Learning Linear Causal Representations from Interventions under General Nonlinear Mixing

Simon Buchholz*[1], Goutham Rajendran*[2], Elan Rosenfeld[2],
Bryon Aragam[3], Bernhard Schölkopf[1], and Pradeep Ravikumar[2]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[2]Carnegie Mellon University, Pittsburgh, USA
[3]University of Chicago, Chicago, USA

June 6, 2023

Abstract

We study the problem of learning causal representations from unknown, latent interventions in a general setting, where the latent distribution is Gaussian but the mixing function is completely general. We prove strong identifiability results given unknown single-node interventions, i.e., without having access to the intervention targets. This generalizes prior works which have focused on weaker classes, such as linear maps or paired counterfactual data. This is also the first instance of causal identifiability from non-paired interventions for deep neural network embeddings. Our proof relies on carefully uncovering the high-dimensional geometric structure present in the data distribution after a non-linear density transformation, which we capture by analyzing quadratic forms of precision matrices of the latent distributions. Finally, we propose a contrastive algorithm to identify the latent variables in practice and evaluate its performance on various tasks.

1 Introduction

Modern generative models such as GPT-4 [57] or DDPMs [25] achieve tremendous performance for a wide variety of tasks [7]. They do this by effectively learning high-level representations which map to raw data through complicated *non-linear maps*, such as transformers [84] or diffusion processes [74]. However, we are unable to reason about the specific representations they learn. Besides their susceptibility to bias [19], they often fail to generalize to out-of-distribution settings [5], rendering them problematic in safety-critical domains. In order to gain a deeper understanding of what representations deep generative models learn, one line of work has pursued the goal of causal representation learning (CRL) [70]. CRL aims to learn high-level representations of data while simultaneously recovering the rich causal structure embedded in them, allowing us to reason about them and use them for higher-level cognitive tasks such as systematic generalization and planning. This has been used effectively in many application domains including genomics [50] and robotics [42, 88].

A crucial primitive in CRL is the fundamental notion of identifiability [35, 65], i.e., the question whether a unique (up to tolerable ambiguities) model can explain the observed data. It is well known that because of non-identifiability, CRL is impossible in general settings in the absence of inductive biases or additional information [28, 48]. In this work, we consider additional information in the form of interventional data [70]. It is common to have access to such data in many application domains such as genomics and robotics stated above (e.g. [16, 56, 55]). Moreover, there is a pressing need in such safety-critical domains to build reliable and trustworthy systems, making identifiable CRL particularly important. Therefore, it's important to study whether we can and also how to perform identifiable CRL from raw observational and interventional data. Here, identifiability opens the possibility to provably recover the true representations with formal guarantees. Meaningfully learning such representations with causal structure enables better interpretability, allows us to reason about fairness, and helps with performing high-level tasks such as planning and reasoning.

*Equal Contribution

In this work, we study precisely this problem of causal representation learning in the presence of interventions. While prior work has studied simpler settings of linear or polynomial mixing [66, 10, 79, 83, 2], we allow for general non-linear mixing, which means our identifiability results apply to deep neural network embeddings and are therefore more applicable to complex real-world systems and datasets which are used in practice. With our results, we make progress on fundamental questions on interventional learning raised by [70].

Concretely, we consider a general model with latent variables Z and observed data X generated as $X = f(Z)$ where $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ is an arbitrary non-linear mixing function. We assume Z satisfies a Gaussian structural equation model (SEM) which is unknown and unobserved. A Gaussian prior is commonly used in practice (which implies a linear SEM over Z) and further, having a simple model for Z allows us to learn useful representations, generate data efficiently, and explore the latent causal relationships, while the non-linearity of f ensures universal approximability of the model. We additionally assume access to interventional data $X^{(i)} = f(Z^{(i)})$ for $i \in I$ where $Z^{(i)}$ is the latent under an intervention on a single node Z_{t_i}. Notably, we allow for various kinds of interventions, we don't require knowledge of the targets t_i, and we don't require paired data, (i.e., we don't need counterfactual samples from the joint distribution $(X, X^{(i)})$, but only their marginals). Having targets or counterfactual data is unrealistic in many practical settings but many prior identifiability results require them, so eliminating this dependence is a crucial step towards CRL in the real world.

Contributions Our main contribution is a general solution to the identifiability problem that captures practical settings with non-linear mixing functions (e.g. deep neural networks) and unknown interventions (since the targets are latent). We study both perfect and imperfect interventions and additionally allow for shift interventions, where perfect interventions remove the dependence of the target variable from its parents, while imperfect interventions (also known as soft interventions) modify the dependencies. Below, we summarize our main contributions:

1. We show identifiability of the full latent variable model, given non-paired interventional data with unknown intervention targets. In particular, we learn the mixing, the targets and the latent causal structure.

2. Compared to prior works that have focused on linear/polynomial f, we allow non-linear f which encompasses representations learned by e.g. deep neural networks and captures complex real-world datasets. Moreover, we study both imperfect (also called soft) and perfect interventions, and always allow shift interventions.

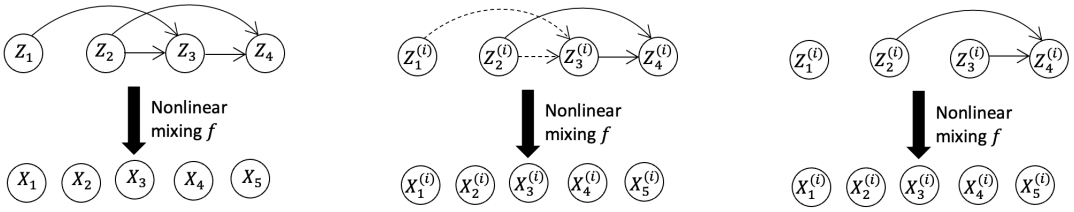
3. We construct illustrative counterexamples to probe tightness of our assumptions which suggest directions for future work.

4. We propose a novel algorithm based on contrastive learning to learn causal representations from interventional data, and we run experiments to validate our theory. Our experiments suggest that a contrastive approach, which so far has been unexplored in interventional CRL, is a promising technique going forward.

2 Related work

Causal representation learning [70, 69] has seen much recent progress and applications since it generalizes and connects the fields of Independent Component Analysis [12, 27, 29], causal inference [76, 59, 61, 62, 77] and latent variable modeling [38, 3, 73, 90, 31]. Fundamental to this approach is the notion of identifiability [34, 14, 87]. Due to non-identifiability in general settings without inductive biases [28, 48], prior works have approached this problem from various angles — using additional auxiliary labels [34, 26, 6, 71]; by imposing sparsity [70, 41, 52, 96]; or by restricting the function classes [39, 8, 97, 21]. See also the survey [31]. Another line of research has investigated similar models in the context of *domain generalization* [66, 10, 67], also demonstrating latent causal subspace identifiability (and recovery) from interventional data. Moreover, a long line of works have proposed practical methods for CRL (which includes causal disentanglement as a special case), [18, 89, 15, 91, 43, 13] to name a few. It's worth noting that most of these approaches are essentially variants of the Variational Autoencoder framework [37, 63].

Of particular relevance to this work is the setting when interventional data is available. We first remark that the much simpler case of fully observed variables (i.e., no latent variables), has been studied in e.g. [22, 60, 78, 33, 17] (see also the survey [77]). In this work, we consider the more difficult setting of structure learning over latent variables, which have been explored in [66, 10, 44, 41, 6, 97, 1, 79, 83, 2].



(a) No interventions (b) An imperfect intervention (c) A perfect intervention

Figure 1: Illustration of an example latent variable model under interventions. (a) No interventions. (b) An imperfect intervention on node $t_i = 3$. Dashed edges indicate the weights could have potentially been modified. (c) A perfect intervention on node $t_i = 3$.

Among these, [44] assumes that the intervention targets are known, [41] specifically consider instance-level pre- and post- interventions for time-series data and [6, 97, 1] assumes access to paired counterfactual data. In contrast, we assume unknown targets and work in general settings with non-paired interventional data, which is important in various real-world applications [80, 93]. The work [46] require several graphical restrictions on the causal graph and also require $2d$ interventions, while we make no graphical restrictions and only require d interventions (which we also show cannot be improved under our assumptions). [66, 10, 79, 83] consider linear mixing functions f whereas we study general non-linear f. Finally, [2] consider polynomial mixing in the more restricted class of deterministic do-interventions. For a more detailed comparison to the related works we refer to Appendix C.

Our proposed algorithm is based on contrastive learning. Contrastive learning has been used in other contexts in this domain [30, 26, 53, 97, 86] (either in the setting of time-series data or paired counterfactual data), however the application to non-paired interventional settings is new to the best of our knowledge.

Notation In this work, we will almost always work with vectors and matrices in d dimensions where d is the latent dimension, and we will disambiguate when necessary. For a vector v, we denote by v_i its i-th entry. Let Id denote the $d \times d$ identity matrix with columns as the standard basis vectors $e_1, \ldots, e_d$. We denote by $N(\mu, \Sigma)$ the multivariate normal distribution with mean μ and covariance Σ. For a set C, let $\mathcal{U}(C)$ denote the uniform distribution on C. For two random variables X, X', we write $X \overset{\mathcal{D}}{=} X'$ if their distributions are the same. We denote the set $\{1, \ldots, d\}$ by $[d]$. For permutation matrices we use the convention that $(P_\omega)_{ij} = \mathbf{1}_{j=\omega(i)}$ for $\omega \in S_d$. We use standard directed graph terminology, e.g. edges, parents. When we use the term "non-linear mixing" in this work, we also include linear mixing as a special case.

3 Setting: Interventional Causal Representation Learning

We now formally introduce our main settings of interest and the required assumptions. We assume that there is a latent variable distribution Z on $\mathbb{R}^d$ and an observational distribution X on $\mathbb{R}^{d'}$ given by $X = f(Z)$ where $d \leq d'$ and f is a non-linear mixing. This encapsulates the most general definition of a latent variable model. As the terminology suggests, we observe X in real-life, e.g. images of cats and Z encodes high-level latent information we wish to learn and model, e.g. Z could indicate orientation and size.

Assumption 1 (Nonlinear f)**.** *The non-linear mixing f is injective, differentiable and embeds into $\mathbb{R}^{d'}$.*

Such an assumption is standard in the literature. Injectivity is needed in order to identify Z from X because otherwise we may have learning ambiguity if multiple Zs map to the same X. Differentiability is needed to transfer densities and is a standard assumption in ML methods based on gradient descent. Note that Assumption 1 allows for a large class of complicated nonlinearities and we discuss in Appendix E why proving results for such a large class of mixing functions is important.

Next, we assume that the latent variables Z encode causal structure which can be expressed through a Structural Causal Model (SCM) on a Directed Acyclic Graph (DAG). We focus on linear SCMs with an underlying causal graph G on vertex set $[d]$, encoded by a matrix A through its non-zero entries, i.e., there is an edge $i \to j$ in G iff $A_{ji} \neq 0$.

Assumption 2 (Linear Gaussian SCM on Latent Variables)**.** *The latent variables Z follow a linear SCM with Gaussian noise, so*

$$Z = AZ + D^{1/2}\epsilon \tag{1}$$

where D is a diagonal matrix with positive entries, A encodes a DAG G and $\epsilon \sim N(0, \mathrm{Id})$.

For an illustration of the model, see Fig. 1a. It is convenient to encode the coefficients of linear SCMs through the matrix $B = D^{-1/2}(\mathrm{Id} - A)$. Then $Z = B^{-1}\epsilon$ and both D and A can be recovered from B, indeed the diagonal entries of B agree with the entries of $D^{-1/2}$. Note that we can and will always assume that the entries of D are positive since $\epsilon \overset{\mathcal{D}}{=} -\epsilon$.

The Gaussian prior assumption is standard in latent variable modeling and enables efficient inference for downstream tasks, among other advantages. Importantly, under the above assumptions, our model has infinite modeling capacity [51, 81, 32], so they're able to model complex datasets such as images and there's no loss in representational power.

The goal of representation learning is to learn the non-linear mixing f and the high-level latent distribution Z from observational data X. In causal representation learning, we wish to go a step further and model Z as well by learning the parameters B which then lets us easily recover A, D and the causal graph G. For general non-linear mixing, this model is not identifiable and hence we cannot hope to recover Z, but in this work we use additional interventional data, similar to the setups in recent works [2, 79, 83].

Assumption 3 (Single node interventions)**.** *We consider interventional distributions $Z^{(i)}$ for $i \in I$ which are single node interventions of the latent distribution Z, i.e., for each intervention i there is a target node t_i and we change the assigning equation of $Z^{(i)}_{t_i}$ to*

$$Z^{(i)}_{t_i} = (A^{(i)}Z^{(i)})_{t_i} + (D^{(i)})^{1/2}_{t_i,t_i}(\epsilon_{t_i} + \eta^{(i)}) \tag{2}$$

while leaving all other equations unchanged. We assume that $A^{(i)}_{t_i,k} \neq 0$ only if $k \to t_i$ in G, i.e., no parents are added and $\eta^{(i)}$ denotes a shift of the mean.

For an illustration, see Fig. 1b. An intervention on node t_i has no effect on the non-descendants of t_i, but will have a downstream effect on the descendants of t_i. In particular, A is modified so that the weight of the edge $k \to t_i$ could potentially be changed if it exists already but no new incoming edges to t_i may be added. Also, the noise variable ϵ_{t_i} is also allowed to be modified via a scale intervention as well as a shift intervention. Note that [79, 2] do not allow shift interventions, whereas we do.

Again, this can be written concisely as $Z^{(i)} = (B^{(i)})^{-1}(\epsilon + \eta^{(i)}e_{t_i})$ where $B^{(i)} = (D^{(i)})^{-1/2}(\mathrm{Id} - A^{(i)})$. Let $r^{(i)}$ denote the row t_i of $B^{(i)}$, then we can write $B^{(i)} = B - e_{t_i}(B^\top e_{t_i} - r^{(i)})^\top$. We assume that interventions are non-trivial, i.e., $Z^{(i)} \overset{\mathcal{D}}{\neq} Z$.

In our model, we observe interventional distributions $X^{(i)} = f(Z^{(i)})$ for various interventions $i \in I$.

Definition 1 (Intervention Types)**.** *We call an intervention perfect (or stochastic hard) if $A^{(i)}_{t_i\cdot} = 0$, and $D^{(i)}_{t_i,t_i} \neq D_{t_i,t_i}$, so that $r^{(i)} = \lambda_i e_{t_i}$ with $\lambda_i > 0, \lambda_i^2 \neq D_{t_i,t_i}$, i.e., we remove all connections to former parents and change the noise variance. Note that we exclude $\lambda_i = 0$ which results in a degenerate distribution. We potentially also change the noise mean (due to potential shift interventions $\eta^{(i)}$). We call an intervention a pure shift intervention if $A^{(i)}_{t_i\cdot} = A_{t_i\cdot}$, and $D^{(i)}_{t_i,t_i} = D_{t_i,t_i}$, so that $r^{(i)} = B^\top e_{t_i}$, i.e., the relation to the parents stays the same, we only change the mean.*

We call a single-node intervention imperfect if it is not perfect (some prior works call it a soft intervention). For an illustration of perfect and imperfect interventions, see Fig. 1c, Fig. 1b respectively. Finally, we require an exhaustive set of interventions.

Assumption 4 ((Coverage of Interventions))**.** *All nodes are intervened upon by at least one intervention, i.e., $\{t_i : i \in I\} = [d]$*

This assumption was also made in the prior works [79, 2, 83]. When the interventions don't cover all the nodes, we have non-identifiability as described in Section 4 and Appendix D. We also extensively discuss that none of our other assumptions can simply be dropped in these sections.

Remark 1. *Our theory also readily extends to noisy observations, i.e., $X = f(Z) + \nu$ where ν is independent noise. In this case, we first denoise via a standard deconvolution argument [34, 41] and then apply our theory.*

To simplify the notation, it is convenient to use $B^{(0)} = B$, $\eta^{(0)} = 0$ for the observational distribution. We also define $\bar{I} = I \cup \{0\}$. Then, all information about the latent variable distributions $Z^{(i)}$ and observed distributions $X^{(i)}$ are contained in $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$.

4 Main Results

We can now state our main results for the setting introduced above.

Theorem 1. *Suppose we are given distributions $X^{(i)}$ generated using a model $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ such that Assumptions 1-4 hold and such that all interventions i are perfect. Then the model is identifiable up to permutation and scaling, i.e., for any model $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in\bar{I}}, \widetilde{f})$ that generates the same data, there is a permutation $\omega \in S_d$ (and associated permutation matrix P_ω) and an invertible pointwise scaling matrix $\Lambda \in \mathrm{Diag}(d)$ such that*

$$\widetilde{t}_i = \omega(t_i), \quad \widetilde{B}^{(i)} = P_\omega^\top B^{(i)} \Lambda^{-1} P_\omega, \quad \widetilde{f} = f \circ \Lambda^{-1} P_\omega, \quad \widetilde{\eta}^{(i)} = \eta^{(i)}. \tag{3}$$

This in particular implies that

$$\widetilde{Z}^{(i)} \overset{\mathcal{D}}{=} P_\omega^\top \Lambda Z^{(i)} \tag{4}$$

and we can identify the causal graph G up to permutation of the labels.

This result says that for the interventional model as described in the previous section, we can identify the non-linear map f, the intervention targets t_i, the parameter matrices B, D, A up to permutations P_ω and diagonal scaling Λ. Moreover, we can recover the shifts $\eta^{(i)}$ exactly and also the underlying causal graph G up to permutations.

Remark 2. *Identifiability and recovery up to permutation and scaling is the best possible for our setting. This is because the latent variables Z are not actually observed, which means we cannot (and in fact don't need to) resolve permutation and scaling ambiguity without further information about Z. See [79, Proposition 1] for the short proof.*

When we drop the assumption that the interventions are perfect, we can still obtain a weaker identifiability result. Define π_G to be the minimal partial order on $[d]$ such that $i \preceq j$ if (i, j) is an edge in G, i.e., $i \preceq j$ if and only if i is an ancestor of j in G. Note that any topological ordering of G is compatible with the partial order π_G. Then, our next result shows that under imperfect interventions, we can still recover the partial order π_G.

Theorem 2. *Suppose we are given the distributions $X^{(i)}$ generated using a model $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ such that the Assumptions 1-4 hold and none of the interventions is a pure shift intervention. Then π_G can be identified up to a permutation of the labels.*

Theorem 1 and Theorem 2 generalize the main results of [79] which assume linear f while we allow for non-linear f. The key new ingredient of our work is the following theorem, which shows identifiability of f up to linear maps.

Theorem 3. *Assume that $X^{(i)}$ is generated according to a model $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ such that the Assumptions 1-4 hold. Then we have identifiability up to linear transformations, i.e., if $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i), \widetilde{f})$ generates the same observed distributions $\widetilde{f}(\widetilde{Z}^{(i)}) \overset{\mathcal{D}}{=} X^{(i)}$, then there is an invertible linear map T such that*

$$\widetilde{f} = f \circ T^{-1}, \quad \widetilde{Z}^{(i)} = T Z^{(i)}. \tag{5}$$

The proof of this theorem is deferred to Appendix A. In Appendix B we then review the results of [79] and show how their results can be extended to obtain our main results, Theorem 1 and Theorem 2. Now we provide some intuition for the proofs of the main theorems.

Proof intuition The recent work [79] studies the special case when f is linear, and the proof is linear algebraic. In particular, they consider row spans of the precision matrices of $X^{(i)}$, project them to certain linear subspaces and use those subspaces to construct a generalized RQ decomposition of the pseudoinverse of the linear mixing matrix f. However, once we are in the setting of non-linear f, such an approach is not feasible because we can no longer reason about row spans of the precision matrices

of X, which have been transformed non-linearly thereby losing all linear algebraic structure. Instead, we take a statistical approach and look at the log densities of the $X^{(i)}$. By choosing a Gaussian prior, the log-odds $\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x)$ of $X^{(i)}$ with respect to $X^{(0)}$ can be written as a quadratic form of difference of precision matrices, evaluated at non-linear functions of x. For simplicity of this exposition, ignore terms arising from shift interventions and determinants of covariance matrices. Then, the log-odds looks like $\theta(x) = f^{-1}(x)^\top(\Theta^{(i)} - \Theta^{(0)})f^{-1}(x)$, where $\Theta^{(i)}$ is the precision matrix of $Z^{(i)}$.

