C}} = 1 \iff \exists p : X_j \to \cdots \to X_i$. $\boldsymbol{\mathcal{C}} = \sum_{k=1}^{d} \boldsymbol{\mathcal{A}}^k$. The rows of $\boldsymbol{\mathcal{C}}$ are ordered by π; thus, $\boldsymbol{\mathcal{C}}$ is strictly lower-triangular.*

Assumption A.1 (Structural faithfulness)**.** *The set of $\boldsymbol{N}$'s that induces additional zeroes (i.e., a sparser DAG) in the Jacobians $\mathbf{J}_{\boldsymbol{f}}$, $\mathbf{J}_{\boldsymbol{f}^{-1}}$ has zero measure, i.e., both Jacobians describe the sparsity structure of the underlying DAG DGP with probability one ($\mathbf{J}_{\boldsymbol{f}}$ w.r.t. $\boldsymbol{\mathcal{C}}$, as shown in Lem. A.1; $\mathbf{J}_{\boldsymbol{f}^{-1}}$ w.r.t. $\boldsymbol{\mathcal{A}}$). Alternatively, the structural independencies are reflected in a functional form via $\mathbf{J}_{\boldsymbol{f}}/\mathbf{J}_{\boldsymbol{f}^{-1}}$. We call this property structural faithfulness.*

Assumption A.2 (SEM assumptions)**.** *We assume that the causal DGP fulfils:*

(i) eq. (1) *describes a DAG;*

(ii) N_i *are jointly independent;*
(iii) There are no hidden confounders (faithfulness/stability); moreover, the Jacobians $\mathbf{J}_{\boldsymbol{f}}$, $\mathbf{J}_{\boldsymbol{f}^{-1}}$ *are structurally faithful (Assum. A.1);*
(iv) each f_i *is a homomorphism (they can be general nonlinear functions); and*
(v) each X_i *depend on* N_i *(i.e.,* $\frac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{N}}\big|_{\boldsymbol{X},\boldsymbol{N}}$ *is diagonal with non-zero elements)*

Definition A.6 (DGP with known π)**.** *The DGP is described by the SEM, when* π *is known. I.e., the flow of information is:* $\boldsymbol{N} \xrightarrow{SEM} \boldsymbol{X}$.

Definition A.7 (DGP with unknown π)**.** *The DGP with unknown* π *is given by the SEM, and by a permutation matrix* π *(with a slight abuse of notation) applied to* $\boldsymbol{X}$*. I.e., the flow of information is:* $\boldsymbol{N} \xrightarrow{SEM} \boldsymbol{X} \xrightarrow{\pi} \hat{\boldsymbol{X}}$.

Lemma A.1 ($\mathbf{J}_{\boldsymbol{f}} \sim_{DAG} (\mathbf{I}_d + \boldsymbol{\mathcal{C}})$)**.** *Given Assum. A.2, the partial derivatives of* f_i *w.r.t.* N_j *provide information about* $\boldsymbol{\mathcal{C}}$*, as*

$$(\mathbf{J}_{\boldsymbol{f}})_{kl} = \frac{\partial f_l}{\partial N_k} = 0 \iff \nexists X_k \to \cdots \to X_l$$

We emphasize that the derivatives are also non-zero in the case of indirect paths, i.e., when $\exists X_i \in p : i \neq k, l$*. Furthermore, the strictly lower triangular part of* $\mathbf{J}_{\boldsymbol{f}}$ *has the describes the same DAG as* $\boldsymbol{\mathcal{C}}$*–or equivalently,* $\mathbf{J}_{\boldsymbol{f}} \sim_{DAG} (\mathbf{I}_d + \boldsymbol{\mathcal{C}})$.

B Extended related work

Table 4: Comparison of CD methods. Column $\mathbf{x}$ indicates multivariability, $do(\emptyset)$ indicates whether the method can be applied only to observational data, $\boldsymbol{f}$ indicates constraints on the function class of the SEM, ∂/∂ indicates differentiability, and the data column lists restrictions on the data distribution.

