Identifying Linearly-Mixed Causal Representations from Multi-Node Interventions

Simon Bing BING@TU-BERLIN.DE
Technische Universität Berlin

Urmi Ninad* URMI.NINAD@TU-BERLIN.DE
Technische Universität Berlin
German Aerospace Center, Institute of Data Science

Jonas Wahl* WAHL@TU-BERLIN.DE
Technische Universität Berlin
German Aerospace Center, Institute of Data Science

Jakob Runge RUNGE@TU-BERLIN.DE
German Aerospace Center, Institute of Data Science
Technische Universität Berlin

Editors: Francesco Locatello and Vanessa Didelez

Abstract

The task of inferring high-level causal variables from low-level observations, commonly referred to as *causal representation learning*, is fundamentally underconstrained. As such, recent works to address this problem focus on various assumptions that lead to identifiability of the underlying latent causal variables. A large corpus of these preceding approaches consider multi-environment data collected under different interventions on the causal model. What is common to almost all of these works is the restrictive assumption that in each environment, only a single variable is intervened on. In this work, we relax this assumption and provide a novel identifiability result for causal representation learning that allows for multiple variables to be targeted by an intervention within one environment. Our approach hinges on a general assumption on the coverage and diversity of interventions across environments, which also includes the shared assumption of single-node interventions of previous works. The main idea behind our approach is to exploit the trace that interventions leave on the variance of the ground truth causal variables and regularizing for a specific notion of sparsity with respect to this trace. In addition to and inspired by our theoretical contributions, we present a practical algorithm to learn causal representations from multi-node interventional data and provide empirical evidence that validates our identifiability results.

Keywords: Causal representation learning, identifiability, interventional data, sparsity

1. Introduction

Across the sciences, data are recorded in the form of low-level measurements of observable physical variables. However, the causal model underlying and robustly explaining the data may operate on high-level variables, sometimes called *causally autonomous* or *causally disentangled* variables. The task of learning such causal representations of data falls under the rubric of *causal representation learning* (Schölkopf et al., 2021). This problem has been shown to be fundamentally underconstrained (Locatello et al., 2019), leading to various approaches that employ inductive biases in order to provably identify the underlying latent causal variables.

* These authors contributed equally.

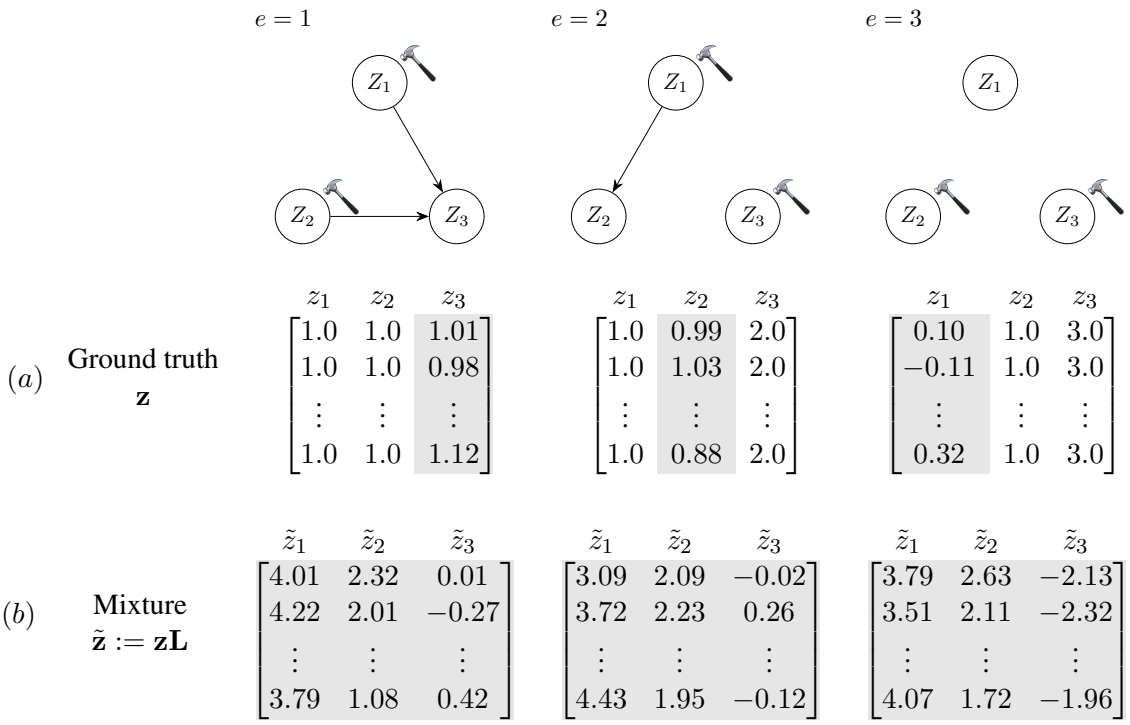


Figure 1: Comparison of samples drawn from the SCM of Example 1 under three different interventions between **(*a*)** the ground truth representation and **(*b*)** a mixed representation. Notice that the density of variables with nonzero variance (indicated by shading) is lower in each environment in the ground truth representation than in the mixed case. We exploit the principle that the ground truth is more sparse in terms of nonzero variance dimensions to achieve identifiability of causal representations using multi-node interventional data.