At this stage, we again shift our viewpoint to geometric and observe that $\Theta^{(i)} - \Theta^{(0)}$ has a certain structure. In particular, for single-node interventions, it has rank at most 2 and for source node targets, it has rank 1. This implies that the level set manifolds of the quadratic forms $\theta(x)$ also have a certain geometric structure in them, i.e., the DAG leaves a geometric signature on the data likelihood. We exploit this carefully and proceed by induction on the topological ordering until we end up showing that f can be identified up to a linear transformation, which is our main Theorem 3. Here, the presence of shift interventions adds additional complexities, and we have to generalize all of our intermediate lemmas to handle these. Once we identify f up to a linear transformation, we can apply the results of [79] to conclude Theorems 1 and 2.

On the optimality and limitations of the assumptions We make a few brief remarks on our assumptions, deferring to Appendix D a full discussion and the technical construction of several illustrative counterexamples.

1. *Number of interventions:* For our main results, Theorems 1 and 2, we assume that there are at least d interventions (Assumption 4). This cannot be weakened even for linear mixing functions [79]. In addition, we also show in Fact 1 that for the linear identifiability proved in Theorem 3, $d-2$ interventions are not sufficient. Thus, the required number of interventions is tight in Theorems 1, 2 and is tight up to at most one intervention in Theorem 3.

2. *Intervention type:* In the setting of imperfect interventions, the weaker identifiability guarantees in Theorem 2 cannot be improved even when the mixing is linear [79, Appendix C]. We also show in Fact 2 that if we drop the condition that interventions are not pure shift interventions, we have non-identifiability. Concretely, we show that when all interventions are of pure shift type, any causal graph is compatible with the observations. Finally, in contrast to the special case of linear mixing, we show in Lemma 8 that we need to exclude non-stochastic hard interventions (i.e., $\mathrm{do}(Z_i = z_i)$) for identifiability up to linear maps.

3. *Distributional assumptions:* We assume Gaussian latent variables, while allowing for very flexible mixing f. This model has universal approximation guarantees and is moreover used ubiquitously in practice. While the result can potentially be extended to more general latent distributions (e.g., exponential families), we additionally show in Lemma 9 that the result is not true when making no assumption on the distribution of ϵ.

5 Experimental methodology

In this section, we explain our experimental methodology and the theoretical underpinning of our approach. Our main experiments for interventional causal representation learning focus on a method based on contrastive learning. We train a deep neural network to learn to distinguish observational samples $x \sim X^{(0)}$ from interventional samples $x \sim X^{(i)}$. Additionally, we design the last layer of the model to model the log-likelihood of a linear Gaussian SCM. Due to representational flexibility of deep neural networks, we will in principle learn the Bayes optimal classifier after optimal training, which we show is related to the underlying causal model parameters. Accordingly, with careful design of the last layer parametric form, we indirectly learn the parameters of the underlying causal model. Similar methods have been used for time-series data or multimodal data [26, 31] but to the best of our knowledge, the contrastive learning approach to interventional learning is novel.

Denote the probability density of $x \sim X^{(i)}$ (resp. $z \sim Z^{(i)}$) by $p_X^{(i)}(x)$ (resp. $p_Z^{(i)}(z)$). The next lemma describes the log-odds of a sample x coming from the interventional distribution $X^{(i)}$ as opposed to the observational distribution $X^{(0)}$. We focus on the identifiable case (see Theorem 1) and therefore only consider perfect interventions. As per the notation in Section 3 and Appendix A.1, let $s^{(i)}$ denote the row t_i of $B^{(0)}$ and let $\eta^{(i)}, \lambda_i$ denote the magnitude of the shift and scale intervention respectively.

Lemma 1. *When we have perfect interventions, the log-odds of a sample x coming from $X^{(i)}$ over coming from $X^{(0)}$ is given by*

$$\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x) = c_i - \frac{1}{2}\lambda_i^2(((f^{-1}(x))_{t_i})^2 + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i}) + \frac{1}{2}\langle f^{-1}(x), s^{(i)}\rangle^2 \tag{6}$$

for a constant c_i independent of x.

The proof is deferred to Appendix F. The form of the log-odds suggests considering the following functions

$$g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta) = \alpha_i - \beta_i h_{t_i}^2(x, \theta) + \gamma_i h_{t_i}(x, \theta) + \langle h(x, \theta), w^{(i)}\rangle^2 \tag{7}$$

where $h(\cdot, \theta)$ denotes a neural net parametrized by θ, parameters $w^{(i)}$ are the rows of a matrix W and α_i, β_i, γ_i are learnable parameters. Note that the ground truth parameters minimize the following cross entropy loss

$$\mathcal{L}_{\text{CE}}^{(i)} = \mathbb{E}_{j\sim\mathcal{U}(\{0,i\})}\mathbb{E}_{x\sim X^{(j)}}\text{CE}(\mathbf{1}_{j=i}, g_i(x)) = -\mathbb{E}_{j\sim\mathcal{U}(\{0,i\})}\mathbb{E}_{x\sim X^{(j)}}\ln\left(\frac{e^{\mathbf{1}_{j=i}g_i(x)}}{e^{g_i(x)}+1}\right). \tag{8}$$

Note that (compare (6) and (7)) W should learn $B = D^{-1/2}(\text{Id} - A)$ thus its off-diagonal entries should form a DAG. To enforce this we add the NOTEARS regularizer [95] given by $\mathcal{R}_{NOTEARS}(W) = \operatorname{tr}\exp(W_0 \circ W_0) - d$ (see Appendix F.3) where W_0 equals W with the main diagonal zeroed out. We also promote sparsity by adding the l_1 regularization term $\mathcal{R}_{REG}(W) = \|W_0\|_1$. Thus, the total loss is given by

$$\mathcal{L}(\alpha, \beta, \gamma, W, \theta) = \sum_{i\in I}\mathcal{L}_{\text{CE}}^{(i)} + \tau_1\mathcal{R}_{NOTEARS}(W) + \tau_2\mathcal{R}_{REG}(W) \tag{9}$$

for hyperparameters τ_1 and τ_2. Our identifiability result Theorem 1 implies that when we assume that the neural network has infinite capacity, the loss in (9) is minimized, τ_1 is large and τ_2 small, and we learn Gaussian latent variables $h(X^{(i)}, \theta)$, then we recover the ground truth latent variables up to the tolerable ambiguities of labeling and scale, i.e., h recovers f^{-1} and W recovers B up to permutation and scale. Thus, we estimate the latent variables using $\hat{Z} = h(X, \theta)$ and estimate the DAG using W_0. Full details of the practical implementation of our approach are given in Appendix G.

Our experimental setup is similar to [79, 2], we consider d interventions with different targets (our theory holds in full generality) and therefore, we can arbitrarily assign $t_i = i$ based on the intervention index i which removes the permutation ambiguity. We focus on non-zero shifts because the cross entropy loss together with the quadratic expression for the log-odds results in a non-convex output layer which makes it hard to find the global loss minimizer, as we will describe in more detail in Appendix F.4. We emphasize that this is no contradiction to the theoretical results stated above. Even when there is no shift intervention the latent variables are identifiable, but our algorithm often fails to find the global minimizer of the loss (9) due to the non-convex loss landscape. For the sake of exposition and to set the stage for future work, we also briefly describe an approach via Variational Autoencoders (VAE). VAEs have been widely used in causal representation learning and while feasible in interventional settings, they are accompanied by certain difficulties, which we detail and suggest how to overcome in Appendix F.5.

6 Experiments

In this section, we implement our approach on synthetic data and image data. Complete details (architectures, hyperparameters) are deferred to Appendix G.

Data generation For all our experiments we use Erdös-Rényi graphs, i.e., we add each undirected edge with equal probability p to the graph and then finally orient them according to some random order of the nodes. We write $\text{ER}(d, k)$ for the Erdös-Rényi graph distribution on d nodes with kd expected edges. For a given graph G we then sample edge weights from $\mathcal{U}(\pm[0.25, 1.0])$ and a scale matrix D. For simplicity we assume that we have n samples from each environment $i \in \bar{I}$. We consider three types of mixing functions. First, we consider linear mixing functions where we sample all matrix entries i.i.d. from a Gaussian distribution. Then, we consider non-linear mixing functions that are parametrized by MLPs with three hidden layers which are randomly initialized, and have Leaky ReLU activations. Finally, we consider image data as described in [2]. Pairs of latent variables (z_{2i+1}, z_{2i+2}) describe the coordinates of balls in an image and the non-linearity f is the rendering of the image. The image generation is based on pygame [72]. A sample image can be found in Figure 3.

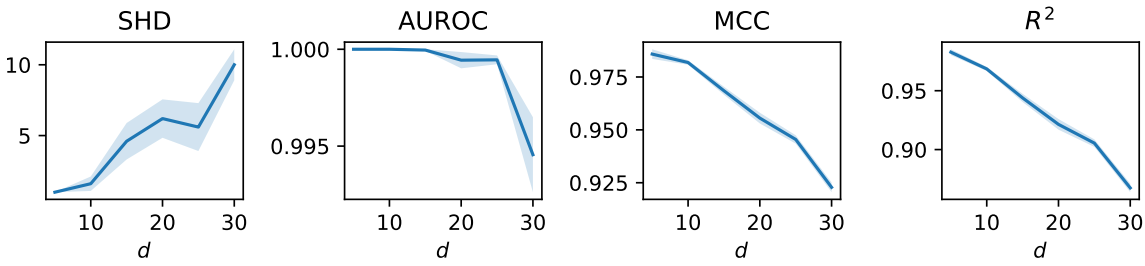


Figure 2: Dependence of performance metrics for $ER(d, 2)$ graphs with $d' = 100$ and nonlinear mixing f on the sample size d.

Evaluation Metrics We evaluate how well we can recover the ground truth latent variables and the underlying DAG. For the recovery of the ground truth latents we use the Mean Correlation Coefficient (MCC) [35, Appendix A.2]. For the evaluation of the learnt graph structure we use the Structural Hamming Distance (SHD) where we follow the convention that we count the number of edge differences of the directed graphs (i.e., an edge with wrong orientation counts as two errors). Since the scale of the variables is not fixed the selection of the edge selection threshold is slightly delicate. Thus, we use the heuristic where we match the number of selected edges to the expected number of edges. As a metric that is independent of this thresholding procedure, we also report the Area Under the Receiver Operating Curve (AUROC) for the edge selection.

Figure 3: Sample image with 3 balls.

Methods We implement our contrastive algorithm as explained in Section F where we use an MLP (with Leaky ReLU nonlinearities) for the function h for the linear and nonlinear mixing functions and a very small convolutional network for the image dataset, which are termed "Contrastive". For linear mixing function we also consider a version of the contrastive algorithm where h is a linear function, termed "Contrastive Linear". As baselines, we consider a variational autoencoder with latent dimension d and also the algorithm for linear disentanglement introduced in [79]. Since the variational autoencoder does not output a causal graph structure we report the result for the empty graph here which serves as a baseline.

Results for linear f First, for the sake of comparison, we replicate exactly the setting from [79], i.e., we consider initial noise variances sampled uniformly from $[2, 4]$ and we consider perfect interventions where the new variance is sampled uniformly from $[6, 8]$. We set $d = 5$, $d' = 10$, $k = 3/2$, and $n = 50000$. Results can be found in Table 1. The linear contrastive method identifies the ground truth latent variables up to scaling. The failure of the nonlinear method to recover the ground truth latents will be explained and further analyzed in Appendix F.4. Note that the nonlinear contrastive and the linear contrastive method recover the underlying graph better than the baseline.

Table 1: Results for linear f with $d = 5, d' = 10$, $k = 3/2$, $n = 50000$.

Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
Contrastive	4.6 ± 0.5	0.84 ± 0.02	0.05 ± 0.02	0.02 ± 0.00
Contrastive Linear	5.4 ± 1.6	0.80 ± 0.07	0.90 ± 0.03	1.00 ± 0.00
Linear baseline	7.0 ± 0.5	0.64 ± 0.05	0.83 ± 0.04	1.00 ± 0.00

Results for nonlinear f We sample all variances from $\mathcal{U}([1, 2])$ (initial variance and resampling for the perfect interventions) and the shift parameters $\eta^{(i)}$ of the interventions from $\mathcal{U}([1, 2])$. Results can be found in Table 2. We find that our contrastive method can recover the ground truth latents and the causal structure almost perfectly, while the baseline for linear disentanglement cannot recover the graph or the latent variables (which is not surprising as the mixing is highly nonlinear). Also, training a vanilla VAE does not recover the latent variables up to a linear map as indicated by the R^2 scores. In Figure 2 we illustrate the dependence on the dimension d of our algorithm in this setting.

Results for image data Finally, we report the results for image data. Here, we generate the graph as before and consider variances sampled from $\sigma^2 \sim \mathcal{U}([0.01, 0.02])$ and shifts $\eta^{(i)}$ from $\mathcal{U}([0.1, 0.2])$ (i.e., the shifts are still of order σ). We exclude samples where one of the balls is not contained in the image

Table 2: Results for nonlinear synthetic data with $n = 10000$.

Setting	Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
ER(5, 2), $d' = 20$	**Contrastive**	1.8 ± 0.5	0.97 ± 0.01	0.97 ± 0.00	0.96 ± 0.00
	VAE	10.0 ± 0.0	0.50 ± 0.00	0.48 ± 0.03	0.57 ± 0.07
	Linear baseline	10.6 ± 1.9	0.48 ± 0.11	0.32 ± 0.03	0.34 ± 0.06
ER(5, 2), $d' = 100$	**Contrastive**	1.0 ± 0.0	1.00 ± 0.00	0.99 ± 0.00	0.98 ± 0.00
	VAE	10.0 ± 0.0	0.50 ± 0.00	0.59 ± 0.02	0.68 ± 0.04
	Linear baseline	3.4 ± 1.2	0.85 ± 0.07	0.18 ± 0.04	0.11 ± 0.04
ER(10, 2), $d' = 20$	**Contrastive**	3.6 ± 1.3	0.98 ± 0.01	0.93 ± 0.00	0.87 ± 0.01
	VAE	18.6 ± 0.9	0.50 ± 0.00	0.59 ± 0.02	0.72 ± 0.02
	Linear baseline	29.6 ± 2.5	0.49 ± 0.02	0.44 ± 0.02	0.51 ± 0.02
ER(10, 2), $d' = 100$	**Contrastive**	1.6 ± 0.5	1.00 ± 0.00	0.98 ± 0.00	0.97 ± 0.00
	VAE	18.6 ± 0.9	0.50 ± 0.00	0.62 ± 0.02	0.78 ± 0.01
	Linear baseline	28.4 ± 2.1	0.51 ± 0.04	0.17 ± 0.03	0.13 ± 0.03

which generates a slight model misspecification compared to our theory. Again we find that we recover the latent graph and the latent variables reasonably well as detailed in Table 3.

Table 3: Results for image data with 2 or 3 balls ($d = 4$ and $d = 6$, respectively), with $n = 25000$.

Setting	Method	SHD ↓	AUROC ↑	MCC ↑	R^2 ↑
ER(4, 2)	**Contrastive**	3.0 ± 0.4	0.76 ± 0.05	0.81 ± 0.02	0.74 ± 0.01
	VAE	4.4 ± 0.4	0.50 ± 0.00	0.45 ± 0.04	0.49 ± 0.04
ER(4, 4)	**Contrastive**	3.4 ± 0.7	0.84 ± 0.06	0.81 ± 0.03	0.73 ± 0.03
	VAE	6.0 ± 0.0	0.50 ± 0.00	0.53 ± 0.04	0.52 ± 0.07
ER(6, 2)	**Contrastive**	4.6 ± 0.6	0.87 ± 0.03	0.69 ± 0.02	0.57 ± 0.03
	VAE	5.2 ± 0.5	0.50 ± 0.00	0.35 ± 0.04	0.36 ± 0.06
ER(6, 4)	**Contrastive**	7.4 ± 1.6	0.81 ± 0.06	0.76 ± 0.03	0.69 ± 0.02
	VAE	11.2 ± 0.6	0.50 ± 0.00	0.36 ± 0.02	0.39 ± 0.05

7 Conclusion

In this work, we extend several prior works and show identifiability for a widely used class of linear latent variable models with non-linear mixing, from interventional data with unknown targets. Counterexamples show that our assumptions are tight in this setting and could only potentially be relaxed under other additional assumptions. We leave to future work to extend our results for other classes of priors, such as non-parametric distribution families, and also to study sample complexity and robustness of our results. We also proposed a contrastive approach to learn such models in practice and showed that it can recover the latent structure in various settings. Finally, we highlight that the results of our experiments are very promising and it would be interesting to scale up our algorithms to large-scale datasets such as [47].

Acknowledgments

We acknowledge the support of AFRL and DARPA via FA8750-23-2-1015, ONR via N00014-23-1-2368, and NSF via IIS-1909816, IIS-1955532. This work was also supported by the Tübingen AI Center.

References

[1] K. Ahuja, J. S. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.

[2] K. Ahuja, Y. Wang, D. Mahajan, and Y. Bengio. Interventional causal representation learning. *arXiv preprint arXiv:2209.11924*, 2022.

[3] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu. A practical algorithm for topic modeling with provable guarantees. In *International conference on machine learning*, pages 280–288. PMLR, 2013.

[4] K. Bello, B. Aragam, and P. Ravikumar. Dagma: Learning dags via m-matrices and a log-determinant acyclicity characterization. *arXiv preprint arXiv:2209.08037*, 2022.

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[6] J. Brehmer, P. De Haan, P. Lippe, and T. Cohen. Weakly supervised causal representation learning. *arXiv preprint arXiv:2203.16437*, 2022.

[7] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*, 2023.

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

[9] F. Castelletti and S. Peluso. Network structure learning under uncertain interventions. *Journal of the American Statistical Association*, pages 1–12, 2022.

[10] Y. Chen, E. Rosenfeld, M. Sellke, T. Ma, and A. Risteski. Iterative feature matching: Toward provable domain generalization with logarithmic environments. *Advances in neural information processing systems*, 35, 2022.

[11] D. M. Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002.

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

[13] J. Cui, W. Huang, Y. Wang, and Y. Wang. Aggnce: Asymptotically identifiable contrastive learning.

[14] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *The Journal of Machine Learning Research*, 23(1):10237–10297, 2022.

[15] N. Dilokthanakul, P. A. Mediano, M. Garnelo, M. C. Lee, H. Salimbeni, K. Arulkumaran, and M. Shanahan. Deep unsupervised clustering with gaussian mixture variational autoencoders. *arXiv preprint arXiv:1611.02648*, 2016.

[16] A. Dixit, O. Parnas, B. Li, J. Chen, C. P. Fulco, L. Jerby-Arnon, N. D. Marjanovic, D. Dionne, T. Burks, R. Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[17] F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *arXiv preprint arXiv:1207.1389*, 2012.

[18] F. Falck, H. Zhang, M. Willetts, G. Nicholson, C. Yau, and C. C. Holmes. Multi-facet clustering variational autoencoders. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] E. Ferrara. Should ChatGPT be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*, 2023.

[20] J. L. Gamella, A. Taeb, C. Heinze-Deml, and P. Bühlmann. Characterization and greedy learning of gaussian structural causal models under unknown interventions. 2022.

[21] L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34, 2021.

[22] A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *The Journal of Machine Learning Research*, 13(1): 2409–2464, 2012.

[23] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] C. Heinze-Deml, J. Peters, and N. Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2), 2018.

[25] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

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

[27] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

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

[29] A. Hyvarinen, J. Karhunen, and E. Oja. Independent component analysis. *Studies in informatics and control*, 11(2):205–207, 2002.

[30] A. Hyvarinen, H. Sasaki, and R. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

[31] A. Hyvärinen, I. Khemakhem, and R. Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *arXiv preprint arXiv:2302.02672*, 2023.

[32] I. Ishikawa, T. Teshima, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Universal approximation property of invertible neural networks. *arXiv preprint arXiv:2204.07415*, 2022.

[33] A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. *Advances in neural information processing systems*, 33:9551–9561, 2020.

[34] I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

[35] I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Ice-beem: Identifiable conditional energy-based deep models. *NeurIPS2020*, 2020.

[36] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[38] B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Learning latent causal graphs via mixture oracles. *arXiv preprint arXiv:2106.15563*, 2021.

[39] B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 2022.