Method	$\mathbf{x}$	$do(\emptyset)$	$\boldsymbol{f}$	∂/∂	Data
(Monti et al., 2019)	✗	✓	✓	✗	non-stationary
(Shimizu et al., 2006)	✓	✓	linear	✗	non-Gaussian
(Guo et al., 2022)	✓	✓	✓	✗	exchageability
(Khemakhem et al., 2021)	✓	✓	affine/additive	✓	✓
(Lachapelle et al., 2020)	✓	✓	additive	✓	Gaussian
(Brouillard et al., 2020)	✓	✗	✓	✓	✓
(Ke et al., 2020)	✓	✗	✓	✓	discrete
(Lippe et al., 2021)	✓	✗	✓	✓	✓
(Ng et al., 2019)	✓	✓	additive	✓	✓
(Schölkopf et al., 2021)	✓	✓	linear/additive	✓	✓
(Zheng et al., 2018)	✓	✓	linear	✓	✓
(Yu et al., 2019)	✓	✓	additive	✓	✓
(Shen et al., 2020)[3]	✓	✓	additive	✓	✓
(Kalainathan et al., 2018)	✓	✓	additive	✓	Gaussian
(Rolland et al., 2022)	✓	✓	additive	✓	✓
Ours	✓	✓	✓	✓	Assum. E.1

C Auxiliary theory

Definition C.1 ($\sim_{DAG}$)**.** *Two matrices* $\mathbf{S}, \mathbf{R}$ *are structurally equivalent if* $(\mathbf{S})_{ij} = 0 \iff (\mathbf{R})_{ij} = 0 : \forall i, j;$. *Structural equivalence, denoted as* $\sim_{DAG}$*, has the following properties (*$\circ$ *denotes composition):*

(i) **D-*invariance:*** *a non-singular diagonal matrix* $\mathbf{D}$ *preserves the sparsity structure; thus,* $(\mathbf{D} \circ \mathbf{S}) \sim_{DAG} \mathbf{S}$
(ii) h_0**-*invariance:*** *for zero-preserving transformations* $h_0 : (h_0(\mathbf{S}))_{ij} = 0 \iff (\mathbf{S})_{ij} = 0$ *then* $h(\mathbf{S}) \sim_{DAG} \mathbf{S}$

[3]Supervised

(iii) π***-equivariance:*** *a permutation π affects the positions of zeros; thus, both operands need to be permuted with the same π to maintain $\sim_{DAG}$, i.e.,* $\mathbf{S} \sim_{DAG} \mathbf{R} \iff (\pi \circ \mathbf{S}) \sim_{DAG} (\pi \circ \mathbf{R})$,
(iv) ***Transitivity:*** $\mathbf{S} \sim_{DAG} \mathbf{P} \wedge \mathbf{P} \sim_{DAG} \mathbf{R} \implies \mathbf{S} \sim_{DAG} \mathbf{R}$
(v) ***Commutativity:*** $\mathbf{S} \sim_{DAG} \mathbf{R} \iff \mathbf{R} \sim_{DAG} \mathbf{S}$.

D Proofs

D.1 Proof of Lem. 1

Lemma 1. *[DAG DGPs resolve the permutation ambiguity of ICA] When $\boldsymbol{f}$ describes a DAG, then the permutation indeterminacy of ICA π_{ICA} can be accounted for such that the Jacobian of the inference network will have a lower-triangular Jacobian, even with unknown causal ordering π.*

Proof. The unknown causal ordering π of N_i implies the right-multiplication of $\mathbf{J}_{\boldsymbol{f}^{-1}}$ with π^{-1}, the permutation indeterminacy of ICA the left-multiplication with π_{ICA}, yielding the estimated Jacobian $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$:

$$\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} = \pi_{\text{ICA}} \circ \mathbf{J}_{\boldsymbol{f}^{-1}} \circ \pi^{-1}, \tag{8}$$

where π_{ICA} and π^{-1} are not necessarily the same.
If π is **unique**, we only need to show that π_{ICA} is also unique. Assume that there exists $\pi_{\text{ICA},1} \neq \pi_{\text{ICA},2}$ such that $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ can be transformed into a lower-triangular $\mathbf{J}_{\boldsymbol{f}^{-1}}$ by both. This implies that the rows of $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ can be permuted such that it yields a lower-triangular $\mathbf{J}_{\boldsymbol{f}^{-1}}$ (when π is already accounted for). Assume that $\pi_{\text{ICA},1}$ yields a lower-triangular $\mathbf{J}_{\boldsymbol{f}^{-1}}$. Then a different $\pi_{\text{ICA},2}$ means that there are at least two rows i, k in $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ that can be permuted differently than in $\pi_{\text{ICA},1}$ such that the resulting matrix is still lower-triangular. $\mathbf{J}_{\boldsymbol{f}^{-1}}$ has a non-zero diagonal (cf. the definition of $\mathbf{B}$ in eq. (15)); thus, using a different ordering $\pi_{\text{ICA},2}$ will violate lower-triangularity, for this means that the i^{th}, k^{th} rows after applying $\pi_{\text{ICA},1}$ will be equal to the k^{th}, i^{th} rows of $\pi_{\text{ICA},2}$ (the former being equal to the true Jacobian $\mathbf{J}_{\boldsymbol{f}}$):