Recent works that provide such identifiability guarantees either restrict the underlying structural causal model (Lachapelle et al., 2022; Buchholz et al., 2023; Liang et al., 2023), or the transformation mapping the causal variables to the observed data (also called the mixing function) (Ahuja et al., 2023a; Zhang et al., 2023), or both (Squires et al., 2023). Many of these methods additionally assume the availability of interventional or counterfactual data collected across environments (Ahuja et al., 2023a,b; Zhang et al., 2023; Squires et al., 2023; Buchholz et al., 2023; Liang et al., 2023; von Kügelgen et al., 2023, 2021; Brehmer et al., 2022), or use supervisory signals such as time structure (Hyvärinen and Morioka, 2017; Hälvä and Hyvärinen, 2020; Yao et al., 2021), sometimes in addition with knowledge of intervention targets (Lippe et al., 2022a,b). What unites nearly all aforementioned methods that require the availability of interventional data—and guarantee full component-wise identifiability of the latent causal variables—is the restrictive assumption requiring interventions to be *atomic* or *single-node*, i.e., when the number of latent variables intervened on per environment is at most one.

In this work, we relax the restrictive single-node intervention assumption, and design a setup that guarantees the element-wise identifiability of latent causal variables by using data that is collected across multiple interventional environments, given the assumption that the mixing function is linear.

However, importantly, we do not require these interventions to be atomic, only that they are *hard* and sufficiently *diverse* in a specific sense outlined below, see in particular Assumption 4. Our assumption on the nature of interventions generalizes previous works in the sense that they allow for more general interventions, most notably those acting on multiple variables in one environment, while still allowing for the case of atomic interventions. Furthermore, we make no assumptions on the underlying structural causal model (SCM). To prove our main result, we introduce a specific notion of sparsity concerning the trace left in the data by interventions. As we detail in Appendix A.2, this general idea of using sparsity to achieve identifiability of representations is motivated by a similar assumption on the predictor of a multi-task prediction problem of Lachapelle et al. (2022). However, our problem setting is not a prediction task, and therefore the nature of our sparsity assumption is fundamentally different. Our results also support the *sparse mechanism shift* hypothesis, that posits that the distribution shifts of the underlying causal variables across environments must be sparse (Perry et al., 2022).

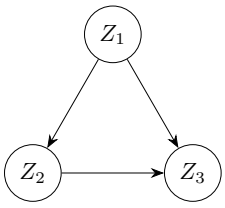
Our **main contributions** can be summarized as follows:

1. We formalize the notion that in the ground truth representation only a *sparse* subset of variables is affected by an intervention, while this effect is dense for variables that have undergone mixing.
2. Based on this, we prove identifiability for the setup of latent variables that have undergone linear mixing, given hard interventional data across environments. The SCM underlying the latent variables can be arbitrarily nonlinear with additive noise. Crucially, the interventional structure is not required to be atomic, but only sparse and diverse in the sense outlined in Assumption 4.
3. We present a proof-of-concept algorithm to recover the latent causal variables up to permutations and rescaling in the setting outlined above, alongside accompanying experiments.

2. Motivating Example

We begin by presenting the following motivating example as an intuitive introduction to our approach to exploit a specific notion of sparsity to disentangle linear mixtures using multi-node interventional data.

Example 1 *Consider the following structural causal model (SCM)*



$$\begin{aligned} Z_1 &:= \eta_1 \\ Z_2 &:= Z_1 + \eta_2 \\ Z_3 &:= Z_1 + Z_2 + \eta_3 \end{aligned} \qquad \eta_j \sim \mathcal{N}(0,1),$$

with $d = 3$ variables $\mathbf{Z} = (Z_1, Z_2, Z_3)$ and independent noise terms η_1, η_2, η_3. Now, consider three environments, with the following corresponding interventions

$$\begin{aligned} I^1 &= \{\text{do}(Z_1 = 1, Z_2 = 1)\}, \\ I^2 &= \{\text{do}(Z_1 = 1, Z_3 = 2)\}, \\ I^3 &= \{\text{do}(Z_2 = 1, Z_3 = 3)\}. \end{aligned}$$

We draw n samples from the above SCM in each environment. The resulting $\mathbf{z}^j \in \mathbb{R}^{n\times d}$ are shown in Fig. 1(a).

Now, consider the same environments as above, but in the case where the variables in $\mathbf{Z}$ are mixed by the invertible matrix

$$\mathbf{L} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix},$$

denoting the mixed variables as $\tilde{\mathbf{Z}} = \mathbf{ZL}$. Again drawing n samples for each environment—we recall that interventions act on the unmixed ground truth variables—the resulting $\tilde{\mathbf{z}}^j \in \mathbb{R}^{n\times d}$ is presented in Fig. 1(b).