[40] B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Identifiability of deep generative models under mixture priors without auxiliary information. *arXiv preprint arXiv:2206.10044*, 2022.

[41] S. Lachapelle, P. Rodriguez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ica. In *Conference on Causal Learning and Reasoning*, pages 428–484. PMLR, 2022.

[42] T. E. Lee, J. A. Zhao, A. S. Sawhney, S. Girdhar, and O. Kroemer. Causal reasoning in simulation for structure and transfer learning of robot manipulation policies. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4776–4782. IEEE, 2021.

[43] S. Li, B. Hooi, and G. H. Lee. Identifying through flows for recovering latent representations. *arXiv preprint arXiv:1909.12555*, 2019.

[44] P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and S. Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

[45] C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2021.12.009. URL `https://www.sciencedirect.com/science/article/pii/S106352032100110X`. Special Issue on Harmonic Analysis and Machine Learning.

[46] Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. v. d. Hengel, K. Zhang, and J. Q. Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

[47] Y. Liu, A. Alahi, C. Russell, M. Horn, D. Zietlow, B. Schölkopf, and F. Locatello. Causal triplet: An open challenge for intervention-centric causal representation learning. In *2nd Conference on Causal Learning and Reasoning (CLeaR)*, 2023. arXiv:2301.05169.

[48] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

[49] I. Loshchilov and F. Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL `https://openreview.net/forum?id=Skq89Scxx`.

[50] M. Lotfollahi, A. K. Susmelj, C. De Donno, Y. Ji, I. L. Ibarra, F. A. Wolf, N. Yakubova, F. J. Theis, and D. Lopez-Paz. Compositional perturbation autoencoder for single-cell response modeling. *BioRxiv*, 2021.

[51] Y. Lu and J. Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. *Advances in neural information processing systems*, 33:3094–3105, 2020.

[52] G. E. Moran, D. Sridhar, Y. Wang, and D. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

[53] H. Morioka and A. Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International Conference on Artificial Intelligence and Statistics*, pages 3399–3426. PMLR, 2023.

[54] J. Moser. On the volume elements on a manifold. *Transactions of the American Mathematical Society*, 120(2):286–294, 1965. ISSN 00029947. URL `http://www.jstor.org/stable/1994022`.

[55] A. Nejatbakhsh, F. Fumarola, S. Esteki, T. Toyoizumi, R. Kiani, and L. Mazzucato. Predicting perturbation effects from resting activity using functional causal flow. *bioRxiv*, pages 2020–11, 2020.

[56] T. M. Norman, M. A. Horlbeck, J. M. Replogle, A. Y. Ge, A. Xu, M. Jost, L. A. Gilbert, and J. S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.

[57] OpenAI. GPT-4 technical report, 2023.

[58] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[59] J. Pearl. *Causality*. Cambridge university press, 2009.

[60] J. Peters, P. Buhlmann, and N. Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. arxiv. *Methodology*, 2015.

[61] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[62] G. Rajendran, B. Kivva, M. Gao, and B. Aragam. Structure learning in polynomial time: Greedy algorithms, bregman information, and exponential families. *Advances in Neural Information Processing*

Systems, 34:18660–18672, 2021.

[63] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[64] J. Rissanen. Modeling by shortest data description. *Autom.*, 14(5):465–471, 1978. doi: 10.1016/0005-1098(78)90005-5. URL `https://doi.org/10.1016/0005-1098(78)90005-5`.

[65] G. Roeder, L. Metz, and D. Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR, 2021.

[66] E. Rosenfeld, P. K. Ravikumar, and A. Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BbNIbVPJ-42`.

[67] E. Rosenfeld, P. Ravikumar, and A. Risteski. Domain-adjusted regression or: Erm may already learn features sufficient for out-of-distribution generalization. *arXiv preprint arXiv:2202.06856*, 2022.

[68] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2): 215–246, 2021.

[69] B. Schölkopf and J. von Kügelgen. From statistical to causal learning. *arXiv preprint arXiv:2204.00607*, 2022.

[70] B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. arXiv:2102.11107.

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

[72] P. Shinners et al. Pygame. *Dostupné z: http://pygame. org/[Online (2011)*, 2011.

[73] R. Silva, R. Scheines, C. Glymour, P. Spirtes, and D. M. Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

[74] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

[75] P. Spirtes and C. Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[76] P. Spirtes, C. N. Glymour, and R. Scheines. *Causation, prediction, and search*. MIT press, 2000.

[77] C. Squires and C. Uhler. Causal structure learning: a combinatorial perspective. *Foundations of Computational Mathematics*, pages 1–35, 2022.

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

[79] C. Squires, A. Seigal, S. Bhate, and C. Uhler. Linear causal disentanglement via interventions, 2023.

[80] S. G. Stark, J. Ficek, F. Locatello, X. Bonilla, S. Chevrier, F. Singer, G. Rätsch, and K.-V. Lehmann. Scim: universal single-cell matching with unpaired feature sets. *Bioinformatics*, 36(Supplement_2): i919–i927, 2020.

[81] T. Teshima, I. Ishikawa, K. Tojo, K. Oono, M. Ikeda, and M. Sugiyama. Coupling-based invertible neural networks are universal diffeomorphism approximators. *Advances in Neural Information Processing Systems*, 33:3362–3373, 2020.

[82] B. Varici, K. Shanmugam, P. Sattigeri, and A. Tajer. Scalable intervention target estimation in linear models. *Advances in Neural Information Processing Systems*, 34:1494–1505, 2021.

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

[84] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[85] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509. URL `https://books.google.de/books?id=hV8o5R7_5tkC`.

[86] J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in Neural Information Processing Systems*, 34, 2021.

[87] Y. Wang, D. Blei, and J. P. Cunningham. Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34:5443–5455, 2021.

[88] S. Weichwald, S. W. Mogensen, T. E. Lee, D. Baumann, O. Kroemer, I. Guyon, S. Trimpe, J. Peters, and N. Pfister. Learning by doing: Controlling a dynamical system using causality, control, and reinforcement learning. In *NeurIPS 2021 Competitions and Demonstrations Track*, pages 246–258. PMLR, 2022.

[89] M. Willetts and B. Paige. I don't need **u**: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

[90] F. Xie, R. Cai, B. Huang, C. Glymour, Z. Hao, and K. Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

[91] M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9593–9602, June 2021.

[92] Y. Yu, J. Chen, T. Gao, and M. Yu. Dag-gnn: Dag structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019.

[93] J. Zhang, C. Squires, and C. Uhler. Matching a desired causal state via shift interventions. *Advances in Neural Information Processing Systems*, 34:19923–19934, 2021.

[94] S. Zhao, J. Song, and S. Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the aaai conference on artificial intelligence*, volume 33, pages 5885–5892, 2019.

[95] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

[96] Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *arXiv preprint arXiv:2206.07751*, 2022.

[97] R. S. Zimmermann, Y. Sharma, S. Schneider, M. Bethge, and W. Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pages 12979–12990. PMLR, 2021.

A Proof of Identifiability up to Linear Maps

In this section we give a full proof of Theorem 3. We try to make this essential self-contained. To make this easy to read, we first introduce and recall the required notation in Section A.1, then we prove two important key relations required for the proof in Section A.2. Finally, we prove first a simpler version of Theorem 3 with more assumptions in Section A.3 and then give the full proof of Theorem 3 in Section A.4.

A.1 General Notation

Let us introduce some notation for our identifiability proofs. For a full list of assumptions we refer to the main text, here we just recall the relevant notation concisely. We will assume that there are two representations that generate the same observations, i.e., we assume that there are two sets of Gaussian SCMs

$$Z^{(i)} = A^{(i)} Z^{(i)} + (D^{(i)})^{\frac{1}{2}}(\epsilon + \eta^{(i)} e_{t_i}) \tag{10}$$

$$\widetilde{Z}^{(i)} = \widetilde{A}^{(i)} \widetilde{Z}^{(i)} + (\widetilde{D}^{(i)})^{\frac{1}{2}}(\widetilde{\epsilon} + \widetilde{\eta}^{(i)} e_{\widetilde{t}_i}) \tag{11}$$

where $\epsilon, \widetilde{\epsilon} \sim N(0, \mathrm{Id})$ and $i = 0$ corresponds to the observational distribution while $i > 0$ correspond to interventional settings where we intervene on a single node t_i (or $\widetilde{t}_i$) by adding a shift and changing the relation to its parents (without adding new parents). Moreover, there are two functions f, $\widetilde{f}$ such that $X^{(i)} \overset{\mathcal{D}}{=} f(Z^{(i)})$ and $X^{(i)} \overset{\mathcal{D}}{=} \widetilde{f}(\widetilde{Z}^{(i)})$. It is convenient to define

$$B^{(i)} = (D^{(i)})^{-\frac{1}{2}}(\mathrm{Id} - A^{(i)}) \tag{12}$$

$$\mu^{(i)} = \eta^{(i)}(B^{(i)})^{-1} e_{t_i} \tag{13}$$

so that $Z^{(i)} = (B^{(i)})^{-1}\epsilon + \mu^{(i)}$ with $\epsilon \sim N(0, \mathrm{Id})$. Note that all model information can be collected in $((B^{(i)}, \eta^{(i)}, t_i), f)$.

As in the main text, it is convenient to use the shorthand $B = B^{(0)}$, $A = A^{(0)}$.

We denote the row t_i of $B^{(i)}$ by $r^{(i)}$, i.e.,

$$r^{(i)} = (B^{(i)})^\top e_{t_i}. \tag{14}$$

Note that for perfect interventions where we cut the relation to all parents of node t_i we have $r^{(i)} = \lambda_i e_{t_i}$ for some $\lambda_i \neq 0$. Similarly, for the observational distribution we refer to the row t_i of $B^{(0)}$ by $s^{(i)}$, i.e.,

$$s^{(i)} = (B^{(0)})^\top e_{t_i}. \tag{15}$$

We thus find

$$B^{(0)} - B^{(i)} = e_{t_i}(e_{t_i}^\top B^{(0)} - e_{t_i}^\top B^{(i)}) = e_{t_i}(s^{(i)} - r^{(i)})^\top. \tag{16}$$

We also consider the covariance matrix $\Sigma^{(i)}$ and the precision matrix $\Theta^{(i)}$ of $Z^{(i)}$ which are given by

$$\Theta^{(i)} = (B^{(i)})^\top B^{(i)}, \quad \Sigma^{(i)} = (\Theta^{(i)})^{-1} \tag{17}$$

Indeed,

$$\Sigma^{(i)} = \mathbb{E}[(B^{(i)})^{-1}\epsilon\epsilon^\top (B^{(i)})^{-\top}] = (B^{(i)})^{-1}\mathbb{E}[\epsilon\epsilon^\top](B^{(i)})^{-\top} = (B^{(i)})^{-1}(B^{(i)})^{-\top}$$

Finally, the mean $\mu^{(i)}$ of $Z^{(i)}$ given by

$$\mu^{(i)} = \mathbb{E}Z^{(i)} = \eta^{(i)}(B^{(i)})^{-1} e_{t_i}. \tag{18}$$

We also define similar quantities for $\widetilde{Z}$. To simplify the notation, we will frequently use the shorthand $v \otimes w = v \cdot w^\top \in \mathbb{R}^{d \times d}$ for $v, w \in \mathbb{R}^d$.

A.2 Two auxiliary lemmas

The key to our identifiability results is the relation shown in the following lemma.

Lemma 2. *Assume that there are two latent variable models* $((B^{(i)}, \eta^{(i)}, t_i)_{i\in I}, f)$ *and* $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in I}, \widetilde{f})$ *as defined above such that* $f(Z^{(i)}) \overset{\mathcal{D}}{=} \widetilde{f}(\widetilde{Z}^{(i)})$. *Then* $h = \widetilde{f}^{-1} \circ f$ *exists and is a diffeomorphism, i.e., bijective and differentiable with differentiable inverse. Moreover, it satisfies the relation*

$$\begin{aligned} &\frac{1}{2} z^\top (\Theta^{(i)} - \Theta^{(0)}) z - \eta^{(i)} (r^{(i)})^\top z \\ &\quad = \frac{1}{2} h(z)^\top (\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) h(z) - \widetilde{\eta}^{(i)} (\widetilde{r}^{(i)})^\top h(z) + c^{(i)} \end{aligned} \tag{19}$$

for all z *and* i *and some constant* $c^{(i)}$. *If* $\mu^{(i)} = \widetilde{\mu}^{(i)} = 0$ *this simplifies to*

$$\frac{1}{2} z^\top (\Theta^{(i)} - \Theta^{(0)}) z = \frac{1}{2} h(z)^\top (\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) h(z) + c^{(i)}. \tag{20}$$

Proof. As $Z^{(0)}$ and $\widetilde{Z}^{(0)}$ have full support and f and $\tilde{f}$ are by assumption embeddings, the equality $f(Z^{(0)}) \overset{\mathcal{D}}{=} \widetilde{f}(\widetilde{Z}^{(0)})$ implies that the two image manifolds agree and $h = \widetilde{f}^{-1} \circ f$ is a differentiable map. Moreover, we find that $h_* Z^{(i)} = \widetilde{Z}^{(i)}$ where h_* denotes the pushforward map. The variables $Z^{(i)}$ are Gaussian with distribution $Z^{(i)} \sim N(\mu^{(i)}, \Sigma^{(i)})$. This implies that its density $p^{(i)}$ can be written as

$$\ln(p^{(i)}(z)) = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma^{(i)}| - \frac{1}{2} (z - \mu^{(i)})^\top \Theta^{(i)} (z - \mu^{(i)}). \tag{21}$$

A similar relation holds for $\widetilde{p}^{(i)}(z)$, the density of $\widetilde{Z}^{(i)}$. Finally, we have the relations

$$p^{(i)}(z) = \widetilde{p}^{(i)}(h(z)) |\det J_h(z)| \tag{22}$$

which is a consequence of $h_* Z^{(i)} = \widetilde{Z}^{(i)}$. Here J_h denotes the Jacobian matrix of partial derivatives. Thus we conclude that

$$\begin{aligned} &-\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\Sigma^{(i)}| - \frac{1}{2} (z - \mu^{(i)})^\top \Theta^{(i)} (z - \mu^{(i)}) \\ &\qquad = -\frac{n}{2} \ln(2\pi) - \frac{1}{2} \ln |\widetilde{\Sigma}^{(i)}| - \frac{1}{2} (h(z) - \widetilde{\mu}^{(i)})^\top \widetilde{\Theta}^{(i)} (h(z) - \widetilde{\mu}^{(i)}) + \ln |\det J_h(z)|. \end{aligned} \tag{23}$$

Calling this relation E^i, we consider the difference $E^0 - E^i$. Then the determinant term cancels, and we obtain for some constant $c^{(i)}$

$$\begin{aligned} &\frac{1}{2} (z - \mu^{(i)})^\top \Theta^{(i)} (z - \mu^{(i)}) - \frac{1}{2} z^\top \Theta^{(0)} z \\ &\quad = \frac{1}{2} (h(z) - \widetilde{\mu}^{(i)})^\top \widetilde{\Theta}^{(i)} (h(z) - \widetilde{\mu}^{(i)}) - \frac{1}{2} h(z)^\top \widetilde{\Theta}^{(0)} h(z) + c'^{(i)}. \end{aligned} \tag{24}$$

Expanding the quadratic expression on the left-hand side we obtain

$$\frac{1}{2} (z - \mu^{(i)})^\top \Theta^{(i)} (z - \mu^{(i)}) - \frac{1}{2} z^\top \Theta^{(0)} z = \frac{1}{2} z^\top (\Theta^{(i)} - \Theta) z - z^\top \Theta^{(i)} \mu^{(i)} + \frac{1}{2} (\mu^{(i)})^\top \Theta^{(i)} \mu^{(i)}. \tag{25}$$

Finally we simplify the last term using (18)

$$\Theta^{(i)} \mu^{(i)} = \eta^{(i)} (B^{(i)})^\top B^{(i)} (B^{(i)})^{-1} e_{t_i} = \eta^{(i)} (B^{(i)})^\top e_{t_i} = \eta^{(i)} r^{(i)} \tag{26}$$

Apply the same simplification on the right hand side. Finally, we collect the constant terms independent of z in $c^{(i)}$, to obtain

$$\begin{aligned} &\frac{1}{2} z^\top (\Theta^{(i)} - \Theta^{(0)}) z - \eta^{(i)} (r^{(i)})^\top z \\ &\quad = \frac{1}{2} h(z)^\top (\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) h(z) - \widetilde{\eta}^{(i)} (\widetilde{r}^{(i)})^\top h(z) + c^{(i)} \end{aligned} \tag{27}$$

which is (19) as desired. □

A second key ingredient for our identifiability results is that the specific structure of the interventions that each target a single node which implies that the difference of precision matrices $\Theta^{(i)} - \Theta^{(0)}$ has rank at most 2. This is also the key ingredient used (in a very different manner) in the proof of the main result in [79]. Here we highlight this simple result because we make use of it frequently.

Lemma 3. *The precision matrices satisfy the relation*

$$\Theta^{(i)} - \Theta^{(0)} = r^{(i)} \otimes r^{(i)} - s^{(i)} \otimes s^{(i)}. \tag{28}$$

Proof. Using (17) we find

$$\begin{aligned}
\Theta^{(i)} - \Theta^{(0)} &= (B^{(i)})^\top B^{(i)} - B^\top B \\
&= (B^{(i)})^\top \left(\sum_{k=1}^{d} e_k e_k^\top \right) B^{(i)} - B^\top \left(\sum_{k=1}^{d} e_k e_k^\top \right) B \\
&= \sum_{k=1}^{d} (B^{(i)})^\top e_k ((B^{(i)})^\top e_k)^\top - \sum_{k=1}^{d} B^\top e_k (B^\top e_k)^\top \\
&= ((B^{(i)})^\top e_{t_i}) \otimes ((B^{(i)})^\top e_{t_i}) - (B^\top e_{t_i}) \otimes (B^\top e_{t_i}) \\
&= r^{(i)} \otimes r^{(i)} - s^{(i)} \otimes s^{(i)}.
\end{aligned} \tag{29}$$

Here we used the fact that all rows except row t_i of $B^{(i)}$ and $B^{(0)}$ agree as the intervention targets a single node, i.e., $(B^{(i)})^\top e_k = (B^{(0)})^\top e_k$ for $k \neq t_i$. □

A.3 Warm-up

To provide some intuition how identifiability arises, we first provide the proof for a simpler result with a much stronger assumption on the set of intervention targets and the type of interventions. But note importantly that we don't relax the assumption of non-linearity of f.

Theorem 4. *Let $f(Z^{(i)}) \overset{\mathcal{D}}{=} \widetilde{f}(\widetilde{Z}^{(i)})$ for two sets of parameters $((B^{(i)}, \eta^{(i)}, t_i)_{i\in I}, f)$ and $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in I}, \widetilde{f})$ as in the above model. We assume that there are $2n$ perfect pairwise different interventions (i.e., $Z^{(i)} \overset{\mathcal{D}}{\neq} Z^{(i')}$ for $i \neq i'$), without shifts (i.e., $\eta^{(i)} = \widetilde{\eta}^{(i)} = 0$). Finally, we assume that there are 2 interventions for each node, and we know the intervention targets of the interventions, i.e., we assume that $t_i = \widetilde{t}_i$ for all i. Then, $Z^{(i)}$ and $\tilde{Z}^{(i)}$ agree up to scaling, i.e., $Z^{(i)} = D\tilde{Z}^{(i)}$ for a diagonal matrix $D \in \mathrm{Diag}(n)$.*

Proof. We consider a node k and two interventions i_1 and i_2 which target node $k = t_{i_1} = t_{i_2}$ (by assumption those interventions exist and i_1 and i_2 agree for both representations). Since we assumed that the interventions are perfect, we have in our notation that

$$r^{(i_1)} = \lambda_{i_1} e_k, \quad r^{(i_2)} = \lambda_{i_2} e_k \tag{30}$$

for two constants $\lambda_{i_1} \neq \lambda_{i_2} > 0$ and similarly for $\widetilde{r}^{(i_j)}$. Indeed, assuming positivity is not a restriction since ϵ follows a symmetric distribution (this is also contained in the parametrisation $(D^{(i)})^{\frac{1}{2}}$) and $\lambda_{i_1} \neq \lambda_{i_2}$ is necessary to ensure that the interventional distributions differ. Combining (20) from Lemma 2 with Lemma 3 we obtain the relation

$$\begin{aligned}
z^\top \left(r^{(i_j)} \otimes r^{(i_j)} - s^{(i_j)} \otimes s^{(i_j)} \right) z &= z^\top (\Theta^{(i_j)} - \Theta^{(0)}) z \\
&= h(z)^\top (\widetilde{\Theta}^{(i_j)} - \widetilde{\Theta}^{(0)}) h(z) + 2c^{(i_j)} \\
&= h(z)^\top \left(\widetilde{r}^{(i_j)} \otimes \widetilde{r}^{(i_j)} - \widetilde{s}^{(i_j)} \otimes \widetilde{s}^{(i_j)} \right) h(z) + 2c^{(i_j)}.
\end{aligned} \tag{31}$$

for $j = 1, 2$.