$$\left[\pi^{-1}_{\text{ICA},1} \circ \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \circ \pi\right]_{[i,k],:} = \left[\mathbf{J}_{\boldsymbol{f}^{-1}}\right]_{[i,k],:} = \left[\pi^{-1}_{\text{ICA},2} \circ \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \circ \pi\right]_{[k,i],:}, \tag{9}$$

which means that for $\pi_{\text{ICA},2}$ the resulting matrix has nonzero elements at indices (i, k) and (k, i). This violates lower-triangularity, since $k \neq i$, so one of the above means that there is at least one non-zero element above the main diagonal, leading to a contradiction.
If π is **not unique**, we can apply the above argument, resulting in a set of permutation matrices, each yielding a lower-triangular Jacobian. □

D.2 Proof of Prop. 1

Proposition 1. *[$\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$]The suitably permuted inference network Jacobian $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ is structurally equivalent to $(\mathbf{I}_D - \boldsymbol{\mathcal{A}})$ if $\boldsymbol{f}^{-1}$ is strongly identified (Khemakhem et al., 2020b, Def.1) up to scalings, sign flips, permutations, and zero-preserving transformations, when Assum. A.2 holds.*

Proof. We start from the functional equation of the SEM and note that if $\boldsymbol{X}$ is the input of $\boldsymbol{f}$ (as $\boldsymbol{Pa}_i$ from eq. (1)), then the output is the same $\boldsymbol{X}$ (which deterministically depends on $\boldsymbol{N}$):

$$\boldsymbol{X} = \boldsymbol{X}(\boldsymbol{N}) := \boldsymbol{f}(\boldsymbol{X}(\boldsymbol{N}), \boldsymbol{N}) = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{N}). \tag{10}$$

For a given $(\boldsymbol{X}, \boldsymbol{N})$ we can evaluate the Jacobian of $\boldsymbol{f}$ via the chain rule—the key point is that since $\boldsymbol{X}$ is a fix point of $\boldsymbol{f}$, $\mathbf{J}_{\boldsymbol{f}}$ will apprear on both sides (evaluated at the same point, expressed with the bar notation):

$$\mathbf{J}_{\boldsymbol{f}}\big|_{\boldsymbol{X},\boldsymbol{N}} = \tfrac{\partial \boldsymbol{X}(\boldsymbol{N})}{\partial \boldsymbol{N}}\big|_{\boldsymbol{X},\boldsymbol{N}} = \tfrac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{N}}\big|_{\boldsymbol{X},\boldsymbol{N}} = \mathbf{A}\tfrac{\partial \boldsymbol{X}}{\partial \boldsymbol{N}}\big|_{\boldsymbol{X},\boldsymbol{N}} + \mathbf{B} = \mathbf{A}\mathbf{J}_{\boldsymbol{f}}\big|_{\boldsymbol{X},\boldsymbol{N}} + \mathbf{B} \tag{11}$$

$$\mathbf{A} := \tfrac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{X}}\big|_{\boldsymbol{X},\boldsymbol{N}};\ \mathbf{B} := \tfrac{\partial \boldsymbol{f}(\boldsymbol{X},\boldsymbol{N})}{\partial \boldsymbol{N}}\big|_{\boldsymbol{X},\boldsymbol{N}}. \tag{12}$$