Notice that in the unmixed case, in each respective environment, the columns of $\mathbf{z}^j$ which have been intervened on take the same constant value for each sample and thus have zero variance, while those dimensions that are not targeted by interventions have nonzero variance and display stochasticity in the values they may take. Contrarily, in the mixed case, for this specific mixing matrix $\mathbf{L}$, in all environments, all columns $\tilde{\mathbf{z}}^j$ have nonzero variance.

The main observation to be made in the above example is the fact that mixing seems to increase the "density" of those dimensions of $\tilde{\mathbf{Z}}$ which are not constant and have nonzero variance, w.r.t the ground truth representation $\mathbf{Z}$.

Inspired by this observation, we propose to regularize the learned representation such that the effect of interventions is most sparse across environments. In the following, we make this specific notion of sparsity precise and show that this constraint indeed allows us to recover the latent causal factors up to permutation and rescaling (cf. Definition 1), under a general linear mixing matrix $\mathbf{L}$.

3. Disentanglement from Multi-Node Interventions

In this section, we present the main results of this work on identifying causal representations from multi-node interventional data. The basic idea of our approach is to exploit the fact that interventions leave a specific trace in the data, which is most sparse in the ground truth representation. In the following, we make this notion of sparsity formal and prove that regularizing for it yields a representation of the latent variables that is equivalent to the ground truth.

3.1. Problem Setting

Notation. Scalar variables are denoted in normal face (x) and vector-valued variables in bold ($\mathbf{x}$). Random variables are capitalized (Y), the values they take we write in lower case (y). Matrices are capitalized and bold ($\mathbf{M}$) and will be introduced as such. We denote the sequence of integers from 1 to n with $[n]$.

Data Generating Process. Consider the random variables $\mathbf{Z} = (Z_1, \ldots, Z_d)$. Assume $\mathbf{Z} \sim P$ where the unknown joint distribution P is induced by a structural causal model (SCM) defined over the random vector $\mathbf{Z}$. This SCM induces the factorization

$$P(\mathbf{Z}) = \prod_{j=1}^{d} P(Z_j \mid Z_{\mathrm{Pa}_j}), \tag{1}$$

where $\text{Pa}_j \subset [d] \setminus \{j\}$ indicates the parents of variable Z_j. Let $\mathcal{G}$ denote the corresponding directed acyclic graph (DAG) of this SCM. For a detailed definition of SCMs, we refer the reader to Pearl (2009).

We make no parametric assumptions on the causal mechanisms of this SCM, i.e., the structural equation for each variable takes on the general form $Z_j := f_j(Z_{\text{Pa}_j}, \eta_j)$, where $\eta_j; j \in [d]$ are the exogenous, independent noise terms. We assume the distributions of the noises have nonzero variance.

Instead of measuring $\mathbf{Z}$ directly, we only have access to observations $\tilde{\mathbf{Z}} \in \mathbb{R}^m$, where $\tilde{\mathbf{Z}} = \mathbf{ZL}$ is generated from the causal variables by the injective linear map $\mathbf{L} : \mathbb{R}^d \to \mathbb{R}^m$ with $m \geq d$.

Additionally, we assume to have access to observations $\tilde{\mathbf{Z}}$ from different environments, where each environment $e \in \mathcal{E}$ corresponds to a setting where subsets $T \subseteq [d]$ of the underlying, latent variables $\mathbf{Z}$ have been intervened upon. We assume that we are given a probability measure P_E on the set of environments $\mathcal{E}$ with full support $\mathcal{E}$, which describes the distribution of a random environment E. Note that in practice we will have distributions of the latent variables corresponding to different interventional environments, that we can then view as having been drawn from P_E, much in the way outlined in Mooij et al. (2020). We explicitly do not include the observational distribution, i.e., where no variables have been intervened on, in $\mathcal{E}$. We consider do-interventions (Pearl, 2009), sometimes called hard interventions, which replace the structural equations of intervened upon variables with constant values. We write

$$Z_j := a_j \quad \text{for } j \in T, \tag{2}$$

where $\mathbf{a} \in \mathbb{R}^{|T|}$. Each environment e is characterized by its corresponding intervention $I^e := \{(T, \mathbf{a}) \mid T \subseteq [d], \mathbf{a} \in \mathbb{R}^{|T|}\}$ and we denote with P^e the distribution of the random vector $\mathbf{Z}^e$ induced by intervention I^e. We stress that interventions are explicitly not assumed to only target single variables. We assume that the mixing function $\mathbf{L}$ remains unchanged across environments.

Objective. Our goal is to recover the latent variables $\mathbf{Z}$ from observations $\tilde{\mathbf{Z}}$. Specifically, we are interested in exploring how observing transformations of latent variables under different interventional environments $e \in \mathcal{E}$ can be leveraged to learn the underlying causal variables.