Note that since $k = t_{i_1} = t_{i_2}$, we have $s^{(i_1)} = s^{(i_2)} = B^\top e_k$ by definition. Thus, we subtract the relation in the last display for i_1 from the relation for i_2 and find using (30),

$$\begin{aligned}
(\lambda_{i_1}^2 - \lambda_{i_2}^2) z_k^2 &= (z^\top r^{(i_1)})^2 - (z^\top r^{(i_2)})^2 \\
&= (h(z)^\top \widetilde{r}^{(i_1)})^2 - (h(z)^\top \widetilde{r}^{(i_2)})^2 + 2(c^{(i_1)} - c^{(i_2)}) \\
&= (\widetilde{\lambda}_{i_1}^2 - \widetilde{\lambda}_{i_2}^2) h_k(z)^2 + 2(c^{(i_1)} - c^{(i_2)}).
\end{aligned} \tag{32}$$

for all z. Since $\widetilde{\lambda}_{i_1} \neq \widetilde{\lambda}_{i_2} > 0$ we can divide and obtain

$$\frac{\lambda_{i_1}^2 - \lambda_{i_2}^2}{\widetilde{\lambda}_{i_1}^2 - \widetilde{\lambda}_{i_2}^2} z_k^2 = h_k(z)^2 + 2\frac{c^{(i_1)} - c^{(i_2)}}{\widetilde{\lambda}_{i_1}^2 - \widetilde{\lambda}_{i_2}^2}. \tag{33}$$

Since h is surjective, we have that h_k is also surjective which implies that the range of the right hand side is $[2(c^{(i_1)} - c^{(i_2)})/(\widetilde{\lambda}_{i_1}^2 - \widetilde{\lambda}_{i_2}^2), \infty)$. The range of the left hand side is depending on the sign of the prefactor and is either $(-\infty, 0]$ or $[0, \infty)$. Since the ranges have to agree we conclude that $c^{(i_1)} = c^{(i_2)}$ and

$$h_k(z) = \pm\sqrt{\frac{\lambda_{i_1}^2 - \lambda_{i_2}^2}{\widetilde{\lambda}_{i_1}^2 - \widetilde{\lambda}_{i_2}^2}} z_k. \tag{34}$$

Since k was arbitrary, this ends the proof. □

A.4 Proof of identifiability up to linear maps

The proof of our main result relies on two essentially geometric lemmas and a slight extension of Lemma 2 which we discuss now. The first lemma alone is sufficient to handle the case where interventions change the scaling matrix D, while the second is required to handle interventions that change B and the third allows us to consider pure shift interventions.

Lemma 4. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous and surjective function, Q a quadratic form on $\mathbb{R}^d$ and $\alpha \in \mathbb{R}$ a constant such that*

$$g(z)^2 = \langle z, Qz \rangle + \alpha. \tag{35}$$

Then Q has rank 1 and g is linear, i.e., $g(z) = v \cdot z$ for some $v \in \mathbb{R}^d$.

Proof. Assume that Q has the eigendecomposition $Q = \sum_{i=1}^r \lambda_i v_i v_i^\top$ where r is the rank of Q and $\lambda_i \neq 0$ and v_i has unit norm. Clearly g surjective implies $r > 0$. Since g is surjective we find that the range of the left hand side is $[0, \infty)$ and therefore $\langle z, Qz \rangle \geq -\alpha$. This implies that $\lambda_i \geq 0$ for all i and therefore, the range of the right hand side is $[\alpha, \infty)$ which implies $\alpha = 0$. Now we fix any $c > 0$ and consider $E_c = \{z \in \mathbb{R}^d | \, \langle z, Qz \rangle = c\}$. It is easy to see that E_c is the product of an ellipsoid and $\mathbb{R}^{d-r}$ and thus is connected if $r \geq 2$. We find by (35) that $g(z) = \pm\sqrt{c}$ for $z \in E_c$ and if $g(z) \in \{-\sqrt{c}, \sqrt{c}\}$ then $z \in E_c$. By continuity of g and since E_c is connected this implies that $g(z) = \sqrt{c}$ for all $z \in E_c$ or $g(z) = -\sqrt{c}$ for all $z \in E_c$ contradicting the surjectivity of g. We conclude that $r = 1$, i.e., $Q = \lambda_1 v_1 v_1^\top = \tilde{v}_1 \tilde{v}_1^\top$ with $\tilde{v}_1 = \sqrt{\lambda_1} v_1$. Thus we find that

$$g(z)^2 = |\tilde{v}_1 \cdot z|^2. \tag{36}$$

Since g is continuous, this implies that $g(z) = \pm\tilde{v}_1 \cdot z$ or $g(z) = \pm|\tilde{v}_1 \cdot z|$. As only the first functions are surjective we conclude that $g(z) = \pm\tilde{v}_1 \cdot z$ and the claim follows with $w = \pm\tilde{v}_1$. □

To handle shift interventions we need a slightly stronger version of the lemma above which contains an additional linear term.

Corollary 1. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be a continuous and surjective function, Q a quadratic form on $\mathbb{R}^d$, $w \in \mathbb{R}^d$ a vector, and $\alpha \in \mathbb{R}$ a constant such that*

$$g(z)^2 = \langle z, Qz \rangle + w \cdot z + \alpha. \tag{37}$$

Then Q has rank 1 and g is affine, i.e., $g(z) = v \cdot z + c$ for some $v \in \mathbb{R}^d$ and $c \in \mathbb{R}$.

Proof. Assume as before that Q has the eigendecomposition $Q = \sum_{i=1}^r \lambda_i v_i v_i^\top$ where r is the rank of Q and $\lambda_i \neq 0$ and $|v_i| = 1$. First, we show that $w \in \langle v_1, \ldots, v_r \rangle$ where $\langle . \rangle$ denotes the span. Indeed, suppose that this is not the case. Then we could find z_0 such that $z_0 \cdot w < 0$ and $z_0 v_i = 0$ for $1 \leq i \leq r$, and therefore $\langle z_0, Qz_0 \rangle = 0$. Plugging $z = \lambda z_0$ for $\lambda \to \infty$ in (37) the right-hand side tends to $-\infty$ while the left-hand side is non-negative. This is a contradiction and therefore $w \in \langle v_1, \ldots, v_r \rangle$ holds. This implies that there is w' such that $Qw' = w/2$ and thus

$$\langle z, Qz \rangle + w \cdot z + \alpha = \langle (z + w'), Q(z + w') \rangle + \alpha - \langle w', Qw' \rangle. \tag{38}$$

Then we consider $\widetilde{g}(\widetilde{z}) = g(\widetilde{z} - w')$, $\widetilde{\alpha} = \alpha - \langle w', Qw' \rangle$ which satisfy

$$\begin{aligned} \widetilde{g}^2(\widetilde{z}) &= g^2(\widetilde{z} - w') && (39)\\ &= \langle (\widetilde{z} - w' + w'), Q(\widetilde{z} - w' + w') \rangle + \alpha - \langle w', Qw' \rangle && (40)\\ &= \langle \widetilde{z}, Q\widetilde{z} \rangle + \widetilde{\alpha}. && (41) \end{aligned}$$

Using Lemma 4 we conclude that Q has rank 1 and $\widetilde{g}(\widetilde{z}) = v \cdot z$ for some v. But then $g(z) = \widetilde{g}(z + w') = z \cdot v + w' \cdot v$ which concludes the proof. □

The second lemma is similar, but slightly simpler.

Lemma 5. *Assume that there is a quadratic form Q, a vector $0 \neq w \in \mathbb{R}^d$, $w' \in \mathbb{R}^d$, $\alpha \in \mathbb{R}$ and a continuous function $g : \mathbb{R}^d \to \mathbb{R}$ such that*

$$g(z)(w \cdot z) = z \cdot Qz + w' \cdot z + \alpha. \tag{42}$$

Then $g(z) = v \cdot z + c$ for some $v \in \mathbb{R}^d$, $c \in \mathbb{R}$, i.e., g is an affine function.

Proof. By rescaling we can assume that w has unit norm. Plugging in $z = 0$ we find that $\alpha = 0$. Denote by Π_w the orthogonal projection on w, i.e., $\Pi_w v = (w \cdot v)w$ and by $\Pi_w^\perp$ the projection on the orthogonal complement of w, i.e., $z = \Pi_w z + \Pi_w^\perp z$ for all $z \in \mathbb{R}^d$. Then we find for all z

$$0 = g(\Pi_w^\perp z) w \cdot \Pi_w^\perp z = (\Pi_w^\perp z) \cdot Q\Pi_w^\perp z + (\Pi_w^\perp z) \cdot w'. \tag{43}$$

By scaling z we find that both terms on the right hand side vanish and thus for all z

$$\begin{aligned} (\Pi_w^\perp z) \cdot Q\Pi_w^\perp z &= 0, && (44)\\ (\Pi_w^\perp z) \cdot w' &= 0. && (45) \end{aligned}$$

This implies (using also the symmetry of Q)

$$\begin{aligned} g(z) w \cdot \Pi_w z &= g(z) w \cdot z \\ &= z \cdot Qz + w' \cdot z \\ &= (\Pi_w z + \Pi_w^\perp z) \cdot Q(\Pi_w z + \Pi_w^\perp z) + w' \cdot (\Pi_w z + \Pi_w^\perp z) \\ &= (\Pi_w z) \cdot Q\Pi_w z + 2(\Pi_w z) Q \Pi_w^\perp z + (\Pi_w^\perp z) \cdot Q \Pi_w^\perp z + \Pi_w z \cdot w' \\ &= (\Pi_w z) \cdot (Q(\Pi_w z + 2\Pi_w^\perp z) + w'). \end{aligned} \tag{46}$$

Now we use $\Pi_w z = (w \cdot \Pi_w z) w$ and find

$$g(z) w \cdot \Pi_w z = (w \cdot \Pi_w z) w \cdot (Q(\Pi_w z + 2\Pi_w^\perp z) + w'). \tag{47}$$

This implies

$$g(z) = w \cdot Q(\Pi_w z + 2\Pi_w^\perp z) + w \cdot w' \tag{48}$$

for all z such that $\Pi_w z \neq 0$. By continuity we conclude that this holds for all z, i.e., g is affine. □

Again we need a slight generalization of this lemma.

Corollary 2. *Assume that there is a quadratic form Q, a vector $0 \neq w \in \mathbb{R}^d$, $w' \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$ and a continuous function $g : \mathbb{R}^d \to \mathbb{R}$ such that*

$$g(z)(w \cdot z + \beta) = z \cdot Qz + w' \cdot z + \alpha. \tag{49}$$

Then $g(z) = v \cdot z + c$ for some $v \in \mathbb{R}^d$, $c \in \mathbb{R}$, i.e., g is an affine function.

Proof. Define $\widetilde{g}(\widetilde{z}) = g(\widetilde{z} - \beta w |w|^{-2})$. Then

$$\begin{aligned} \widetilde{g}(\widetilde{z})(w \cdot \widetilde{z}) &= g(\widetilde{z} - \beta w |w|^{-2})(w \cdot (\widetilde{z} - \beta w |w|^{-2}) + \beta) \\ &= (\widetilde{z} - \beta w |w|^{-2}) \cdot Q(\widetilde{z} - \beta w |w|^{-2}) + w' \cdot (\widetilde{z} - \beta w |w|^{-2}) + \alpha \\ &= \widetilde{z} \cdot Q\widetilde{z} + \widetilde{w} \cdot \widetilde{z} + \widetilde{\alpha} \end{aligned} \tag{50}$$

for some $\widetilde{w}$ and $\widetilde{\alpha}$. So, we conclude from Lemma 5 that $\widetilde{g}$ is affine and thus the same is true for g. □

We now discuss a small extension of Lemma 2 by showing that we can complete the squares in (19) which then allows us to obtain a relation that is very similar to (20).

Lemma 6. *Assume the general setup as introduced above and also $\Theta^{(i)} \neq \Theta$. Then there is a vector $\mu'^{(i)}$ and a constant c such that*

$$(z-\mu^{(i)})^\top \Theta^{(i)}(z-\mu^{(i)}) - z^\top \Theta^{(0)} z = (z-\mu'^{(i)})^\top(\Theta^{(i)}-\Theta^{(0)})(z-\mu'^{(i)}) + c. \tag{51}$$

Proof. Using the symmetry of the precision matrices and expanding all the terms, we can express the difference of the left-hand side and the right-hand side of (51) as

$$\begin{aligned}&(z-\mu^{(i)})^\top \Theta^{(i)}(z-\mu^{(i)}) - z^\top \Theta^{(0)} z - (z-\mu'^{(i)})^\top(\Theta^{(i)}-\Theta^{(0)})(z-\mu'^{(i)}) - c \\ &= -2z^\top \Theta^{(i)}\mu^{(i)} + 2z^\top(\Theta^{(i)}-\Theta^{(0)})\mu'^{(i)} + (\mu^{(i)})^\top\Theta^{(i)}\mu^{(i)} - (\mu'^{(i)})^\top(\Theta^{(i)}-\Theta^{(0)})\mu'^{(i)} - c.\end{aligned} \tag{52}$$

Hence, we can set c at the end such the constant terms cancel and all we have to show is that there exists $\mu'^{(i)}$ such that

$$(\Theta^{(i)}-\Theta)\mu'^{(i)} = \Theta^{(i)}\mu^{(i)}. \tag{53}$$

While this is clearly not true for arbitrary $\mu^{(i)}$ because $\Theta^{(i)}-\Theta$ has only rank 2, such a $\mu'^{(i)}$ does exist for $\mu^{(i)} = \eta^{(i)}(B^{(i)})^{-1}e_{t_i}$ (see equation (18)). Indeed, we can simplify using (17) and Lemma 3

$$(\Theta^{(i)}-\Theta)\mu'^{(i)} = r^{(i)}((r^{(i)})^\top\mu'^{(i)}) - s^{(i)}((s^{(i)})^\top\mu'^{(i)}) \tag{54}$$

$$\Theta^{(i)}\mu^{(i)} = \eta^{(i)}(B^{(i)})^\top B^{(i)}(B^{(i)})^{-1}e_{t_i} = \eta^{(i)}(B^{(i)})^\top e_{t_i} = \eta^{(i)}r^{(i)}. \tag{55}$$

Now there are two cases. When $r^{(i)}$ and $s^{(i)}$ are collinear (but not identical, by assumption) we can choose $\mu'^{(i)}) = \lambda r^{(i)}$ for a suitable λ. Otherwise, we can pick any vector $\mu'^{(i)}$ such that $\mu'^{(i)}s^{(i)} = 0$, $\mu'^{(i)}r^{(i)} = \eta^{(i)}$. This is always possible for non-collinear vectors (project $r^{(i)}$ on the orthogonal complement of $s^{(i)}$). □

Let us now first discuss a rough sketch of the proof of our main result. This allows us to present the main ideas without discussing all the technical details required for the full proof of the result. We restate the theorem for convenience.

Theorem 3. *Assume that $X^{(i)}$ is generated according to a model $((B^{(i)},\eta^{(i)},t_i)_{i\in\bar{I}},f)$ such that the Assumptions 1-4 hold. Then we have identifiability up to linear transformations, i.e., if $((\widetilde{B}^{(i)},\widetilde{\eta}^{(i)},\widetilde{t}_i),\widetilde{f})$ generates the same observed distributions $\widetilde{f}(\widetilde{Z}^{(i)}) \overset{\mathcal{D}}{=} X^{(i)}$, then there is an invertible linear map T such that*

$$\widetilde{f} = f\circ T^{-1}, \quad \widetilde{Z}^{(i)} = TZ^{(i)}. \tag{5}$$

Proof Sketch. For simplicity of the sketch we ignore shift interventions here. We assume that we have two sets of parameters $((B^{(i)},t_i)_{i\in\bar{I}},f)$, $((\widetilde{B}^{(i)},\widetilde{t}_i)_{i\in\bar{I}}),\widetilde{f})$ generating the same observations. We can relabel the interventions i and the variables $\widetilde{Z}_k$ such that $\widetilde{t}_i = i$ and such that $1,2,\dots$ is a valid causal order of the graph. The general idea is to show by induction that $(h_1(z),\dots,h_k(z))$ is a linear function of z. Let us sketch how this is achieved. The key ingredients here are Lemma 2 and Lemma 3. Applying (20) from Lemma 2 in combination with Lemma 6 we obtain

$$\frac{1}{2}z^\top(\Theta^{(k+1)}-\Theta^{(0)})z = \frac{1}{2}h(z)^\top(\widetilde{r}^{(k+1)}\otimes\widetilde{r}^{(k+1)} - \widetilde{s}^{(k+1)}\otimes\widetilde{s}^{(k+1)})h(z) + c^{(k+1)}. \tag{56}$$

By our assumption that the variables $\widetilde{Z}^{(i)}$ follow the causal order of the underlying graph, only the entries $\widetilde{r}_r^{(k+1)}$ for $r \leq k+1$ are non-zero and similarly for $\widetilde{s}^{(k+1)}$. Therefore, the right-hand side is a quadratic form of $(h_1(z),\dots,h_{k+1}(z))$. Now we apply our induction hypothesis which states that $(h_1(z),\dots,h_k(z))$ is a linear function of z. Then we find expanding the quadratic form on the right hand side that there is a matrix Q, a vector v, a constant δ (which we assume to be non-zero) and another quadratic form Q' such that

$$\frac{1}{2}z^\top Q'z = \frac{1}{2}z^\top Qz + (v\cdot z)h_{k+1}(z) + \delta h_{k+1}^2(z) + c^{(k+1)} \tag{57}$$

$$= \frac{1}{2}z^\top Qz + \delta\left(h_{k+1}(z) + \frac{(v\cdot z)}{2\delta}\right)^2 - \frac{(v\cdot z)^2}{4\delta} + c^{(k+1)}. \tag{58}$$

Now we can apply Corollary 1 to conclude that $h_{k+1}(z)$ is a linear function of z. Adding the shifts and considering also the case $\delta = 0$ adds several technical difficulties, which we handle in the full proof. □

After all those preparations and presenting the proof sketch, we will finally prove our main theorem in full generality.

Full proof of Theorem 3. As the proof has to consider different cases and is a bit technical we structure it in a series of steps.

Labeling assumptions We can assume without loss of generality (by relabeling variables and interventions) that the variables $\widetilde{Z}_i$ are ordered such that their natural order $i = 1, 2, \dots$ is a valid topological order of the underlying DAG and that intervention $1 \leq i \leq n$ targets node Z_i, i.e., $\widetilde{t}_i = i$. We make no assumption on the ordering of the Z_i or targets of t_i which are not necessarily assumed to be pairwise different. Recall that we defined (and showed existence of) $h = \widetilde{f}^{-1} f$ in Lemma 2.

Induction claim We now prove by induction that $h_k(z)$ is linear for $k \geq 0$. The base case $k = 0$ is trivially true. For the induction step, assume that this is the case for $j \leq k$, i.e., $h_j(z) = m_j \cdot z$ for some $m_j \in \mathbb{R}^d$. We define $M_k = (m_1, \dots, m_k)^\top$ so that we can write concisely

$$(h_1(z), \dots, h_k(z))^\top = M_k z. \tag{59}$$

To prove the induction step, we need to show that there exists $m_{k+1} \in \mathbb{R}^d$ such that $h_{k+1}(z) = m_{k+1} \cdot z$.