The above equation can be reordered to yield the expression for $\mathbf{J}_{\boldsymbol{f}}$ (note that $\mathbf{A}$, $\mathbf{B}$ depend on $\boldsymbol{X}$, $\boldsymbol{N}$):

$$\mathbf{J}_{\boldsymbol{f}}\big|_{\boldsymbol{X},\boldsymbol{N}} = (\mathbf{I}_d - \mathbf{A})^{-1}\mathbf{B}, \tag{13}$$

where $\mathbf{A}$ describes the $X_i - X_j$ edges in the DAG (i.e., $\mathbf{A} \sim_{DAG} \boldsymbol{\mathcal{A}}$), $\mathbf{B}$ is diagonal (as the $\boldsymbol{X}$ values are fixed). Since we reason about the Jacobian point-wise, we can invoke the inverse function theorem (by assumption, $\boldsymbol{f}$ is bijective) to express $\mathbf{J}_{\boldsymbol{f}^{-1}}$:

$$\mathbf{J}_{\boldsymbol{f}^{-1}} = \mathbf{J}_{\boldsymbol{f}}^{-1} = \mathbf{B}^{-1}\left(\mathbf{I}_d - \mathbf{A}\right). \tag{14}$$

$\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$ follows as $\mathbf{A} \sim_{DAG} \boldsymbol{\mathcal{A}}$ and $\mathbf{B}$ is diagonal (the invariance of $\sim_{DAG}$ follows from Def. C.1(i)). For proving that $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$, we need $\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ (Def. C.1(iv)), which requires us to account for all indeterminacies of strong identifiability: i) Def. C.1(i) accounts for scalings and sign flips; ii) Def. C.1(ii) for zero-preserving transformations; and iii) Def. C.1(iii) for permutations, which can be extracted as shown in Lem. 1.

□

Alternative proof

Proof. The proof consists of two steps: 1) leveraging the iterative formulation of the SEM (2), proving that $\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$ and 2) relying on the properties of $\sim_{DAG}$ and Lem. 1, showing $\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$.
We start by formulating $\mathbf{J}_{\boldsymbol{f}}$ (recall that $\boldsymbol{X} = \boldsymbol{X}^d$) based on the iterative SEM expression (2):

$$\mathbf{J}_{\boldsymbol{f}}\big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}} = \frac{\partial \boldsymbol{X}^d}{\partial \boldsymbol{N}}\Big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}} = \frac{\partial \boldsymbol{f}(\boldsymbol{X}^{d-1},\boldsymbol{N})}{\partial \boldsymbol{N}}\Big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}} = \mathbf{A}^{d-1}\frac{\partial \boldsymbol{X}^{d-1}}{\partial \boldsymbol{N}}\Big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}} + \mathbf{B}^{d-1} \tag{15}$$

$$\mathbf{A}^{d-1} := \frac{\partial \boldsymbol{f}(\boldsymbol{X}^{d-1},\boldsymbol{N})}{\partial \boldsymbol{X}^{d-1}}\Big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}};\ \mathbf{B}^{d-1} := \frac{\partial \boldsymbol{f}(\boldsymbol{X}^{d-1},\boldsymbol{N})}{\partial \boldsymbol{N}}\Big|_{\boldsymbol{X}^{d-1},\boldsymbol{N}}, \tag{16}$$

where $\mathbf{A}$ describes the $X_i - X_j$ edges in the DAG (i.e., $\mathbf{A} \sim_{DAG} \boldsymbol{\mathcal{A}}$), $\mathbf{B}$ is diagonal (as the $\boldsymbol{X}^{d-1}$ values are fixed). Although both $\mathbf{A}, \mathbf{B}$ are dependent from t (superscript), unless $\boldsymbol{f}$ is linear, they are independent when seen through the lens of structural equivalence. By Assum. A.1, it holds that $\mathbf{A}^k \sim_{DAG} \mathbf{A}^j \wedge \mathbf{B}^k \sim_{DAG} \mathbf{B}^j : \forall j, k$. Thus, we will omit superscripts for both.
Realizing that (15) gives us a recursive formula, and recalling that $\boldsymbol{X}^0 = \mathbf{0}$, we can unroll (15) iteratively for $t = d-1, d-2, \ldots, 0$:

$$\mathbf{J}_{\boldsymbol{f}} = \mathbf{A}\frac{\partial \boldsymbol{X}^{d-1}}{\partial \boldsymbol{N}} + \mathbf{B} \sim_{DAG} \mathbf{A}\left[\mathbf{A}\frac{\partial \boldsymbol{X}^{d-2}}{\partial \boldsymbol{N}} + \mathbf{B}\right] + \mathbf{B} \sim_{DAG} \mathbf{A}\left[\mathbf{A}\left[\ldots\left[\mathbf{A}\underbrace{\frac{\partial \boldsymbol{X}^0}{\partial \boldsymbol{N}}}_{=\mathbf{0}} + \mathbf{B}\right]\right] + \mathbf{B}\right] + \mathbf{B} \tag{17}$$