Formally, recovering the latent variables $\mathbf{Z}$ from observations $\tilde{\mathbf{Z}}$ amounts to learning the inverse of the mixing function $\mathbf{L}$. We are only interested in latent variables that are equal to the ground truth up to permutation, element-wise rescaling and possible redundancies if the dimensionality of the learned representation is greater than that of the ground truth variables. Consequently, we define an equivalence class over such latents. While weaker notions exist (Hyvärinen and Morioka, 2017), we refer to identifiability as being able to recover the latent variables up to this equivalence. We call equivalent representations *causally disentangled up to redundancies*.

Definition 1 (Causal Disentanglement up to Redundancies) *A learned representation $\hat{\mathbf{Z}} \in \mathbb{R}^m$ is causally disentangled up to redundancies w.r.t. the ground truth representation $\mathbf{Z} \in \mathbb{R}^d$ if there exists a matrix $\mathbf{L} \in \mathbb{R}^{d \times m}$, where*

1. *each column of $\mathbf{L}$ contains at most one nonzero element,*
2. *there are at least d columns in $\mathbf{L}$ with a nonzero element,*

s.t. $\hat{\mathbf{Z}} = \mathbf{ZL}$ almost surely.

This definition ensures that each learned variable $\hat{Z}_i$ is either a scalar multiple of a ground truth variable Z_j, or zero, and that all ground truth variables in $\mathbf{Z}$ appear as scalar multiples in $\hat{\mathbf{Z}}$ at least once. Notice for the special case where $m = d$, the matrix $\mathbf{L}$ in the above definition can be decomposed into a diagonal invertible matrix $\mathbf{D}$ and a permutation matrix $\mathbf{P}$, i.e., $\mathbf{L} = \mathbf{DP}$, and we recover the standard definition of causal disentanglement (Khemakhem et al., 2020; Lachapelle et al., 2022).

3.2. Assumptions

Before presenting our main theorem, we introduce some additional objects and assumptions that are required to make our considerations precise.

First, let us formally define a quantity that allows us to measure the notion of sparsity we are interested in.

Definition 2 (Variance density.) *Let $\mathbf{A} \in \mathbb{R}^d$ be a random vector. Then*

$$\|\mathbf{A}\|_{\mathrm{Var}} := \sum_{j=1}^{d} \mathbb{1}(\mathrm{Var}(A_j) \neq 0),$$

where $\mathbb{1}(\cdot)$ is the indicator function.

The next lemma states that there is no mixing matrix $\mathbf{L}$, such that in the resulting mixture, any variable has variance zero.

Lemma 3 (Non-vanishing variance under mixing.) *For all invertible $\mathbf{L} \in \mathbb{R}^{d\times d}$,*

$$\forall j \in [d], \mathrm{Var}((\mathbf{ZL})_j) \neq 0.$$

The proof is presented in Appendix A.1.

Further, we denote with S^e the set that contains the indices the elements of $\mathbf{Z}^e$ that have nonzero variance in environment e, that is

$$S^e := \{j \in [d] \mid \mathrm{Var}(Z_j^e) \neq 0\}.$$

Because we consider hard interventions that set the values of variables to constant values, S contains all variables that have *not* been intervened upon in this environment.

We consider again the environment distribution P_E which we can factorize as

$$P_E = \sum_{S\in\mathcal{P}([d])} p(S) P_{E|S},$$

where $\mathcal{P}([d])$ denotes all subsets of $[d]$, $p(S) = P_E(\{e \in \mathcal{E} \mid S^e = S\})$ and $P_{E|S}$ is the conditional distribution $P_{E|S}(\,\cdot\,) := P_E(\,\cdot\,|S) = \frac{P_E(\,\cdot\,\cap S)}{p(S)}$ if $p(S) \neq 0$ (and 0 otherwise). We denote with $\mathcal{S}$ the support of the distribution $p(S)$, i.e., $\mathcal{S} := \{S \in \mathcal{P}([d]) \mid p(S) > 0\}$. The next assumption concerns $\mathcal{S}$ and in words states that the environments we observe must be diverse enough, such that for each latent dimension j, if we consider all environments where we intervene on variable Z_j, we do not always simultaneously intervene on another variable Z_i.

Assumption 4 (Sufficient coverage of interventions.) *For all* $j \in [d]$

$$\bigcup_{S \in \mathcal{S} | j \notin S} S = [d] \setminus \{j\}.$$

This assumption is arguably quite general, therefore some remarks on its implications are in order. First, notice that if a given variable Z_j is intervened on in *all* environments $\mathcal{E}$, the assumption cannot be fulfilled. Further, the set of environments that contain single-node interventions for all variables are subsumed and permitted by the above assumption, showing that it generalizes the most common assumption of preceding works.