Precision matrix decomposition For the DAG $\widetilde{G}$ underlying $\widetilde{Z}_i$, let PA_i denote the indices of the set of parents of Z_i and let $\overline{\mathrm{PA}}_i = \{i\} \cup \mathrm{PA}_i$. Recall the notation $\widetilde{s}^{(i)} = \widetilde{B}^\top e_i$ (using $\widetilde{t}_i = i$) and $\widetilde{r}^{(i)} = (\widetilde{B}^{(i)})^\top e_i$. Now $\widetilde{s}^{(i)}_j \neq 0$ only holds if $j \in \overline{\mathrm{PA}}_i$ and we ordered the variables such that $\overline{\mathrm{PA}}_i \subset [i]$. In particular, $\widetilde{s}^{(i)}_j = 0$ and $\widetilde{r}^{(i)}_j = 0$ for $j > i$ (interventions do not create new parents). By Lemma 3 we have

$$(\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) = \widetilde{r}^{(i)} \otimes \widetilde{r}^{(i)} - \widetilde{s}^{(i)} \otimes \widetilde{s}^{(i)} \tag{60}$$

We find that

$$(\widetilde{\Theta}^{(k+1)} - \widetilde{\Theta}^{(0)})_{rs} = \widetilde{r}^{(k+1)}_r \widetilde{r}^{(k+1)}_s - \widetilde{s}^{(k+1)}_r \widetilde{s}^{(k+1)}_s = 0 \tag{61}$$

if $r > k+1$ or $s > k+1$. In particular, there exists a symmetric $A \in \mathbb{R}^{(k+1)\times(k+1)}$ such that

$$h(z)^\top (\widetilde{\Theta}^{(i)} - \widetilde{\Theta}^{(0)}) h(z) = \sum_{r,s=1}^{k+1} h_r(z) A_{rs} h_s(z). \tag{62}$$

Now we decompose A as

$$A = \begin{pmatrix} A' & b \\ b^\top & \delta \end{pmatrix} \tag{63}$$

for some $A' \in \mathbb{R}^{k\times k}$, $b \in \mathbb{R}^k$ and $\delta \in \mathbb{R}$. Using the induction hypothesis we find

$$\begin{aligned} h(z)^\top (\tilde{\Theta}^{(k+1)} - \tilde{\Theta}^{(0)}) h(z) &= \sum_{r,s=1}^{k+1} h_r(z) A_{rs} h_s(z) \\ &= M_k z \cdot A' M_k z + 2 h_{k+1}(z) b^\top M_k z + \delta h_{k+1}(z)^2. \end{aligned} \tag{64}$$

We use the notation $v_{:r} \in \mathbb{R}^r$ for the vector $(v_1, \dots, v_r)^\top$. Then we can write

$$\eta^{(k+1)} \widetilde{r}^{(k+1)} h(z) = \eta^{(k+1)} \widetilde{r}^{(k+1)}_{:k} M_k z + \eta^{(k+1)} \widetilde{r}^{(k+1)}_{k+1} h_{k+1}(z) \tag{65}$$

Now we have to consider different cases.

Case $\delta \neq 0$ We assume first that $\delta \neq 0$. In this case we find using the last two displays

$$\begin{aligned} &h(z)^\top(\tilde{\Theta}^{(k+1)} - \tilde{\Theta}^{(0)})h(z) + 2\eta^{(k+1)}\widetilde{r}^{(k+1)}h(z) + c^{(k+1)} \\ &= z \cdot M_k^\top A' M_k z + \delta\left(h_{k+1}(z) + \frac{v^\top M_k z + \eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)}}{\delta}\right)^2 - \frac{1}{\delta} z \cdot M_k^\top b b^\top M_k z \\ &- \frac{1}{\delta}\eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)} b^\top M_k z - \frac{1}{\delta}(\eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)})^2 + 2\eta^{(k+1)}\widetilde{r}_{:k}^{(k+1)} \cdot M_k z + c^{(k+1)}. \end{aligned} \tag{66}$$

Next we set

$$g(z) = h_{k+1}(z) + \delta^{-1} b \cdot M_k z + \delta^{-1}\eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)}. \tag{67}$$

Looking carefully at (66) we find that there is a symmetric matrix $\widetilde{Q}$, a vector $\widetilde{w}$, and a constant $\widetilde{c}$ such that

$$h(z)^\top(\tilde{\Theta}^{(k+1)} - \tilde{\Theta}^{(0)})h(z) - 2\eta^{(k+1)}\widetilde{r}^{(k+1)}h(z) = \delta g(z)^2 + z \cdot \widetilde{Q} z + \widetilde{w} z + \widetilde{c}. \tag{68}$$

We claim that g is continuous and surjective. Continuity follows directly from continuity of h. To show that g is surjective we can ignore the constant shift and we note that $\delta^{-1} v^\top M_k z = \delta^{-1}\sum_{r=1}^k v_r h_r(z)$ and thus surjectivity of h implies that g is surjective (for every c pick z such that $h_{k+1}(z) = c$ and $h_r(z) = 0$ for $r \leq k$).

Using (19) from Lemma 2 we conclude that

$$\delta g(z)^2 = z^\top(\Theta^{(k+1)} - \Theta^{(0)})z - 2\eta^{(k+1)} r^{(k+1)} z - z \cdot \widetilde{Q} z - \widetilde{w} z - \widetilde{c} \tag{69}$$

$$= z \cdot Q z + w \cdot z + \alpha \tag{70}$$

for all z and some symmetric $Q \in \mathbb{R}^{d \times d}$. Corollary 1 then implies that

$$h_{k+1}(z) + \delta^{-1} v M_k z + \delta^{-1}\eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)} = g(z) = v \cdot z + c \tag{71}$$

for some $v \in \mathbb{R}^d$, $c \in \mathbb{R}$. Thus h_{k+1} is an affine function, i.e., $h_{k+1}(z) = m_{k+1} z + \alpha$. But

$$0 = \mathbb{E}\widetilde{Z}_{k+1} = \mathbb{E}Z \cdot m_{k+1} + \alpha = \alpha \tag{72}$$

and therefore h_{k+1} is linear.

Case $\delta = 0$, $b \neq 0$ Now we consider the case that $0 = \delta$. Using (64), (65), and (19) we find similarly to above

$$\begin{aligned} M_k z \cdot A M_k z + 2h_{k+1}(z) b^\top M_k z - 2\eta^{(k+1)}\widetilde{r}_{:k}^{(k+1)} \cdot M_k z - 2\eta^{(k+1)}\widetilde{r}_{k+1}^{(k+1)} h_{k+1}(z) + \widetilde{c}^{(k+1)} \\ = z^\top(\Theta^{(k+1)} - \Theta^{(0)})z - 2\eta^{(k+1)} r^{(k+1)} z \end{aligned} \tag{73}$$

Collecting all terms we find that h_{k+1} satisfies a relation of the type

$$h_k(z)(w \cdot z + \beta) = z \cdot Q z + w' \cdot z + \alpha \tag{74}$$

for suitable Q, β, w, w', and α. Note in particular that $w = 2b^\top M_k$. If $w \neq 0$ then Corollary 2 implies that $h_k(z)$ is affine and, as argued above, we conclude that h_k is linear.

Case $\delta = 0$, $b = 0$ It remains to address the case $w = 2b^\top M_k = 0$ and $\delta = 0$. Since h is invertible M_k has full rank, so we conclude that then $b = 0$. Going back to the definition of b and δ (see (62) and (63)) we find

$$(b^\top, \delta)^\top = \widetilde{r}_{:(k+1)}^{(k+1)}\widetilde{r}_{k+1}^{(k+1)} - \widetilde{s}_{:(k+1)}^{(k+1)}\widetilde{s}_{k+1}^{(k+1)} \tag{75}$$

Now $\delta = 0$ implies $\widetilde{s}_{k+1}^{(k+1)} = \widetilde{r}_{k+1}^{(k+1)}$ (because they are both positive). But then we conclude $\widetilde{r}^{(k+1)} = \widetilde{s}^{(k+1)}$ from $b = 0$ and $\widetilde{r}_r^{(k+1)} = \widetilde{s}_r^{(k+1)} = 0$ for $r > k+1$. This implies

$$\widetilde{\Theta}^{(k+1)} - \widetilde{\Theta}^{(0)} = \widetilde{r}^{(k+1)} \otimes \widetilde{r}^{(k+1)} - \widetilde{s}^{(k+1)} \otimes \widetilde{s}^{(k+1)} = 0 \tag{76}$$

That is, this implies that intervention $k+1$ is a pure shift intervention and $\eta^{(k+1)} \neq 0$ (otherwise we do not actually intervene). In this case, we conclude from Lemma 2 and Lemma 6 (if $\Theta^{(k+1)} \neq \Theta$)

$$\eta^{(k+1)}\widetilde{r}^{(k+1)}h(z) + c^{(k+1)} = (z - \mu'^{(k+1)})^\top(\Theta^{(k+1)} - \Theta)(z - \mu'^{(k+1)}) + c. \tag{77}$$

But then $\eta^{(k+1)}\widetilde{r}^{(k+1)}\nabla h(\mu'^{(k+1)}) = 0$, this implies that ∇h is not invertible which is a contradiction to h being a diffeomorphism. Thus, we conclude that also $\Theta^{(k+1)} = \Theta$, i.e., intervention $k+1$ is also a pure shift intervention for the Z variables. But then we find

$$\widetilde{r}_{k+1}^{(k+1)}h_{k+1}(z) + \widetilde{r}_{:k}^{(k+1)}M_k z = (\widetilde{\eta}^{(k+1)})^{-1}(\eta^{(k+1)}r^{(k+1)}z + c^{(k+1)}). \tag{78}$$

So we finally conclude that also in this case h_{k+1} is a linear function, this ends the proof. □

B From Linear Identifiability to Full Identifiability

In this section, we provide the proof of Theorem 1 which is a combination of our linear identifiability result Theorem 3 proved in the previous section and the main result of [79] which shows identifiability for linear mixing f. However, we need to carefully address their normalization choices and in addition consider shifts.

We emphasize that the full identifiability results could also be derived from our proof of Theorem 3 by carefully considering ranks of quadratic forms and showing that we can inductively identify source nodes. We do not add these arguments as our reasoning is already quite complex and because for linear mixing functions this is not substantially different from the original proof in [79].

For the convenience of the reader, we will now restate the main results of [79]. For our work it is convenient to slightly deviate from their conventions, i.e., we define the causal order of a graph by $i \prec j$ iff $i \in \text{an}(j)$, in particular $i \to j$ implies $i \prec j$.

We rephrase their results using our notation. Let $S(G)$ denote the set of permutations of the vertices of G that preserve the causal ordering, i.e., all permutations ρ such that $\rho(i) < \rho(j)$ if $i \prec j$. They consider linear mixing functions as specified next.

Assumption 5. *Assume $f(z) = Lz$ for some matrix $L \in \mathbb{R}^{d' \times d}$ with trivial kernel and pseudoinverse $H = L^+$ such that the absolute value of each row of H is one and the leftmost is positive.*

Then the following result holds.

Theorem 5 (Linear setting, Theorems 1 and 2 in [79])**.** *Suppose latent variables $Z^{(i)}$ are generated following Assumptions 2 and 3 where $\eta^{(i)} = 0$ (i.e., no shift) with DAG G and all $A^{(i)}$ are lower triangular. Suppose in addition that Assumption 4 holds and the mixing function satisfies Assumption 5. Then we can identify the partial order π_G of the underlying DAG G up to permutation of the node labels. If we in addition assume that all interventions are perfect the problem is identifiable in the following sense. For every representation $(\widetilde{B}^{(i)}, \widetilde{H})$ such that all $\widetilde{B}^{(i)}$ are lower triangular and $\widetilde{L}(\widetilde{Z}^{(i)}) \overset{\mathcal{D}}{=} X^{(i)}$ where $\widetilde{L} = \widetilde{H}^+$ there is a permutation $\sigma \in S(G)$ such that*

$$\widetilde{H} = P_\sigma^\top H, \quad \widetilde{B}^{(i)} = P_\sigma^\top B^{(i)} P_\sigma. \tag{79}$$

Remark 3. *A few remarks are in order.*

1. *Compared to their statement in Theorem 2 we replaced P_σ by $P_\sigma^\top$ because we used the transposed convention for the permutation matrices.*
2. *Note that their result per se doesn't mention identifiability up to scaling but that's because part of their assumptions involve normalizing f. Since we don't normalize the linear map f, we add scaling in the theorem statement.*
3. *For the case of imperfect interventions, their result talks about identifiability of the transitive closure $\overline{G}$ of the graph G, but this is the same as identifiability of the partial order π_G that we state here.*

We are now ready to prove our main theorems.

Theorem 1. *Suppose we are given distributions $X^{(i)}$ generated using a model $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ such that Assumptions 1-4 hold and such that all interventions i are perfect. Then the model is identifiable up to permutation and scaling, i.e., for any model $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in\bar{I}}, \widetilde{f})$ that generates the same data, there is a permutation $\omega \in S_d$ (and associated permutation matrix P_ω) and an invertible pointwise scaling matrix $\Lambda \in \mathrm{Diag}(d)$ such that*

$$\widetilde{t}_i = \omega(t_i), \quad \widetilde{B}^{(i)} = P_\omega^\top B^{(i)} \Lambda^{-1} P_\omega, \quad \widetilde{f} = f \circ \Lambda^{-1} P_\omega, \quad \widetilde{\eta}^{(i)} = \eta^{(i)}. \tag{3}$$

This in particular implies that

$$\widetilde{Z}^{(i)} \overset{\mathcal{D}}{=} P_\omega^\top \Lambda Z^{(i)} \tag{4}$$

and we can identify the causal graph G up to permutation of the labels.

Proof of Theorem 1. Suppose there are two representations $((B^{(i)}, \eta^{(i)}, t_i)_{i\in I}, f)$ and $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in I}, \widetilde{f})$ of the distributions $X^{(i)}, i \geq 0$. As explained in the proof of Theorem 3 we can assume by relabeling the nodes that the matrices $\widetilde{B}^{(i)}$ are lower triangular and the corresponding DAG $\widetilde{G}$ has the property that for every edge $i \to j$ we have $i < j$. We now apply Theorem 3 and find that there is an invertible linear map T such that $\widetilde{f} = f \circ T$ and $T^{-1} Z^{(i)} \overset{\mathcal{D}}{=} \widetilde{Z}^{(i)}$. Then we can find a unique invertible $\Lambda \in \mathrm{Diag}(d)$ such that ΛT has largest absolute value entry 1 in every row (and in case of ties the leftmost entry is 1). We can also find a permutation $\rho \in S_d$ such that $\bar{B}^{(i)} = P_\rho B^{(i)} \Lambda^{-1} P_\rho^\top$ is lower triangular for all i (note the distinction between $\bar{B}^{(i)}$ and $\widetilde{B}^{(i)}$). Indeed, this is possible since the underlying DAG G is acyclic by assumption, interventions do not add edges and Λ is diagonal. Then we find

$$(T^{-1}\Lambda^{-1}P_\rho^\top)(\bar{B}^{(i)})^{-1}\epsilon \overset{\mathcal{D}}{=} (T^{-1}\Lambda^{-1}P_\rho^\top)(\bar{B}^{(i)})^{-1}P_\rho\epsilon \tag{80}$$

$$= T^{-1}(B^{(i)})^{-1}\epsilon \tag{81}$$

$$\overset{\mathcal{D}}{=} (\widetilde{B}^{(i)})^{-1}\epsilon. \tag{82}$$

Here the last step follows from the fact that $T^{-1}Z^{(i)} \overset{\mathcal{D}}{=} \widetilde{Z}^{(i)}$ which implies that the centered distributions agree.

Now we can apply Theorem 5 (with B replaced by $\widetilde{B}$ and identity mixing) and find that

$$P_\rho \Lambda T = P_\sigma^\top \mathrm{Id}, \quad \bar{B}^{(i)} = P_\sigma^\top \widetilde{B}^{(i)} P_\sigma \tag{83}$$

for some $\rho \in S(\widetilde{G})$ because by assumption $\bar{B}^{(i)}$ are lower triangular. This implies that

$$T = \Lambda^{-1} P_\rho^\top P_\sigma^\top, \tag{84}$$

$$B^{(i)} = P_\rho^\top \bar{B}^{(i)} P_\rho \Lambda = P_\rho^\top P_\sigma^\top \widetilde{B}^{(i)} P_\sigma P_\rho \Lambda. \tag{85}$$

To summarize, there is a permutation $\omega = (\rho \circ \sigma)^{-1}$ (note that $P_\sigma P_\rho = P_{\rho\circ\sigma}$) and a diagonal matrix Λ such that

$$B^{(i)} = P_\omega \widetilde{B}^{(i)} P_\omega^\top \Lambda, \quad T = \Lambda^{-1} P_\omega, \quad Z^{(i)} = T\widetilde{Z}^{(i)} = \Lambda^{-1} P_\omega \widetilde{Z}^{(i)}, \tag{86}$$

We also find the relation $\omega(t_i) = \widetilde{t}_i$ from here because $P_\omega e_i = e_{\omega^{-1}(i)}$. Equating the means of $T^{-1}Z^{(i)}$ and $\widetilde{Z}^{(i)}$ we find

$$\widetilde{\eta}^{(i)} (\widetilde{B}^{(i)})^{-1} e_{\widetilde{t}_i} = \eta^{(i)} T^{-1} (B^{(i)})^{-1} e_{t_i}. \tag{87}$$

Multiplying by $\widetilde{B}^{(i)}$ we find

$$\widetilde{\eta}^{(i)} e_{\widetilde{t}_i} = \eta^{(i)} \widetilde{B}^{(i)} T^{-1} (B^{(i)})^{-1} e_{t_i} = P_\omega^\top e_{t_i} \tag{88}$$

and we conclude that $\widetilde{\eta}^{(i)} = \eta^{(i)}$.

□

Theorem 2. *Suppose we are given the distributions $X^{(i)}$ generated using a model $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ such that the Assumptions 1-4 hold and none of the interventions is a pure shift intervention. Then π_G can be identified up to a permutation of the labels.*

Proof of Theorem 2. Suppose there are two representations $((B^{(i)}, \eta^{(i)}, t_i)_{i\in I}, f)$ and $((\widetilde{B}^{(i)}, \widetilde{\eta}^{(i)}, \widetilde{t}_i)_{i\in I}, \widetilde{f})$ of the observational distribution. As shown in Theorem 3 there is a linear map such that $T^{-1}Z^{(i)} = \widetilde{Z}^{(i)}$. Then the same is true for the centered variables. Viewing $\widetilde{Z}^{(i)}$ as new observations obtained through a linear T mixing we conclude from Theorem 5 that π_G is identifiable. □

C A comparison to related works

In this section, we will discuss the relation of our results to the most relevant prior works and put them into context.

[79] This work considers linear mixing functions f, and we directly generalize their work to non-linear f. In particular, their techniques are linear-algebraic and do not generalize to non-linear f (see more technical details in proof intuition in Section 4); we handle this using a mix of statistical and geometric techniques. Notably, their finding that the linearly mixed latent variables are recoverable after observing one intervention per latent node is closely related to an earlier result in *domain generalization*, which showed that a number of environments equal to the number of "non-causal" latent dimensions can ensure recovery of the remaining latents [66], and later work showing that this requirement can be reduced further under linear mixing with additional modeling structure [10, 67]. This connection is not coincidental—that analysis was for invariant prediction with latent variables, which was originally presented as a method of causal inference using distributions resulting from unique interventions [60]. Later this approach was applied for the purpose of robust prediction in nonlinear settings [24], but the identifiability of the latent causal structure under general nonlinear mixing via interventional distributions remained an open question until this work.

[83] This work also considers linear mixing f, but allow for non-linearity in the latent variables. They use score functions to learn the model and they raise as an open question the setting of non-linear mixing, which we study in this work. While the models are not directly comparable, our model is more akin to real-world settings since it's not likely that high-level latent variables are linearly related to the observational distribution, e.g. pixels in an image are not linearly related to high-level concepts. Moreover, Gaussian priors and deep neural networks are extensively used in practice, and our theory as well as experimental methodology apply to them.

[46] While they do not discuss interventions, this work considers the setting of multi-environment CRL with Gaussian priors and their results can be applied to interventional learning. When applied to interventional data as in our setting (as also shown in [79, Appendix D]), they require additional restrictive assumptions such as $d = d'$, a bijective mixing f and $e \leq d$ where e is the number of edges in the underlying causal graph whereas we make no such restrictions on d, f and e (therefore, the maximum value of e can be as large as $\approx d^2/2$ in our setting). And when these assumptions are satisfied, they require $2d$ interventions whereas we show that d interventions are sufficient as well as necessary.