$$= \sum_{i=0}^{d-1} \mathbf{A}^i \mathbf{B} = \left(\mathbf{I}_d - \mathbf{A}\right)^{-1} \mathbf{B}, \tag{18}$$

where the structural equivalences follow by the structural faithfulness of $\mathbf{J}_{\boldsymbol{f}}$ (Assum. A.1), the last equality expresses the sum of the geometric series with elements $\mathbf{A}^i$ (the sum is finite as $\mathbf{A}$ is strictly lower triangular). By invoking the inverse function theorem (by assumption, $\boldsymbol{f}$ is bojective), we can express $\mathbf{J}_{\boldsymbol{f}^{-1}}$:

$$\mathbf{J}_{\boldsymbol{f}^{-1}} = \mathbf{J}_{\boldsymbol{f}}^{-1} \sim_{DAG} \mathbf{B}^{-1}\left(\mathbf{I}_d - \mathbf{A}\right). \tag{19}$$

$\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$ follows as $\mathbf{A} \sim_{DAG} \boldsymbol{\mathcal{A}}$ and $\mathbf{B}$ is diagonal (the invariance of $\sim_{DAG}$ follows from Def. C.1(i)). For proving that $\mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}} \sim_{DAG} (\mathbf{I}_d - \boldsymbol{\mathcal{A}})$, we need $\mathbf{J}_{\boldsymbol{f}^{-1}} \sim_{DAG} \mathbf{J}_{\widehat{\boldsymbol{f}}^{-1}}$ (Def. C.1(iv)), which requires us to account for all indeterminacies of strong identifiability: i) Def. C.1(i) accounts for scalings and sign flips; ii) Def. C.1(ii) for zero-preserving transformations; and iii) Def. C.1(iii) for permutations, which can be extracted as shown in Lem. 1. □

E Experimental details

Assumption E.1 (NLICA assumptions)**.** *We assume the setting of (Zimmermann et al., 2021), specifically that of Thm. 6, under which, an encoder which minimizes a contrastive loss was proven to estimate the noise variables (often referred to as "sources" in the ICA literature) up to a composition of input independent permutations, sign flips, and rescaling. For completeness, we restate the assumptions below:*

Table 6: Results for **sparse** linear and nonlinear SEMs. Mean Correlation Coefficient (MCC) measures identifiability, $|\mathcal{E}^*|$ is the maximum number of edges in a DAG, Acc is accuracy, Ours is our proposal, HSIC refers to using HSIC independence tests, and SHD is the Structural Hamming Distance

		Linear				Nonlinear			
$\lvert\mathcal{E}^*\rvert$	d	MCC	Acc(Ours)	Acc(HSIC)	SHD	MCC	Acc(Ours)	Acc(HSIC)	SHD
6	3	1.	$0.917_{\pm 0.108}$	$0.708_{\pm 0.11}$	0.111	1.	$0.889_{\pm 0.111}$	$0.75_{\pm 0.144}$	0.111
15	5	$0.961_{\pm 0.062}$	$0.768_{\pm 0.121}$	$0.784_{\pm 0.111}$	$0.256_{\pm 0.132}$	$0.972_{\pm 0.059}$	$0.76_{\pm 0.095}$	$0.84_{\pm 0.098}$	$0.208_{\pm 0.0873}$
36	8	$0.844_{\pm 0.184}$	$0.709_{\pm 0.084}$	$0.711_{\pm 0.122}$	$0.322_{\pm 0.109}$	$0.783_{\pm 0.155}$	$0.656_{\pm 0.059}$	$0.708_{\pm 0.119}$	$0.375_{\pm 0.081}$
55	10	$0.8_{\pm 0.217}$	$0.648_{\pm 0.059}$	$0.715_{\pm 0.1}$	$0.336_{\pm 0.055}$	$0.734_{\pm 0.206}$	$0.618_{\pm 0.044}$	$0.69_{\pm 0.086}$	$0.37_{\pm 0.082}$

Table 7: Results for **permuted** (i.e., π is not the identity) linear and nonlinear SEMs. Mean Correlation Coefficient (MCC) measures identifiability, $|\mathcal{E}^*|$ is the maximum number of edges in a DAG, Acc is accuracy, Ours is our proposal, HSIC refers to using HSIC independence tests, and SHD is the Structural Hamming Distance