A particularly interesting implication of this assumption lies in its connection to the notion of *separating systems*, first introduced by Katona (1966). This concept has proven useful for experimental design for causal discovery (Hyttinen et al., 2013; Shanmugam et al., 2015; Kocaoglu et al., 2017a,b) and also allows us to derive a lower bound on the number of environments required to fulfill the above assumption and therefore ultimately learn the underlying causal representation. Specifically, Assumption 4 is equivalent to the definition of *strongly* separating systems in Kocaoglu et al. (2017b, Definition 1), which by Kocaoglu et al. (2017b, Lemma 2) allows us to directly conclude that we can satisfy our central assumption with a minimum of $2\lceil \log d \rceil$ environments. This stands in contrast to requiring d environments when only single-node interventions are considered and highlights that allowing multi-node interventions can lead to requiring less environments overall, when the interventions can be chosen freely. Note that Lippe et al. (2022c) show a similar result for a lower bound on required environments for causal representation learning, however these are specific to their time series setting that additionally requires knowledge of intervention targets in each environment.

3.3. Identifiability Result

We are now ready to present our main identifiability result. Intuitively, it exploits the fact that in the ground truth representation, interventions set certain dimensions to a constant value, leaving only a sparse subset of all dimensions with nonzero variance. Conversely, under some linear mixing, generally *all* latent dimensions will have nonzero variance terms. The idea is then to find a transformation that yields a representation that is as sparse as possible in the aforementioned sense, which we show results in recovering latent variables that are equivalent to the ground truth.

Theorem 5 (Disentanglement via intervention sparsity.) *Assume the data generating process described in Section 3.1. Let* $\mathbf{L} \in \mathbb{R}^{d \times m}$ *be an injective matrix and* $\hat{\mathbf{Z}} = \mathbf{ZL}$*. Suppose Assumption 4 holds. Then, if*

$$\mathbb{E}_{P_E} \|\hat{\mathbf{Z}}^E\|_{\text{Var}} \leq \mathbb{E}_{P_E} \|\mathbf{Z}^E\|_{\text{Var}},$$

$\hat{\mathbf{Z}}$ *is causally disentangled up to redundancies w.r.t.* $\mathbf{Z}$ *(cf. Definition 1), where* $\|\mathbf{Z}^{\cdot}\|_{\text{Var}}$ *is defined as the function* $\|\mathbf{Z}^{\cdot}\|_{\text{Var}} : (\mathcal{E}, P_E) \to \mathbb{R},\ e \mapsto \|\mathbf{Z}^e\|_{\text{Var}}$.

In particular, if $m = d$*, it holds that* $\mathbf{L} = \mathbf{DP}$*, where* $\mathbf{D}$ *is a diagonal invertible matrix and* $\mathbf{P}$ *is a permutation matrix.*

In words, Theorem 5 shows that under our assumptions, any representation that is as sparse as the ground truth representation (w.r.t. the variance density) must already be causally disentangled. This will motivate the optimization approach of the following section. The proof of Theorem 5 is inspired by that of Lachapelle et al. (2023, Theorem B.5), and is presented in Appendix A.2.

4. Experiments

We illustrate the result presented in Theorem 5 by providing empirical evidence that we can recover the ground truth causal variables from observed data when our assumptions are met. We provide code to reproduce all of our experimental results[1].

4.1. Practical Algorithm

Based on our theoretical results presented in Section 3 we present an algorithm that implements the main principles of our theorem in terms of practical constraints, to recover the underlying causal representation from data.

Since the observed data $\tilde{\mathbf{Z}}$ is generated from the ground truth data $\mathbf{Z}$ according to $\tilde{\mathbf{Z}} = \mathbf{ZL}$, our algorithm aims to learn a linear map $\hat{\mathbf{L}} : \mathbb{R}^m \to \mathbb{R}^d$, such that the resulting $\hat{\mathbf{Z}} = \mathbf{Z}\hat{\mathbf{L}}\hat{\mathbf{L}}$ is disentangled w.r.t the ground truth $\mathbf{Z}$ (cf. Definition 1). We learn $\hat{\mathbf{L}}$ via stochastic gradient descent by optimizing over the loss function

$$\min_{\hat{\mathbf{L}}} \mathcal{L}_{\text{Var}} + \lambda_e \mathcal{L}_e + \lambda_m \mathcal{L}_m + \lambda_{\text{diag}} \mathcal{L}_{\text{diag}} + \lambda_{\text{norm}} \mathcal{L}_{\text{norm}}, \tag{3}$$

where the individual loss terms are introduced in the following.

Since our theoretical results exploit assumptions on the sparsity of nonzero variance terms across environments, we introduce the matrix $\mathbf{V} \in \mathbb{R}^{e \times m}$, where the elements in row $\mathbf{V}_{i:}$ contain the variance of $\tilde{\mathbf{Z}}$ in environment $e = i$; $\mathbf{V}_{i,j} := \text{Var}(\tilde{Z}_j^i)$. In words, $\mathbf{V}$ contains the stacked variance terms of $\tilde{\mathbf{Z}}$ in each environment.