[2] This work also considers the setup of interventional causal representation learning. However, their setting is different in various ways and we now outline the key differences.

1. Their main results assume that the mixing f is an *injective polynomial* of finite degree p (see Assumption 2 in their paper). The injectivity requires their coefficient matrix to be full rank, which in turn implies $d' \geq \sum_{r=0}^{p} \binom{r+d-1}{d-1} = \binom{p+d}{d}$ (see the discussion below Assumption 2 in their work). This means that we cannot set the degree of the decoder polynomial to be arbitrarily large for fixed output dimension d' (which is required for universal approximation). In contrast, we only require $d' \geq d$. For a general discussion of the relation between approximability and identifiability we refer to Appendix E.

2. All their results for general nonlinear mixing functions rely on deterministic do-interventions which is a type of hard intervention that assigns a constant value to the target. However, we focus on randomized soft interventions (and also allow shift interventions which have found applications [68, 56]), which as also pointed out by [79, 82] is less restricted. Moreover, they require multiple interventions for each latent variable (as they use an ϵ-net argument) to show approximate identifiability, while the structural assumptions on the causal model allow us to show full identifiability with just one intervention for each latent variable. Thus, their setting is not directly comparable to ours and neither is strictly more general than the other. In particular, their proof techniques and experimental methodology don't translate to our setting, and we use completely different ideas for our proofs and experiments.

D Counterexamples and Discussions

In this section, we first present several counterexamples to relaxed versions of our main results in Section D.1 and then provide the missing proofs in Section D.2.

D.1 Main counterexamples

In this section, we will discuss the optimality and limitations of various assumptions of our main results. In particular, we consider the number of interventions, the distribution of the latent variables, and the types of interventions separately.

Number of interventions Our main results all require d interventions, i.e., one intervention per node. This is necessary even in the simpler setting of linear f as was shown in [79, Proposition 4], directly implying it is also necessary for the more general class of non-linear f that we handle here. Therefore, the number of interventions in our main theorems 1, 2 is both necessary and sufficient. Going a step further, it is natural to ask whether Theorem 3 remains true for less than d interventions. However, as we show next, $d-2$ interventions are not sufficient.

Fact 1. *Suppose we are given distributions $X^{(i)}$ generated by $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ satisfying Assumption 1-3. If the number of interventions satisfies $|I| \leq d-2$ it is not possible to identify f up to linear maps. Consider, e.g., $d=2$ and $Z = N(0, \mathrm{Id})$ and no interventions. Then $h(Z) \overset{\mathcal{D}}{=} Z$ for (nonlinear) radius dependent rotations as defined in [28, 8].*

Let us next consider the case with a total of d environments which corresponds to $d-1$ interventional distributions plus one observational distribution. We can provide a non-identifiability result for a general class of heterogeneous latent variable models that satisfy an algebraic property.

Lemma 7. *Assume that $d=2$ and $Z^{(0)} \sim N(0, \Sigma^{(0)})$, $Z^{(1)} \sim N(0, \Sigma^{(1)})$ and $\Sigma^{(1)} \succ \Sigma^{(0)}$ or $\Sigma^{(1)} \prec \Sigma^{(0)}$ in Löewner order (i.e., the difference is positive definite). Then there is a non-linear map h such that $h(Z^{(i)}) \overset{\mathcal{D}}{=} Z^{(i)}$ for $i = 0, 1$.*

We provide a proof of this lemma in Appendix D.2. Note that this non-identifiability lemma requires the assumption on $\Sigma^{(i)}$. In Lemma 3 in Appendix A we have seen that for the interventions considered in this work the relation $\Sigma^{(i)} > \Sigma^{(0)}$ or $\Sigma^{(i)} < \Sigma^{(0)}$ generally does not hold. This is some weak indication that for Theorem 3, $d-1$ interventions might be sufficient. We leave this for future work.

As an additional note, in the special case when f is a permutation matrix, we are in the setting of traditional causal discovery and here, $d-1$ interventions are necessary and sufficient [9, 17, 79].

Type of intervention Even when we restrict ourselves to only linear mixing functions as considered as in [79], we need perfect interventions in Theorem 1. Otherwise, only the weaker identifiability guarantees in Theorem 2 can be obtained. This is shown in [79]for linear f. Therefore, in the more general setting of nonlinear f, we cannot hope to obtain Theorem 1 with imperfect interventions, showing that this assumption is needed.

Next, we clarify that for the identifiability result in Theorem 2 , the condition that the interventions are not pure shift interventions is also necessary.

Fact 2. *Suppose $X^{(i)}$ is generated by $((B^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, f)$ satisfying Assumptions 1-4 where all interventions are pure shift interventions. By definition, this implies $B^{(i)} = B$ for all i. Consider any matrix $\widetilde{B}$ such that the corresponding DAG is acyclic. Then $((\widetilde{B}^{(i)}, \eta^{(i)}, t_i)_{i\in\bar{I}}, \widetilde{f})$ with $\widetilde{f} = f \circ (B^{-1}\widetilde{B})$ and $\widetilde{B}^{(i)} = \widetilde{B}$ for all i generates the same distributions $X^{(i)}$. Indeed,*

$$\widetilde{f}(\widetilde{Z}^{(i)}) = f \circ (B^{-1}\widetilde{B})(\widetilde{B}^{-1}(\epsilon + \eta^{(i)} e_{t_i}) = f(B^{-1}(\epsilon + \eta^{(i)} e_{t_i})) = f(Z^{(i)}). \tag{89}$$

This implies that any underlying causal graph could generate the observations.

Finally, we remark that in contrast to the case with linear mixing functions we need to exclude non-stochastic hard interventions for linear identifiability.

Lemma 8. *Consider $d=2$ and $Z^{(0)} = N(0, \mathrm{Id})$ and the non-stochastic hard interventions $\mathrm{do}(Z_1 = 0)$ and $\mathrm{do}(Z_2 = 0)$ resulting in the degenerate Gaussian distributions $Z^{(1)} = N(0, e_2 \otimes e_2)$ and $Z^{(2)} = N(0, e_1 \otimes e_1)$. Then there is a nonlinear h such that $h(Z^{(i)}) \overset{\mathcal{D}}{=} Z^{(i)}$ for all i.*

Intuitively, to show this, we choose h to be identity close to the supports $\{Z_1 = 0\}$ and $\{Z_2 = 0\}$ but some nonlinear measure preserving deformation in the four quadrants. The full argument can be found in Appendix D.2. A similar counterexample was discussed in [2]. However, here we in addition constrain the distribution of the latent variables, which adds a bit of complexity.

Distribution of the Noise Variables Our key additional assumption compared to the setting considered in [79, 2, 83] is that we assume that the latent variables are Gaussian which allows us to put only very mild assumptions on the mixing function f. However, we do this in a manner so that we still preserval universal representability guarantees.

[83] consider arbitrary noise distributions but restrict to linear mixing functions. We believe that our result can be generalized to more general distributions of the noise variables, e.g., exponential families, but those generalizations will require different proof strategies, which we leave for future work. However, the result will not be true for arbitrary noise distributions, as the following lemma shows.

Lemma 9. *Consider* $Z^{(0)} = (\epsilon_1, \epsilon_2)^\top$, $Z^{(1)} = (\epsilon_1', \epsilon_2)$, *and* $Z^{(2)} = (\epsilon_1, \epsilon_2')$ *where* $\epsilon_1, \epsilon_2 \sim \mathcal{U}([-1, 1])$ *and* $\epsilon_1', \epsilon_2' \sim \mathcal{U}([-3, 3])$. *Also define* $\widetilde{Z}^{(0)} = (\epsilon_1, \epsilon_1 + \epsilon_2)$, $\widetilde{Z}^{(1)} = (\epsilon_1', \epsilon_1' + \epsilon_2)$, *and* $\widetilde{Z}^{(2)} = (\epsilon_1, \epsilon_2')$. *Then there is* h *such that*

$$h(Z^{(i)}) \stackrel{\mathcal{D}}{=} \widetilde{Z}^{(i)}. \tag{90}$$

A proof of this lemma can be found in Appendix D.2. The result shows that even with d perfect interventions and linear SCMs it is not possible to identify the underlying DAG (empty graph for Z and $\widetilde{Z}_1 \to \widetilde{Z}_2$) and we also cannot identify f up to linear maps. While our example relies on distributions with bounded support, we conjecture that full support is not sufficient for identifiability.

Structural Assumptions A final assumption used in our results is the restriction to linear SCMs. This is used for most works on causal representation learning (see Section 2), even when they restrict to linear mixing functions. Although this may potentially be a nontrivial restriction, the advantage is that the simplicity of the latent space will enable us to meaningfully probe the latent variables, intervene and reason about them. Nevertheless, it's an interesting direction to study to what generality this can be relaxed. Indeed, even for the arguably simpler problem of causal structure learning (i.e., when $Z^{(i)}$ is observed) there are many open questions when moving beyond additive noise models. We leave this for future work.

D.2 Technical constructions

In this section, we provide the proofs for the counterexamples in Section D.1. For the first two counterexamples we construct functions h such that $h(Z^{(i)}) \stackrel{\mathcal{D}}{=} Z^{(i)}$ by setting $h = \Phi_t$ where Φ_t is defined as the flow of a suitable vectorfield X, i.e.,

$$\partial_t \Phi_t(z) = X(\Phi_t(z)). \tag{91}$$

This is a common technique in differential geometry and was introduced to construct counterexamples to identifiability ICA with volume preserving functions in [8]. To find a suitable vectorfield X we rely on the fact that the density p_t of $\mathbb{P}_t = (\Phi_t)_* \mathbb{P}$ satisfies the continuity equation

$$\partial_t p_t(z) + \mathrm{Div}(p_t(z) X(z)) = 0. \tag{92}$$

In particular $\mathbb{P}_t = \mathbb{P}$ holds iff

$$\mathrm{Div}(p_t(z) X(z)) = 0. \tag{93}$$

Therefore it is sufficient to find a vectorfield X such that $\mathrm{Div}(p^{(i)} X) = 0$ for all environments i to construct a suitable $h = \Phi_1$.

Lemma 7. *Assume that* $d = 2$ *and* $Z^{(0)} \sim N(0, \Sigma^{(0)})$, $Z^{(1)} \sim N(0, \Sigma^{(1)})$ *and* $\Sigma^{(1)} \succ \Sigma^{(0)}$ *or* $\Sigma^{(1)} \prec \Sigma^{(0)}$ *in Löewner order (i.e., the difference is positive definite). Then there is a non-linear map* h *such that* $h(Z^{(i)}) \stackrel{\mathcal{D}}{=} Z^{(i)}$ *for* $i = 0, 1$.

Proof of Lemma 7. Let us first discuss the high-level idea of the proof. The general strategy is to construct a vectorfield X whose flow preserves $Z^{(i)}$ for $i = 0, 1$. This holds if and only if the densities p_0, p_1 of $Z^{(0)}$, $Z^{(1)}$ satisfy

$$\operatorname{Div}(p_0 X) = \operatorname{Div}(p_1 X) = 0 \tag{94}$$

Using $\operatorname{Div}(fX) = \nabla f \cdot X + f \operatorname{Div} X$ we find

$$X \cdot \nabla(\ln p_0(z) - \ln p_1(z)) = \frac{X \cdot \nabla p_0(z)}{p_0(z)} - \frac{X \cdot \nabla p_1(z)}{p_1(z)} - \frac{p_0(z) \operatorname{Div} X}{p_0(z)} + \frac{p_1(z) \operatorname{Div} X}{p_1(z)} = 0. \tag{95}$$

We conclude that any X satisfying (94) must be orthogonal to $\nabla(\ln p_0(z) - \ln p_1(z))$ and thus parallel to the level lines of $\ln p_0(z) - \ln p_1(z)$. This already fixes the direction of X and the magnitude can be inferred from (94). To simplify the calculations we can first linearly transform the data so that the directions of X, i.e., equivalently the level lines of $\ln p_0(z) - \ln p_1(z)$ have a simple form. We assume that $\Sigma_0 \prec \Sigma_1$. Clearly, it is sufficient to show the result for any linear transformation of the Gaussian distributions $Z^{(i)}$. Note that generally for $G \sim N(\mu, \Sigma)$ we have $AG \sim N(A\mu, A\Sigma A^\top)$ Applying $\Sigma_0^{-\frac{1}{2}}$ we can reduce the problem to $\mathrm{Id} \prec \Sigma_1' = \Sigma_0^{-\frac{1}{2}} \Sigma_1 \Sigma_0^{-\frac{1}{2}}$. Diagonalizing Σ_1' by $U \in \mathrm{SO}(d)$ it is sufficient to consider $\mathrm{Id} \prec \Lambda = U \Sigma_1' U^\top$ where $\Lambda = \operatorname{Diag}(\lambda_1, \lambda_2)$. Finally, we can rescale z_1 and z_2 such that the resulting covariance matrices $\Lambda^0 = \operatorname{Diag}(\lambda_1^0, \lambda_2^0)$, $\Lambda^1 = \operatorname{Diag}(\lambda_1^1, \lambda_2^1)$ satisfy

$$\frac{1}{\lambda_1^1} - \frac{1}{\lambda_1^0} = \frac{1}{\lambda_2^1} - \frac{1}{\lambda_2^0} = 1. \tag{96}$$

Thus, it is sufficient to show the claim for such covariances Λ^0, Λ^1 and the result for arbitrary covariances follows by applying suitable linear transformation.

For such covariances we find for some constant c

$$\ln(p_0(z)) - \ln(p_1(z)) = -\frac{z_1^2}{2\lambda_1^0} - \frac{z_2^2}{2\lambda_2^0} + \frac{z_1^2}{2\lambda_1^1} + \frac{z_2^2}{2\lambda_2^1} + c = z_1^2 + z_2^2 + c. \tag{97}$$

The level lines are circles. Thus, we consider the vector field

$$X_0(z) = \begin{pmatrix} -z_2 \\ z_1 \end{pmatrix}. \tag{98}$$

We see directly that $\operatorname{Div} X_0 = 0$. We consider the Ansatz $X(z) = f(|z|) X_0(z)/p_0(z)$. We now fix the radial function $f : \mathbb{R}_+ \to \mathbb{R}$ such that it has compact support away from 0. We find using $\operatorname{Div} X_0 = 0$

$$\begin{aligned} \operatorname{Div}(X(z)p_0(z)) &= \operatorname{Div}(X(z)p_0(z)) && (99)\\ &= X_0(z) \cdot (\nabla |z|) f'(z) && (100)\\ &= \begin{pmatrix} -z_2 \\ z_1 \end{pmatrix} \cdot \frac{z}{|z|^2} f'(z) && (101)\\ &= 0. && (102) \end{aligned}$$

Note, that close to 0 where $|z|$ is not differentiable the vector field vanishes by our construction of f. We find similarly using (97)

$$\begin{aligned} \operatorname{Div}(X(z)p_1(z)) &= X_0(z) \cdot \nabla \left(f(|z|) \frac{p_1(z)}{p_0(z)} \right) && (103)\\ &= X_0(z) \cdot \nabla \left(f(|z|) e^{-|z|^2 - c} \right) && (104)\\ &= X_0(z) \cdot \nabla \tilde{f}(|z|) && (105)\\ &= 0. && (106) \end{aligned}$$

Finally we remark that the flow Φ_t of the vectorfield X generates a family of functions h as desired, i.e., for all t we have $(\Phi_t)_* Z^{(i)} = Z^{(i)}$ and Φ_t is clearly nonlinear as it is the identity close to 0 but not globally. □

The proof of the next lemma is similar but simpler.

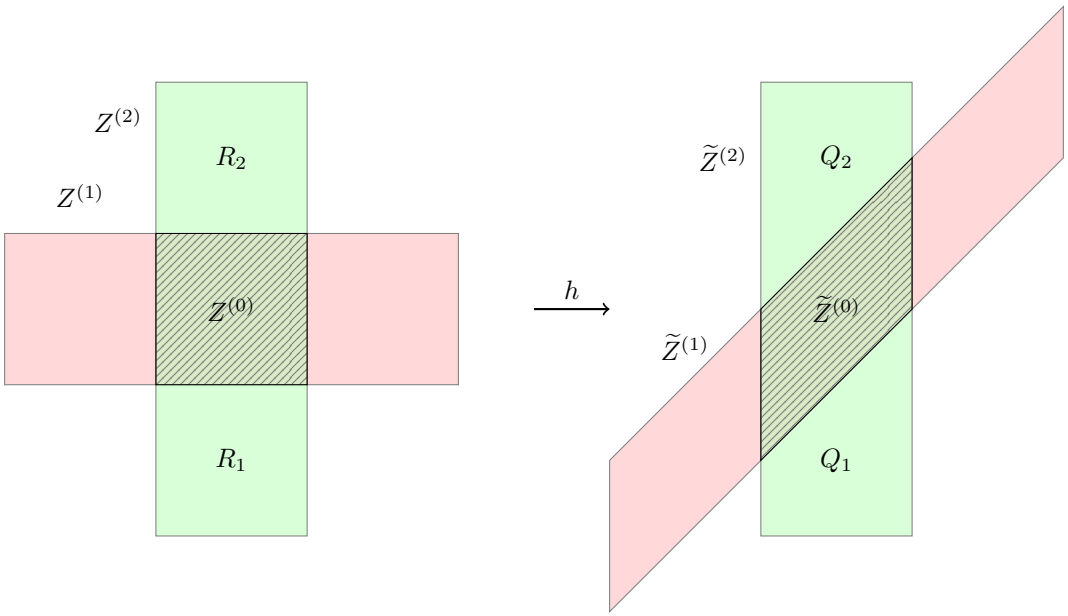


Figure 4: Sketch of the setting in Lemma 9. The map h agrees with $(z_1, z_2) \to (z_1, z_1 + z_2)$ on the red rectangle and maps R_i to Q_i such that the uniform measure is preserved.

Lemma 8. *Consider $d = 2$ and $Z^{(0)} = N(0, \mathrm{Id})$ and the non-stochastic hard interventions $\mathrm{do}(Z_1 = 0)$ and $\mathrm{do}(Z_2 = 0)$ resulting in the degenerate Gaussian distributions $Z^{(1)} = N(0, e_2 \otimes e_2)$ and $Z^{(2)} = N(0, e_1 \otimes e_1)$. Then there is a nonlinear h such that $h(Z^{(i)}) \overset{\mathcal{D}}{=} Z^{(i)}$ for all i.*

Proof of Lemma 8. Pick any smooth divergence free vectorfield X_0 with compact support in $[\epsilon, \infty)^2$ for some $\epsilon > 0$. Then the vector field $X = X_0/p_0$ satisfies

$$\mathrm{Div}\, X p_0 = \mathrm{Div}\, X_0 = 0. \tag{107}$$

Thus the flow Φ_t preserves the distribution of $Z^{(0)}$. But it also preserved the interventional distributions $Z^{(i)}$ for $i = 1, 2$ because $X(z) = 0$ and thus $\Phi_t(z) = z$ for z close to the supports of $Z^{(i)}$. □

Finally we prove Lemma 9 that shows that we cannot completely drop the assumption on the distribution of the noise variables ϵ.