		Linear				Nonlinear			
$\lvert\mathcal{E}^*\rvert$	d	MCC	Acc(Ours)	Acc(HSIC)	SHD	MCC	Acc(Ours)	Acc(HSIC)	SHD
6	3	1.	1.	0.667	0.	1.	1.	0.667	0.
15	5	$0.989_{\pm 0.039}$	$0.949_{\pm 0.098}$	$0.866_{\pm 0.088}$	$0.051_{\pm 0.098}$	$0.988_{\pm 0.039}$	$0.94_{\pm 0.087}$	0.863	$0.06_{\pm 0.087}$
36	8	$0.837_{\pm 0.252}$	$0.834_{\pm 0.162}$	$0.624_{\pm 0.127}$	$0.166_{\pm 0.162}$	$0.752_{\pm 0.232}$	$0.794_{\pm 0.138}$	$0.687_{\pm 0.139}$	$0.206_{\pm 0.138}$
55	10	$0.852_{\pm 0.251}$	$0.761_{\pm 0.213}$	$0.578_{\pm 0.086}$	$0.239_{\pm 0.213}$	$0.794_{\pm 0.255}$	$0.705_{\pm 0.16}$	$0.573_{\pm 0.05}$	$0.295_{\pm 0.159}$

(i) the space of sources/latent/noise variables, is a convex body in $\mathbb{R}^d$, i.e. a hyperrectangle/cube.
(ii) $p(\boldsymbol{N})$, the marginal distribution, is uniform
(iii) $p(\tilde{\boldsymbol{N}}|\boldsymbol{N})$, the conditional distribution, is a rotationally asymmetric generalized normal distribution (Subbotin, 1923), i.e. a Laplace distribution.
(iv) the observations are generated by a smooth, bijective (i.e., invertible) mapping
(v) the contrastive objective uses the same metric as $p(\tilde{\boldsymbol{N}}|\boldsymbol{N})$, i.e., L_1 for Laplace (cf. (Zimmermann et al., 2021, Def. 1)).

Table 5: Hyperparameters for our experiments (§ 4)

Parameter	Values
$\widehat{\boldsymbol{f}}^{-1}$	6-layer MLP
Activation	Leaky ReLU
Batch size	6144
Learning rate	1e−4
$\mathbb{R}^d$	$[0;1]^d$
C_p	1
m_p	0
C_{param}	0.05
m_{param}	1
p	1
τ	1
α	0.5

F Notation

Acronyms

ANM Additive Noise Model

CD Causal Discovery
CL Contrastive Learning

DAG Directed Acyclic Graph
DGP Data Generating Process

i.i.d. independent and identically distributed
ICA Independent Component Analysis
ICM Independent Causal Mechanisms

LiNGAM Linear Non-Gaussian Acyclic Model
LSTM Long Short-Term Memory

MCC Mean Correlation Coefficient
MLP Multi-Layer Perceptron

NLICA nonlinear Independent Component Analysis

SEM Structural Equation Model
SHD Structural Hamming Distance

Nomenclature

$\mathcal{L}_{CL}$ contrastive loss function
$\mathcal{L}_{\pi}$ regularizer for learning π
$\mathbf{S}$ Sinkhorn network
$\mathcal{E}$ edge set of a graph
$\mathcal{L}$ loss function
d problem dimensionality

Algebra
α scalar field
$\mathbf{D}$ diagonal matrix
$\mathbf{I}_D$ D-dimensional identity matrix
$\mathbf{J}$ Jacobian matrix

Causality
N noise (independent) variable component
X observation component
$\boldsymbol{N}$ noise (independent) variable vector
$\boldsymbol{Pa}$ parent set of $\boldsymbol{X}$
$\boldsymbol{X}$ observation vector
$\boldsymbol{\mathcal{A}}$ adjacency matrix of a SEMs
$\boldsymbol{\mathcal{C}}$ connectivity matrix of a SEMs
$\boldsymbol{f}$ structural assignment in SEMs
$\mathcal{I}$ index set
π causal ordering
$\sim_{DAG}$ structural equivalence
f a component of $\boldsymbol{f}$

Latents
$\mathcal{Z}$ latents
d dimensionality of the latent space $\mathcal{Z}$

Observations
D dimensionality of the observation space $\mathcal{X}$
$\mathcal{X}$ observation space