Total Variance Support. The first term in Eq. (3), $\mathcal{L}_{\text{Var}}$, captures the assumption that the support of nonzero variance terms, i.e., $\sum_{i,j} \mathbb{1}(\mathbf{V}_{i,j} \neq 0)$, is minimal across environments. Since the indicator function is not differentiable, we use the sigmoid function $\sigma(x) := \frac{1}{1+e^{-x}}$ and define

$$\mathcal{L}_{\text{Var}} := \sum_{i,j}^{e,m} \sigma(\mathbf{V}_{i,j}).$$

Per-Environment Variance Support. The next term, $\mathcal{L}_e$, enforces that there is no environment in which all variables are intervened upon at the same time. In terms of $\mathbf{V}$, this means that every row $\mathbf{V}_{i:}$ contains at least one nonzero entry, which implies that the sum across each row i is nonzero, i.e., $\mathbb{1}((\sum_j^m \mathbf{V}_{i,j}) \neq 0)$. Again replacing the counting operation of the indicator function with a sum over sigmoid functions, we write

$$\mathcal{L}_e := -\sum_i^e \sigma\big(\sum_j^m \mathbf{V}_{i,j}\big),$$

where the negative sign comes from the fact that we want this term to be nonzero.

Per-Dimension Variance Support. This term reflects the deliberation that there is no dimension $j \in [m]$ which is intervened upon in all environments. Analogous to $\mathcal{L}_e$, in terms of $\mathbf{V}$ this means that we enforce that each column $\mathbf{V}_{:j}$ contains at least one nonzero entry and we write

$$\mathcal{L}_m := -\sum_j^m \sigma\big(\sum_i^e \mathbf{V}_{i,j}\big).$$

1. https://github.com/simonbing/Multi-Node-CRL

Diagonal Sparsity. Consider the exemplary variance matrix for $m = 3$ and with three environments

$$\mathbf{V} = \begin{bmatrix} 0 & 0 & * \\ 0 & 0 & * \\ * & * & 0 \end{bmatrix},$$

where nonzero entries are denoted with $*$. $\mathbf{V}$ satisfies the considerations represented by $\mathcal{L}_e$ and $\mathcal{L}_m$ (no nonzero rows or columns), but clearly is in violation of Assumption 4, since Z_1 and Z_2 are always intervened on together. Considering the same number of environments, a variance matrix that satisfies all of the above constraints could, e.g., be

$$\mathbf{V}' = \begin{bmatrix} 0 & 0 & * \\ * & 0 & 0 \\ 0 & * & 0 \end{bmatrix},$$

where all but one variable is intervened on per environment. Let the k-th diagonal of the $(d \times d)$-matrix $\mathbf{A}$ be defined as $\text{diag}_k(\mathbf{A}) := \{\mathbf{A}_{i,j} \,\forall\, i, j \in [d] \mid (i+k-1) \mod d = j \mod d\}$. In words, $\text{diag}_k(\mathbf{A})$ describes the diagonal elements of the matrix $\mathbf{A}$ that "wrap around", where any $k > 1$ denotes the offset from the main diagonal. For a 3×3-matrix, consider the following example:

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{bmatrix}, \qquad \begin{aligned} \text{diag}_1(\mathbf{A}) &= \{a_{1,1}, a_{2,2}, a_{3,3}\}, \\ \text{diag}_2(\mathbf{A}) &= \{a_{1,2}, a_{2,3}, a_{3,1}\}, \\ \text{diag}_3(\mathbf{A}) &= \{a_{1,3}, a_{2,1}, a_{3,2}\}. \end{aligned}$$

If we compare the diagonals of $\mathbf{V}$ and $\mathbf{V}'$ we notice that $\mathbf{V}'$—which is aligned with our assumptions—has more diagonals that contain only zeros than $\mathbf{V}$. Based on this observation, we formulate the loss term $\mathcal{L}_{\text{diag}}$.

We operationalize the above argument that $\mathbf{V}$ should contain as many $\mathbf{0}$-diagonals as possible by recalling the $\ell_{2,1}$ norm for matrices, defined as $\|A\|_{2,1} := \sum_{j=1}^{d} \|A_{:j}\|$, where $\|\cdot\|$ denotes the Euclidean norm for vectors. Since regularizing with the $\ell_{2,1}$ norm is known to promote column sparsity (Argyriou et al., 2008), we define an analogous regularization to promote sparsity of diagonals:

$$\mathcal{L}_{\text{diag}} := \sum_{j=1}^{m} \|\text{diag}_j(\mathbf{V})\|.$$

Norm Regularization. To prevent $\mathbf{V}$ from collapsing to all zeros, we enforce the Frobenius norm of $\hat{\mathbf{L}}$ to take a fixed value $a \in \mathbb{R}_{>0}$. In practice, we choose $a = 1$ and define

$$\mathcal{L}_{\text{norm}} := (\|\hat{\mathbf{L}}\| - a)^2.$$

4.2. Synthetic Data Generation

We generate synthetic data by sampling from randomly generated DAGs according to the Erdős–Rényi model (Erdős and Rényi, 1959), where we change the number of nodes d and the probability of an edge being present in the graph p across experiments. Unless stated otherwise, we consider linear