Lemma 9. *Consider $Z^{(0)} = (\epsilon_1, \epsilon_2)^\top$, $Z^{(1)} = (\epsilon_1', \epsilon_2)$, and $Z^{(2)} = (\epsilon_1, \epsilon_2')$ where $\epsilon_1, \epsilon_2 \sim \mathcal{U}([-1, 1])$ and $\epsilon_1', \epsilon_2' \sim \mathcal{U}([-3, 3])$. Also define $\widetilde{Z}^{(0)} = (\epsilon_1, \epsilon_1 + \epsilon_2)$, $\widetilde{Z}^{(1)} = (\epsilon_1', \epsilon_1' + \epsilon_2)$, and $\widetilde{Z}^{(2)} = (\epsilon_1, \epsilon_2')$. Then there is h such that*

$$h(Z^{(i)}) \overset{\mathcal{D}}{=} \widetilde{Z}^{(i)}. \tag{90}$$

Proof of Lemma 9. A sketch of the setting can be found in Figure 4. We define $h_0(z) = (z_1, z_1 + z_2)$. If $h(z) = h_0(z)$ for $-1 \le z_2 \le 1$ we find $h(Z^{(i)}) \overset{\mathcal{D}}{=} \widetilde{Z}^{(i)}$ for $i = 0$ and $i = 1$. It remains to modify h_0 such that h preserves the uniform measure on $[-1, 1] \times [-3, 3]$. We do this in two steps, first we construct a modification $\widetilde{h}$ of the map h_0 such that $[-1, 1] \times [-3, 3]$ is mapped bijectively to itself but $\widetilde{h} = h_0$ for $|z_2| \le 1$ (so that $\widetilde{h}(Z^{(i)}) \overset{\mathcal{D}}{=} \widetilde{Z}^{(i)}$ for $i = 0$ and $i = 1$ holds) and then we apply a general result to make the map measure preserving. We start with the first step. Let $\psi : \mathbb{R} \to [0, 1]$ be differentiable such that $\psi(t) = 1$ for $-1 \le z \le 1$ and $\psi(t) = 0$ for $-5/2 \le z \le 5/2$ with $|\psi'(t)| < 1$. Then

$$\tilde{h}(z) = \psi(z_2) z + (1 - \psi(z_2)) h_0(z) = \begin{pmatrix} z_1 \\ z_2 + \psi(z_2) z_1 \end{pmatrix} \tag{108}$$

is injective on $[-1, 1] \times [-3, 3]$ (note that the second coordinate is increasing in z_2) and maps $[-1, 1] \times [-3, 3]$ bijectively to itself and agrees with h_0 for $|z_2| \le 1$. However, it is not necessarily volume preserving. But applying Moser's theorem (see [54]) to the image of the rectangles $R_1 = [-1, 1] \times [-3, -1]$ and $R_2 = [-1, 1] \times [1, 3]$ which are the quadrilaterals $Q_1 = \tilde{h}(R_1)$ with endpoints $(-1, -3)$, $(1, -3)$, $(1, 0)$, $(-1, -2)$ and $Q_2 = \tilde{h}(R_2)$ with endpoints $(1, 3)$, $(-1, 3)$, $(-1, 0)$, $(1, 2)$ with the same size as R_1 and R_2 we infer that there is φ supported in $Q_1 \cup Q_2$ such that $h = \varphi \circ \tilde{h}$ satisfies $h(Z^{(2)}) \overset{\mathcal{D}}{=} \widetilde{Z}^{(2)} \overset{\mathcal{D}}{=} \mathcal{U}([-1, 1] \times [-3, 3])$. □

E Identifiability and Approximation

In this section, we explain that approximation properties of function classes cannot be leveraged to obtain stronger identifiability results. The general setup is that we have two function classes $\mathcal{G} \subset \mathcal{F}$ and we assume that $\mathcal{G}$ is dense in $\mathcal{F}$ (in the topology of uniform convergence on Ω), i.e., for $f \in \mathcal{F}$ and any $\epsilon > 0$ there is $g \in \mathcal{G}$ such that

$$\sup_{z \in \Omega} |f(z) - g(z)| \leq \epsilon. \tag{109}$$

We now investigate the relationship of identifiability results for mixings in either $\mathcal{G}$ or in $\mathcal{F}$.

While this point is completely general, we will discuss this for concreteness in the context of nonlinear ICA and considering polynomial mixing functions for $\mathcal{G}$. We also assume that $\mathcal{F}$ are all diffeomorphisms (onto their image). Suppose we have latents $Z \sim \mathcal{U}([-1,1]^d)$ and observe a mixture $X = h(Z)$. We use the shorthand $\Omega = [-1,1]^d$ from now on. Let us suppose that $h \in \mathcal{G}$. Then the following result holds.

Lemma 10. *Suppose $\widetilde{h}(Z) \overset{\mathcal{D}}{=} X \overset{\mathcal{D}}{=} h(Z)$ where $Z \sim \mathcal{U}(\Omega)$, $h, \widetilde{h} \in \mathcal{G}$ and h satisfies an injectivity condition as defined in Assumption 2 in [2]. Then h and $\widetilde{h}$ agree up to permutation of the variables.*

Proof. This is a consequence of Theorem 4 in [2]. Essentially, they show in Theorem 1 that h and $\tilde{h}$ agree up to a linear map and then use independence of supports to conclude (in our simpler setting here, one could also resort to the identifiability of linear ICA). □

The injectivity condition is a technical condition that ensures that the embedding h is sufficiently diverse, but this is not essential for the discussion here. Let us nevertheless mention here that it is not clear whether this condition is sufficiently loose to ensure identifiability in a dense subspace of $\mathcal{F}$ because it requires an output dimension depending on the degree of the polynomials.

We now investigate the implications for $\mathcal{F}$. It is well known that ICA in general nonlinear function is not identifiable, i.e., for any $f \in \mathcal{F}$ there is a $\widetilde{f} \in \mathcal{F}$ such that $f(Z) \overset{\mathcal{D}}{=} \widetilde{f}(Z)$. To find such an $\widetilde{f}$ one use, e.g., the Darmois construction, radius dependent rotations or, more generally, measure preserving transformations m such that $m(Z) \overset{\mathcal{D}}{=} Z$. Then we have the following trivial lemma.

Lemma 11. *For every $\epsilon > 0$ there are functions $g, \widetilde{g} \in \mathcal{G}$ such that*

$$\sup_{\Omega} |g - f| < \epsilon, \quad \sup_{\Omega} |\widetilde{g} - \widetilde{f}| < \epsilon \tag{110}$$

and then the Wasserstein distance satisfies $W_2(f(Z), g(Z)) < \epsilon$, $W_2(\widetilde{f}(Z), \widetilde{g}(Z)) < \epsilon$.

Remark 4. *For the definition and a discussion of the Wasserstein metric we refer to [85]. Alternatively we could add a small amount of noise, i.e. $X = f(Z) + \nu$ and then consider the total variation distance.*

Proof. The first part follows directly from the Stone-Weierstrass Theorem which shows that polynomials are dense in the continuous function on compact domains. The second part follows from the coupling $Z \to (f(Z), g(Z))$ and (110). □

The main message of this Lemma is that whenever there are spurious solutions in the larger function class $\mathcal{F}$ we can equally well approximate the ground truth mixing f and the spurious solution $\widetilde{f}$ by functions in $\mathcal{G}$. In this sense, the identifiability of ICA in $\mathcal{G}$ does not provide any guidance to resolve the ambiguity of ICA in $\mathcal{F}$. So identifiability results in $\mathcal{G}$ are not sufficient when there is no reason to believe that the ground truth mixing is exactly in $\mathcal{G}$.

Actually, one could view the approximation capacity of $\mathcal{G}$ also as a slight sign of warning as we will explain now. Suppose the ground truth mixing satisfies $g \in \mathcal{G}$. Since ICA in $\mathcal{F}$ is not identifiable we can find $\widetilde{f} \in \mathcal{F}$ which can be arbitrarily different from g but such that $g(Z) = \widetilde{f}(Z)$. Then we can approximate $\widetilde{f}$ by $\widetilde{g}$ with arbitrarily small error. Thus, we find almost spurious solutions, i.e., $\widetilde{g} \in \mathcal{G}$ such that $W_2(\widetilde{g}(Z), g(Z)) < \epsilon$ for an arbitrary $\epsilon > 0$ but $\widetilde{g}$ and g correspond to very different data representations, in this sense the identifiability result is not robust.

When only considering the smaller space $\mathcal{G}$ this problem can be resolved by considering norms or seminorms on $\mathcal{G}$, for polynomials a natural choice is the degree of the polynomial. Indeed, suppose that we observe data generated using a mixing $g \in \mathcal{G}$. Then there might be spurious solutions $\widetilde{f} \in \mathcal{F}$ which can be approximated by $\widetilde{g} \in \mathcal{G}$ but this approximation will generically have a larger norm (or degree for polynomials) than g. Therefore, the minimum description length principle [64] favors g over $\widetilde{g}$.

This reasoning can only be extended to the larger class $\mathcal{F}$ if the ground truth mixing f (for the relevant applications) can be better approximated (i.e., with smaller norm) by functions in $\mathcal{G}$ than all spurious solutions $\widetilde{f}$. This is hard to verify in practice and difficult to formalize theoretically. Nevertheless, this motivates to look for identifiability results for function classes that are known to be useful representation learners and used in practice, in particular neural nets. We emphasize that alternatively one can also directly consider norms on the larger space $\mathcal{F}$ penalizing, e.g., derivative norms, to perform model selection.

F Additional details on experimental methodology

In this appendix, we give more details about our experimental methodology. First, we derive the log-odds and prove Lemma 1 in Appendix F.1. Then, we use it to design our model and theoretically justify our contrastive approach in Appendix F.2, by connecting it to the Bayes optimal classifier. Then, we describe the NOTEARS regularizer in Appendix F.3, describe the limitations of the contrastive approach in Appendix F.4 and finally, in Appendix F.5, we describe the ingredients needed for an approach via Variational Autoencoders, the difficulties involved and how to potentially bypass them.

F.1 Proof of Lemma 1

Lemma 1. *When we have perfect interventions, the log-odds of a sample x coming from $X^{(i)}$ over coming from $X^{(0)}$ is given by*

$$\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x) = c_i - \frac{1}{2}\lambda_i^2(((f^{-1}(x))_{t_i})^2 + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i}) + \frac{1}{2}\langle f^{-1}(x), s^{(i)}\rangle^2 \tag{6}$$

for a constant c_i independent of x.

Proof. We will write the log likelihood of $X^{(i)} = f(Z^{(i)})$ using standard change of variables. As we elaborate in Appendix A.1, we have

$$Z^{(i)} = (B^{(i)})^{-1}\epsilon + \mu^{(i)} \text{ with } B^{(i)} = (D^{(i)})^{-1/2}(I - A^{(i)}), \mu^{(i)} = \eta^{(i)}(B^{(i)})^{-1}e_{t_i} \tag{111}$$

Also, denote by $\Theta^{(i)} = (B^{(i)})^\top B^{(i)}$ the precision matrix of $Z^{(i)}$ (see Appendix A.1 for full derivation). By change of variables,

$$\begin{aligned}\ln p_X^{(i)}(x) &= \ln|J_{f^{-1}}| + \ln p_Z^{(i)}(f^{-1}(x)) \\ &= \ln|J_{f^{-1}}| - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma^{(i)}| - \frac{1}{2}(f^{-1}(x) - \mu^{(i)})^T\Theta^{(i)}(f^{-1}(x) - \mu^{(i)})\end{aligned}$$

where $J_{f^{-1}}$ denotes the Jacobian matrix of partial derivatives. Let $s^{(1)}, s^{(2)}, \ldots, s^{(d)}$ denote the rows of $B^{(0)}$. We then compute the log-odds with respect to $X^{(0)}$, the base distribution without interventions (and where $\mu^{(0)} = 0$) to get

$$\begin{aligned}&\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x) \\ &= \left(-\frac{1}{2}\ln\frac{|\Sigma^{(i)}|}{|\Sigma^{(0)}|} - \frac{1}{2}(\mu^{(i)})^\top\Theta^{(i)}\mu^{(i)}\right) - \frac{1}{2}(f^{-1}(x))^\top(\Theta^{(i)} - \Theta^{(0)})(f^{-1}(x)) + (f^{-1}(x))^\top\Theta^{(i)}\mu^{(i)} \\ &= c_i - \frac{1}{2}(f^{-1}(x))^\top(r^{(i)} \otimes r^{(i)} - s^{(i)} \otimes s^{(i)})(f^{-1}(x)) + \eta^{(i)} \cdot (f^{-1}(x))^\top(B^{(i)})^\top e_{t_i} \\ &= c_i - \frac{1}{2}(f^{-1}(x))^\top(\lambda_i^2 e_{t_i}e_{t_i}^\top - (s^{(i)})(s^{(i)})^\top)(f^{-1}(x)) + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))^\top e_{t_i} \\ &= c_i - \frac{1}{2}(f^{-1}(x))^\top(\lambda_i^2 e_{t_i}e_{t_i}^\top - (s^{(i)})(s^{(i)})^\top)(f^{-1}(x)) + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i} \\ &= c_i - \frac{1}{2}\lambda_i^2(((f^{-1}(x))_{t_i})^2 + \eta^{(i)}\lambda_i \cdot (f^{-1}(x))_{t_i}) + \frac{1}{2}\langle f^{-1}(x), s^{(i)}\rangle^2\end{aligned}$$

for a constant c_i independent of z. For the second equality, we used Lemma 3. □

F.2 A theoretical justification for the log-odds model

In this section, we provide more details on our precise modeling choices for the contrastive approach, and show theoretically why it encourages the model to learn the true underlying parameters.

Recall that, motivated by Lemma 1, we model the log-odds as

$$g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta) = \alpha_i - \beta_i h_{t_i}^2(x, \theta) + \gamma_i h_{t_i}(x, \theta) + \langle h(x, \theta), w^{(i)} \rangle^2 \tag{112}$$

where $h(\cdot, \theta)$ denotes a neural net parametrized by θ, parameters $w^{(i)}$ are the rows of a matrix W and α_i, β_i, γ_i are learnable parameters. Moreover, we have $g_0(x) = 0$ as we compute the log-odds with respect to $X^{(0)}$. Therefore, the cross entropy loss given by

$$\mathcal{L}_{\text{CE}}^{(i)} = -\mathbb{E}_{j \sim \mathcal{U}(\{0,i\})} \mathbb{E}_{x \sim X^{(j)}} \ln \left(\frac{e^{\mathbf{1}_{j=i} g_i(x)}}{e^{g_i(x)} + 1} \right). \tag{113}$$

Let $C(x) \in \{0, 1\}$ denote the label of the datapoint x indicating whether it was an observational datapoint or an interventional datapoint respectively. By using the cross entropy loss, we are treating the model outputs as logits, therefore we can write down the posterior probability distribution using a logistic regression model as

$$\Pr[C(x) = 1 | x] = \frac{\exp(g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta))}{1 + \exp(g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta))} \tag{114}$$

Moreover, by using a sufficiently wide or deep neural network with universal approximation capacity and sufficiently many samples, our model will learn the Bayes optimal classifier, as given by Lemma 1. Therefore, we can equate the probabilities to arrive at

$$\frac{\exp(g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta))}{1 + \exp(g_i(x, \alpha_i, \beta_i, \gamma_i, w^{(i)}, \theta))} = \frac{\exp(\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x))}{1 + \exp(\ln p_X^{(i)}(x) - \ln p_X^{(0)}(x))} \tag{115}$$

implying

$$\alpha_i - \beta_i h_{t_i}^2(x, \theta) + \gamma_i h_{t_i}(x, \theta) + \langle h(x, \theta), w^{(i)} \rangle^2 \tag{116}$$

$$= c_i - \frac{1}{2}\lambda_i^2 (((f^{-1}(x))_{t_i})^2 + \eta^{(i)} \lambda_i \cdot (f^{-1}(x))_{t_i}) + \frac{1}{2} \langle f^{-1}(x), s^{(i)} \rangle^2 \tag{117}$$

This suggests that in optimal settings, our model will learn (up to scaling)

$$\alpha_i = c_i, \quad \beta_i = \frac{1}{2}\lambda_i^2, \quad \gamma_i = \eta^{(i)} \lambda_i, \quad w^{(i)} = s^{(i)}, \quad h(x, \theta) = f^{-1}(x) \tag{118}$$

thereby inverting the nonlinearity and learning the underlying parameters.

F.3 NOTEARS [95]

In our algorithm, using the parameters $w^{(i)}$ as rows, we learn the matrix W, which we saw in optimal settings will be $B = D^{-1/2}(\text{Id} - A)$. Note that the off-diagonal entries of B form a DAG. Therefore, the graph encoded by W_0, which is defined to be W with the main diagonal zeroed out, must also be a DAG.

However, in our algorithm, we don't explicitly enforce this acyclicity. There are two ways to get around this. One way is to assume a default causal ordering on the Z_is, which is feasible as we don't directly observe Z. Then, we can simply enforce W_0 to be triangular and train via standard gradient descent and W_0 is guaranteed to be a DAG. However, this approach doesn't work because the interventional datasets are given in arbitrary order and so we are not able to match the datasets to the vertices in the correct order. So this merely defers the issue.

Instead, the other approach that we take in this work is to regularize the learnt W_0 to model a directed acyclic graph. Learning an underlying causal graph from data is a decades old problem that has been widely studied in the causal inference literature, see e.g. [11, 95, 62, 75, 4] and references therein. In particular, the work [95] proposed an analytic expression to measure the DAGness of the causal graph, thereby making the problem continuous and more efficient to optimize over.

Lemma 12 ([95]). *A matrix $W \in \mathbb{R}^{d\times d}$ is a DAG if and only if*

$$\mathcal{R} := \operatorname{tr} \exp(W \circ W) - d = 0 \tag{119}$$

where $\circ$ is the matrix Hadamard product and exp *is the matrix exponential. Moreover, $\mathcal{R}$ has gradient*

$$\nabla \mathcal{R} = \exp(W \circ W)^\top \circ 2W \tag{120}$$

Therefore, we add $\mathcal{R}_{NOTEARS}(W) = \operatorname{tr} \exp(W_0 \circ W_0) - d$ as a regularization to our loss function. As in prior works, we could also consider the augmented Lagrangian formulation however it leads to additional training complexity. There have been some follow-up works to NOTEARS such as [92] (see [4] for an overview). We leave exploring such alternatives to future work.

F.4 Limitations of the contrastive approach

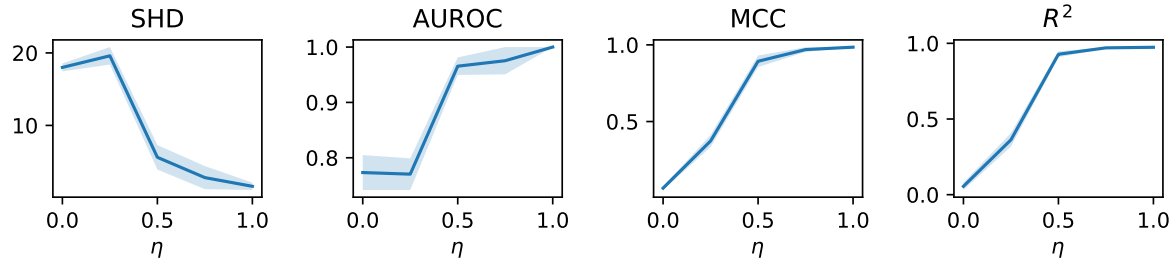


Figure 5: Dependence of the performance metrics on the shift η. Other settings are as for nonlinear mixing functions (see Table 5) with ER$(10, 2)$ graphs and $d' = 100$.

As mentioned in the main part of the paper, our contrastive algorithm struggles to recover the ground truth latent variables when considering interventions without shifts. We illustrate the dependence on the shift strength in Figure 5. In this section, we provide some evidence why this is the case. Apart from setting the stage for future works, this also enables practitioners to be aware of potential pitfalls when applying our techniques.

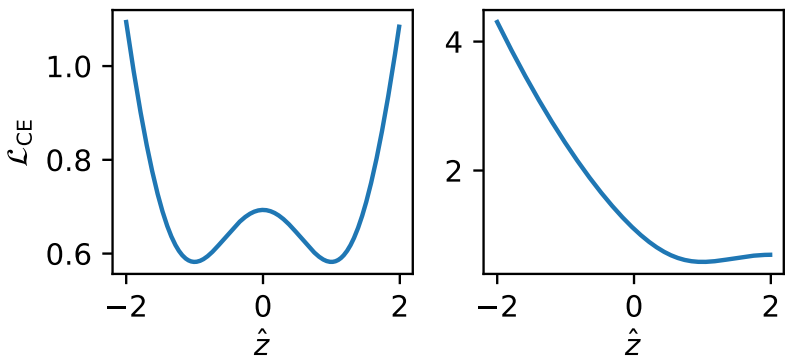


Figure 6: Cross entropy loss for $a = 1$, $c = 0$, $z_0 = 1$ and $b = 0$ (left), $b = 2$ (right) as a function of the estimated latent variable $\hat{z}$.

We suspect that the main reason for this behavior is the non-convexity of the parametric output layer. Note that this is very different from the well known non-convexity of (overparametrized) neural networks. While neural networks parametrize non-convex functions their final layer is typically a convex optimization problem, e.g., a least squares regression or a logistic regression on the features. Moreover, it is well understood that convergence to a global minimizer using gradient descent is possible under suitable conditions, see, e.g., [45].

This is different for our quadratic log-odds expression where gradient descent is not sufficient to find the global minimizer. To illustrate this we consider the case of a single latent, i.e., $d = 1$, then the ground truth parametric form of the log-odds in 6 can be expressed as

$$\ln p_X^{(1)} - \ln p_X^{(0)}(x) = g(x) = h(f^{-1}(x)) = a(f^{-1}(x) - b)^2 + c \tag{121}$$

for some constants a, b, c and $b = 0$ if $\eta^{(1)} = 0$, i.e., there is no shift. We moreover assume, for the sake of argument, that the final layer implementing $h(z) = a(z - b)^2 + c$ is fixed to the ground truth parametric form of the log-odds (then there is also no scaling ambiguity left). Now, let us fix a point x_0 and define $z_0 = f^{-1}(x_0)$ and

$$p = \mathbb{P}(i = 1|X = x_0) = \frac{p_X^{(1)}(x_0)}{p_X^{(1)}(x_0) + p_X^{(0)}(x_0)} = \frac{e^{g(x_0)}}{1 + e^{g(x_0)}}. \tag{122}$$

Then the (sample conditional) cross entropy loss as a function of $\hat{z} = \hat{f}^{-1}(x_0)$ for our learned function $\hat{f}$ is given by

$$\mathcal{L}_{\text{CE}} = -p \ln \left(\frac{e^{h(\hat{z})}}{1 + e^{h(\hat{z})}} \right) - (1 - p) \ln \left(\frac{1}{1 + e^{h(\hat{z})}} \right) = -ph(\hat{z}) + \ln(1 + e^{h(\hat{z})}) \tag{123}$$

We here drop the intervention index, since we focus on a one dimensional illustration. We plot two examples of such loss functions for $b = 0$ and $b \neq 0$ in Figure 6.

To verify that this is indeed the reason for our failure to recover the ground truth latent variables, we investigate the relation between estimated latent variables and ground truth latent variables. The result can be found in Figure 7 and, as we see, when $\eta = 0$, the distribution of the recovered latents is skewed. Moreover, we find that we essentially recover the latent variable up to a sign, i.e., $\hat{Z} = |Z|$ (there is a small offset because we enforce $\hat{Z}$ to be centered). Note that this explains the tiny MCC scores in Table 1, since $\mathbb{E}(Z|Z|) = 0$ for symmetric distributions. This observation might also be a potential starting point to improve our algorithm. For non-zero shifts we recover the latent variables almost perfectly.

F.5 A Variational Autoencoder approach

In this section, we describe the technical details involved with adapting Variational Autoencoders (VAEs) [37, 63] for our setting. We highlight some difficulties that we will encounter and also suggest possible ways to work around them.

Indeed, most earlier approaches for causal representation learning have relied on maximum likelihood estimation via variational inference. In particular, they have relied on autoencoders or variational autoencoders [34, 2]. In our setting, VAEs are a viable approach. More concretely, we can encode the distribution $X^{(0)}$ to ϵ, then apply the $B^{(i)}$ parameter matrix before finally decoding to $X^{(i)}$. To handle the fact that we don't have paired interventional data as in [6], we could potentially modify the Evidence Lower Bound (ELBO) to include some divergence measure between the predicted and measured interventional data. Finally, this can be trained end-to-end as in traditional VAEs. We now describe this in more detail.

We will use VAEs to encode the observational distribution $X^{(0)}$ to ϵ, so the encoder will formally model $B^{(0)} \circ (f^{-1}(.) - \mu^{(i)})$ where we use the notation from Appendix A.1. Note here that we cannot directly design the encoder to map X to Z because we do not know the prior distribution on Z. Indeed, the objective is to learn it.

Let I denote the set of intervention targets. Assume the targets are chosen uniformly at random, which can be done in practice by subsampling each interventional distribution to have the same size. Suppose we had paired counterfactual data $\mathcal{D} = \cup_{i \in I} \mathcal{D}_i$ of paired interventional datasets, i.e. $\mathcal{D}_i$ consists of pairs $(X^{(0)}, X^{(i)})$ with corresponding probability density denoted $p^{(i)}(x^{(0)}, x^{(i)})$. Following [37, 94], define an amortized inference distribution as $q(\epsilon|x^{(0)})$. Then, we can bound the expected log-likelihood as

$$\mathbb{E}_{\mathcal{D}}[\ln p^{(i)}(x^{(0)}, x^{(i)})] = \mathbb{E}_{i \sim \mathcal{U}(I)} \mathbb{E}_{\mathcal{D}_i}[\ln p^{(i)}(x^{(0)}, x^{(i)})] \tag{124}$$

$$= \mathbb{E}_{i \sim \mathcal{U}(I)} \mathbb{E}_{\mathcal{D}_i} \ln \int_{\epsilon} p^{(i)}(x^{(0)}, x^{(i)}, \epsilon) \, d\epsilon \tag{125}$$

$$= \mathbb{E}_{i \sim \mathcal{U}(I)} \mathbb{E}_{\mathcal{D}_i} \ln \int_{\epsilon} \frac{p^{(i)}(x^{(0)}, x^{(i)}, \epsilon) q(\epsilon|x^{(0)})}{q(\epsilon|x^{(0)})} \, d\epsilon \tag{126}$$

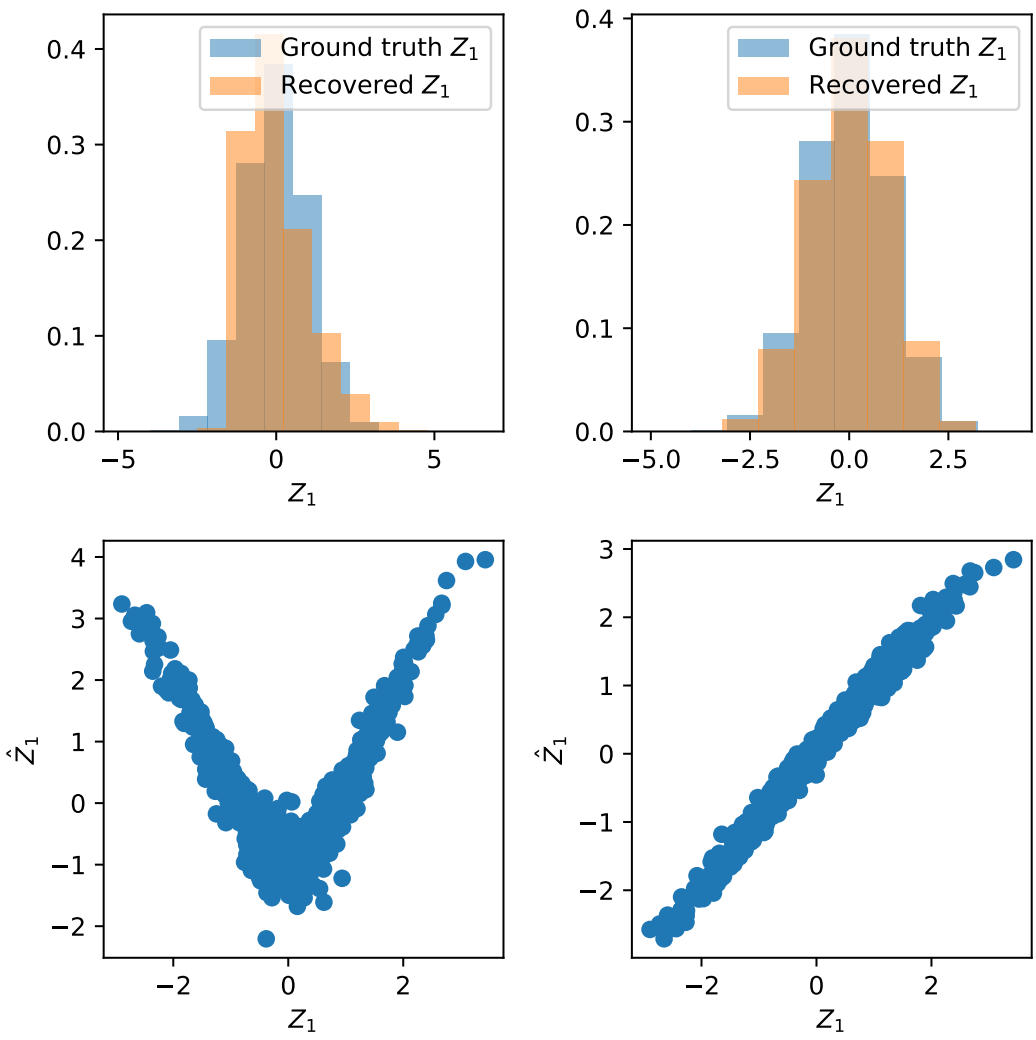


Figure 7: Density of standardized ground truth latents and recovered latents for $\eta = 0$ (top left) and $\eta = 1$ (top right), scatter plot of the ground truth latent variables Z_1 and recovered latent variables $\hat{Z}_1$ for $\eta = 0$ (bottom left) and $\eta = 1$ (bottom right). Results shown for ER$(10, 2)$ graphs and $d' = 100$, all further parameters as in Table 5.

$$= \mathbb{E}_{i\sim\mathcal{U}(I)}\mathbb{E}_{\mathcal{D}_i} \ln \mathbb{E}_{\epsilon\sim q(\epsilon|x^{(0)})} \frac{p^{(i)}(x^{(0)}, x^{(i)}, \epsilon)}{q(\epsilon|x^{(0)})} \tag{127}$$

$$\geq \mathbb{E}_{i\sim\mathcal{U}(I)}\mathbb{E}_{\mathcal{D}_i}\mathbb{E}_{\epsilon\sim q(x^{(0)})} \ln \frac{p^{(i)}(x^{(0)}, x^{(i)}, \epsilon)}{q(\epsilon|x^{(0)})} \qquad \text{(Jensen's inequality)} \tag{128}$$

$$= \mathbb{E}_{i\sim\mathcal{U}(I)}\mathbb{E}_{\mathcal{D}_i}\mathbb{E}_{\epsilon\sim q(x^{(0)})} \ln \frac{p^{(i)}(x^{(0)}, x^{(i)}|\epsilon)p(\epsilon)}{q(\epsilon|x^{(0)})} \tag{129}$$

$$= \mathbb{E}_{i\sim\mathcal{U}(I)}\mathbb{E}_{\mathcal{D}_i}\left(\mathbb{E}_{\epsilon\sim q(x^{(0)})} \ln p^{(i)}(x^{(0)}, x^{(i)}|\epsilon) - \mathbb{E}_{\epsilon\sim q(x^{(0)})} \ln \frac{q(\epsilon|x^{(0)})}{p(\epsilon)}\right) \tag{130}$$

$$= \mathbb{E}_{i\sim\mathcal{U}(I)}\mathbb{E}_{(x^{(0)},x^{(i)})\sim\mathcal{D}_i}\Big[\mathbb{E}_{\epsilon\sim q(\epsilon|x^{(0)})}[\ln p^{(i)}(x^{(0)}, x^{(i)}|\epsilon)] \tag{131}$$

$$- \mathrm{KL}(q(\epsilon|x^{(0)})||N(0, I))\Big] \tag{132}$$

where the last term is the standard Evidence Lower Bound (ELBO). This can then be trained end-to-end via the reparameterization trick. A similar approach was taken in [6] who had access to paired counterfactual data.

However, we do not have access to such paired interventional data, but instead we only observe the marginal distributions $X^{(i)}$. To this end, conditioned on ϵ, we can split the term $\ln p^{(i)}(x^{(0)}, x^{(i)}|\epsilon) = \ln \widetilde{p}^{(0)}(x^{(0)}|\epsilon) + \ln \widetilde{p}^{(i)}(x^{(i)}|\epsilon)$ where $\widetilde{p}^{(i)}$ denotes the density of the intervened marginal of $p^{(i)}$, which corresponds to using the interventional decoder $f \circ ((B^{(i)})^{-1}(.) + \mu^{(i)})$. Nevertheless, $\mathbb{E}_{\epsilon \sim q(x^{(0)})}[\ln \widetilde{p}^{(i)}(x^{(i)}|\epsilon)]$ is still not tractable in unpaired settings. As a heuristic, we could consider a modified ELBO by modifying this term to measure some sort of divergence metric between the true interventional data $X^{(i)}$ and the generated interventional data $\widetilde{p}^{(i)}(x^{(i)}|\epsilon)$, similar to [94]. We leave it for future work to explore this direction.

G Additional details on experiments

In this section, we give additional details on the experiments. We use Pytorch [58] for all our experiments.

Data generation To generate the latent DAG and sample the latent variables we use the sempler package [20]. We always sample $\mathrm{ER}(d, k)$ graphs and the non-zero weights are sampled from the distribution $w_{ij} \sim \mathcal{U}(\pm[0.25, 1.0])$. To generate linear synthetic data, we sample all entries of the mixing matrix i.i.d. from a standard Gaussian distribution. For non-linear synthetic data f is given by the architecture in Table 7 with weights sampled from $\mathcal{U}([-\frac{1}{\sqrt{in_{feat}}}, \frac{1}{\sqrt{in_{feat}}}])$. For image data, similar to [2], the latents Z are the coordinates of balls in an image (but we sample them differently as per our setting) and a $64 \times 64 \times 3$ RGB image is rendered using Pygame [72] which encompasses the nonlinearity. Other parameter choices such as dimensions, DAGs, variance parameters, and shifts are already outlined in Section 6 but for clarity we also outline them in Tables 4, 5, and 6.

Table 4: Parameters used for linear synthetic data.

d	5
d'	10
k	3/2
n	$\{2500, 5000, 10000, 25000, 50000\}$
σ^2_{obs}	$\mathcal{U}([2, 4])$
σ^2_{int}	$\mathcal{U}([6, 8])$
$\eta^{(i)}$	0
runs	5
mixing	linear

Table 5: Parameters used for nonlinear synthetic data.

d	$\{5, 10\}$
d'	$\{20, 100\}$
k	$\{1, 2\}$
n	10000
σ^2_{obs}	$\mathcal{U}([1, 2])$
σ^2_{int}	$\mathcal{U}([1, 2])$
$\eta^{(i)}$	$\mathcal{U}(\pm[1, 2])$
runs	5
mixing	3 layer mlp

Table 6: Parameters used for image data.

d	$\{4, 6\}$
d'	$3 \cdot 64^2$
k	$\{1, 2\}$
n	25000
σ^2_{obs}	$\mathcal{U}([0.01, 0.01])$
σ^2_{int}	$\mathcal{U}([0.01, 0.02])$
$\eta^{(i)}$	$\mathcal{U}(\pm[0.1, 0.2])$
runs	5
mixing	image rendering

Models For synthetic data, the model architecture for $h(x, \theta)$ is a one hidden-layer MLP with 512 hidden units and LeakyReLU activation functions. For the parametric final layer, we decompose the matrix W as $W = D(\mathrm{Id} - A)$ where D is a diagonal matrix and A is a matrix with a masked diagonal and they are both learned. This factorization shall improve stability. We initialize $A = 0$ and $D = \mathrm{Id}$. Also, the sparsity penalty and the DAGness penalty are only applied to A. For the baseline VAE we use the same architecture for both encoder and decoder. The code for the linear baseline is from [79].

Table 7: Synthetic data generation

Architecture	Layer Sequence
MLP Embedding	Input: $z \in \mathbb{R}^d$
	FC 512, LeakyReLU(0.2) FC 512, LeakyReLU(0.2) FC 512, LeakyReLU(0.2) FC $d' \longleftarrow$ Output x

For image data, the model architectures for $h(x, \theta)$ are small convolutional networks. For the contrastive approach the architecture can be found in Table 8. For the VAE model, we use the encoder architecture as in Table 9 and the decoder architecture as in Table 10 respectively.

Table 8: Contrastive model architecture for image data.

Architecture	Layer Sequence
Conv Embedding	Input: $x \in \mathbb{R}^{64\times64\times3}$
	Conv (3, 1, kernel size = 5, stride = 3), ReLU() MaxPool (kernel size = 2, stride = 2) FC 64, LeakyReLU() FC $d \longleftarrow$ Output $\hat{z}$

Table 9: VAE Encoder architecture for image data

Architecture	Layer Sequence
Conv Encoder	Input: $x \in \mathbb{R}^{64\times64\times3}$
	Conv(3, 16, kernel size = 4, stride = 2, padding = 1), LeakyReLU() Conv(16, 32, kernel size = 4, stride = 2, padding = 1), LeakyReLU() Conv(32, 32, kernel size = 4, stride = 2, padding = 1), LeakyReLU() Conv(32, 64, kernel size = 4, stride = 2, padding = 1), LeakyReLU() FC $2d \longleftarrow$ Output (mean, logvar)

Table 10: VAE Decoder architecture for image data

Architecture	Layer Sequence
Deconv Decoder	Input: $\hat{z} \in \mathbb{R}^d$
	FC $(3 \times 3) \times d$ DeConv (d, 32, kernel size = 4, stride = 2, padding = 0), LeakyReLU() DeConv (32, 16, kernel size = 4, stride = 2, padding = 1), LeakyReLU() DeConv (16, 8, kernel size = 4, stride = 2, padding = 1), LeakyReLU() DeConv (8, 3, kernel size = 4, stride = 2, padding = 1), LeakyReLU() Sigmoid() $\longleftarrow$ Output $\hat{x}$

Training For training we use the hyperparameters outlined in Table 11. We use a 80–20 split for train and validation/test set respectively. After subsampling each dataset to the same size, we copy the observation samples for each interventional dataset in order to have an equal number of observational and interventional samples so they can be naturally paired during the contrastive learning. We select the model with the smallest validation loss on the validation set, where we only use the cross entropy loss for validation. For the VAE baseline we use the standard VAE validation loss for model selection.

Table 11: Hyperparameters used for training.

τ_1 (see Eq. (9))	10^{-5}
τ_2 (see Eq. (9))	10^{-4}
learning rate	$5 \cdot 10^{-4}$
batch size	512
epochs	200 (synthetic data), 100 (image data)
optimizer	Adam [36]
learning rate scheduler	CosineAnnealing [49]

Compute and runtime The experiments using synthetic data were run on 2 CPUs with 16Gb RAM on a compute cluster. For image data we added a GPU to increase the speed. The runtime per epoch for the synthetic data and 10000 samples was roughly 10s. The total run-time of the reported experiments was about 100h (i.e. 200 CPU hours) and 20h on a GPU, but hyperparameter selection and preliminary experiments required about 20 times more compute.

Mean Correlation Coefficient (MCC) Mean Correlation Coefficient (MCC) is a metric that has been utilized in prior works [35, 40] to quantify identifiability. MCC measures linear correlations up to permutation of the components. To compute the best permutation, a linear sum assignment problem is solved and finally, the correlation coefficients are computed and averaged over. A high MCC indicates that the true latents have been recovered. In this work, we compute the transformation using half of the samples and then report the MCC using the other half, i.e. the out-of-distribution samples.

Further Experimental Attempts and Dead Ends We will briefly summarize a few additional settings we tried in our experiments.

- We fixed $D = \mathrm{Id}$ in the parametrization of W. This fixes the scaling of the latent variables in some non-trivial way. This resulted in very similar results.
- We tried to combine the contrastive algorithm with a VAE approach. Since the ground truth latent variables Z are highly correlated, they are not suitable for an autoencoder that typically assumes a factorizing prior. Therefore, we map the estimated latent variables for the observational distribution to the noise distribution ϵ which factorizes. Note that for the ground truth latent variables the relation $\epsilon = B^{(0)}Z^{(0)}$ holds so we consider $\hat{\epsilon} = W\hat{Z}$. We use $\hat{\epsilon}$ as the latent space of an autoencoder on which we then stack a decoder. We train using the observational samples and the ELBO for the VAE part and with the contrastive loss (and interventional and observational samples) for the $\hat{Z}$ embedding. This resulted in a slightly worse recovery of the latent variables (as measured by the MCC score) but much worse recovery of the graph. We suspect that this originates from the fact that the matrix W is used both in the parametrization of the log-odds and to map $\hat{Z}$ to ϵ which might make the learning more error-prone. Note that this is different from the suggestion in Section F.5.
- Choosing a larger learning rate for the parametric part than for the non-parametric part seemed to help initially, but using the same sufficiently small learning rate for the entire model was more robust.
- We tried larger architectures for both synthetic and image data. For synthetic data, they turned out to be more difficult to train without offering any benefit, while for image data we quickly observed overfitting. The overfitting persisted even when we used a pre-trained ResNet [23] with frozen layers. We suspect that for image data the performance can be further improved by carefully tuning regularization and model size.