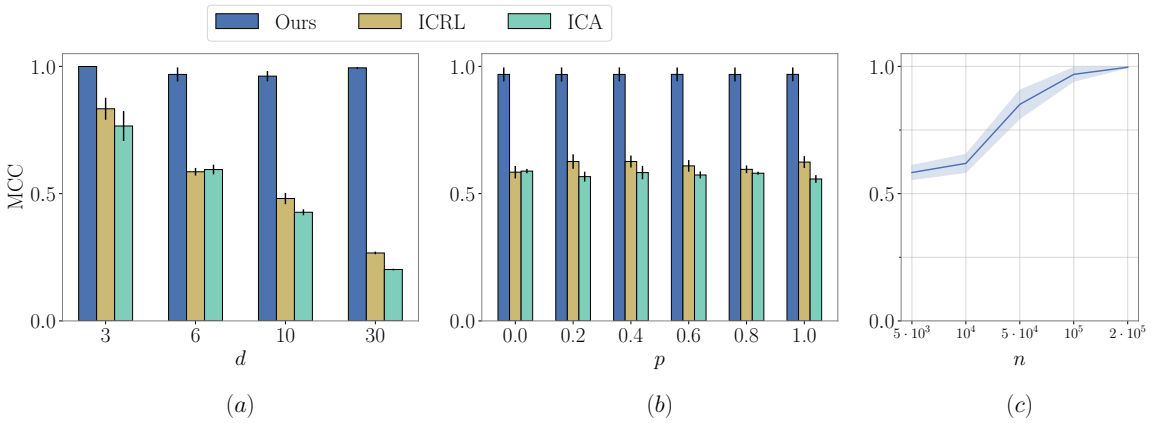


Figure 2: We report the mean MCC score across various experimental conditions, over five random seeds. Error bars or shaded regions indicate the standard error. **(*a*)** Our model performs well across all considered number of latent variables d, even up to $d = 30$. **(*b*)** MCC across different probabilities of an edge being present p. Our method achieves near-perfect score across all settings, indicating that we do not implicitly rely on assumptions on the density of the underlying graph. **(*c*)** MCC for different sample sizes n. Performance increases with sample size and saturates at $n = 2 \cdot 10^5$.

causal mechanisms $f_j := \sum_{j \in \text{Pa}_j} \alpha_j Z_j + \eta_j$ for all variables Z_j, where we sample the coefficients α_j independently from $\mathcal{U}[-0.1, 1.0]$. The random noise η_j is independently sampled from $\mathcal{N}(0, 0.1)$ for all j. For all experiments, except the one in which we investigate the effect of varying sample size, we sample $n = 10^5$ data points per environment.

To generate interventional data, we consider environments where for each latent variable Z_j we intervene on all *other* variables $Z_i, i \neq j$, satisfying Assumption 4.

Since we only assume access to a mixture $\tilde{\mathbf{Z}} = \mathbf{ZL}$ of the causal variables, after sampling according to the procedure detailed above, we randomly sample an invertible matrix $\mathbf{L}$ and apply this mixing. For the results shown in this section we sample one $\mathbf{L}$ and keep it fixed across runs, to focus on the randomness induced by different initializations. In practice we found that different mixing matrices $\mathbf{L}$ do not greatly influence the reported results, which we show with additional experiments in Appendix B.5.

4.3. Results

We provide the results of various experimental settings and compare against two exemplary baselines; linear independent component analysis (ICA) (Comon, 1994) and interventional causal representation learning (ICRL) (Ahuja et al., 2023a). These are unfair comparisons, since both methods' assumptions are misaligned with our setting; ICA assumes independent, non-gaussian latent variables and ICRL assumes single-node interventions per environment. This comparison is not meant to show that our proposed method outperforms these approaches per se, but rather that our problem setting is not a contrived reformulation of a possibly simpler case, where ICA and ICRL are representative baselines for simpler settings. The main conclusion we can draw from our empirical findings is that our method recovers the ground truth latent variables in a more general problem setting than previous approaches.

Model	MCC
Nonlin. SCM 1	0.96 ± 0.03
Nonlin. SCM 2	0.97 ± 0.05

Table 1: MCC (mean $\pm$ standard error) over five different random seeds for two different nonlinear models. Our method performs equally well as for linear SCMs.

To quantify how well we achieve our objective of recovering the equivalence class described in Definition 1 we report the mean correlation coeeficient (MCC) (Hyvärinen and Morioka, 2016; Khemakhem et al., 2020) (cf. Appendix B.4), which is precisely aligned with our notion of identifiability.

As suggested by our theoretical findings, across all experimental settings, our method achieves an almost perfect MCC score.

Effect of SCM Size. We investigate the effect of different underlying SCM sizes by considering $d \in \{3, 6, 10, 30\}$. Shown in Fig. 2(a), we see that our model's performance does not noticeably fall off, even up to $d = 30$. Both methods we compare against, ICA and ICRL, do not perform well, corroborating that our considered setting is indeed more general than their respective settings, and cannot be solved by either model.

SCM Density. In this experiment, we investigate the sensitivity of our approach to the density of the underlying ground truth causal graph. We vary the probability p of an edge being present in a graph with $d = 6$ nodes from 0 to 1, interpolating between a completely disconnected and a fully connected graph. The results are reported in Fig. 2(b). We see that the performance of our approach is equally high across all values of p, underlining that we do not rely on assumptions regarding the sparsity of the underlying causal graph. Interestingly, our method can disentangle independent Gaussian latent variables ($p = 0$), which is precisely the case in which ICA fails.

Nonlinear SCMs. We consider two different SCMs, both with nonlinear mechanism functions f_j for each Z_j. Both models consist of $d = 6$ variables, with an adjacency matrix randomly sampled according to the procedure described in Section 4.2. In the first model, the mechanism functions are defined as $f_j := \sum_{j \in \text{Pa}_j} Z_j^2 + \eta_j$ for all Z_j, while the second model consists of more complex nonlinear mechanisms. See Appendix B.3 for a detailed description of both models. We report the MCC scores for both models in Table 1 and see that our method achieves similarly good results as on linear SCMs, underlining that we do not rely on parametric assumptions on the underlying causal model.

Number of Samples. We investigate the influence of sample size by increasing the number of data points n sampled per environment, for an SCM with $d = 6$. The results are visualized in Fig. 2(c), indicating that disentanglement is achieved after observing around 10^5 samples per environment, while the score saturates and perfect results are achieved after $2 \cdot 10^5$ samples.

5. Related Work

First exemplified by the impossibility result of nonlinear independent component analysis (ICA) (Hyvärinen and Pajunen, 1999), the general problem of representation learning is known to be heavily underconstrained (Locatello et al., 2019). Just as research in ICA has progressed by making

assumptions explicit and subsequently exploiting these assumptions for inference (Hyvärinen and Morioka, 2017; Hälvä and Hyvärinen, 2020; Gresele et al., 2021; Morioka and Hyvärinen, 2023), recent works in causal representation learning works propose various assumptions to enable identifiability. One such inductive bias is to assume and exploit time structures, e.g., in Lippe et al. (2022b,a, 2023) by assuming knowledge of interventions targets or types, in Yao et al. (2021, 2022) by the assumption of nonstationarity or in Lachapelle et al. (2022, 2024) by imposing sparsity constraints on the underlying SCM. A complementary line of works focuses on constraining the kind of mixing the latent variables have undergone, with a particular interest in the case of linear mixing. Squires et al. (2023) show results for linear SCMs that undergo linear mixing, generalized by Buchholz et al. (2023) to nonparametric SCMs, and Varıcı et al. (2023) provide a score-based approach for learning representations under linear mixtures. One common assumption made across approaches is some kind of heterogeneity assumption on the available data, either induced by counterfactual pairs (von Kügelgen et al., 2021; Brehmer et al., 2022) or interventional data, e.g., hard do-interventions as in Ahuja et al. (2023a) or soft interventions as in Zhang et al. (2023). What unites almost all works that include identifiability results based on some form of interventional (or counterfactual) data is the assumption that per environment, interventions only target a *single* node (Brehmer et al., 2022; Squires et al., 2023; Buchholz et al., 2023; Ahuja et al., 2023a; Zhang et al., 2023; Varıcı et al., 2023; Liang et al., 2023; von Kügelgen et al., 2023). While there exist identifiability results with multi-node interventions in the time series setting they come with additional assumptions or limitations that we do not require. The results of Lachapelle et al. (2022, 2024) only hold for an empty graph when only a single time step is considered and the results of Lippe et al. (2022a,b, 2023) require observing an auxiliary variable related to interventions at a given time step to show identifiability. Ahuja et al. (2023b) and Liang et al. (2023) also consider multi-node interventions, however they require additional assumptions and only guarantee identifiability up to blocks of variables, as opposed to our stronger component-wise identifiability.

An approach that utilizes a closely related assumption of sparsity to ours is presented by Lachapelle et al. (2023). The authors consider a prediction problem in the context of representation learning, where the source of heterogeneity does not stem from interventions on the underlying SCM as in our case, but from multiple prediction tasks that share the same representation as potential predictors. By assuming that only a sparse subset of all covariates are used per task, the latents can be identified up to permutation and rescaling. Analogously, we guarantee identifiability by assuming that a sparse subset of the ground truth latent variables have nonzero variance under interventions.

Another line of works frames causal representation learning as being closer to the setting of classical causal structure learning (Spirtes et al., 2001) with latent variables, commonly assuming that at least some nodes of the target graph are observed. While we asssume the observed variables to be deterministic functions of the latents, these works consider observational noise. Most approaches here rely on a variation of the *pure children assumption*, stating that each observed variable has only a single latent parent (Silva et al., 2006). Cai et al. (2019) consider the case where observations are linear transformations of the latents and each latent has two pure children. Xie et al. (2020) generalize the preceding assumptions to allow latents with multiple children, formalized in their Generalized Independent Noise condition and in later work use this condition to learn hierarchical latent variable models (Xie et al., 2022).

6. Discussion

In this work, we present a novel identifiability result for learning causal representations under linear mixing. Our main contribution lies in the generalization of a large body of preceding works that assume that the environments across which the data is collected correspond to single-node or atomic interventions on the latent variables, to the case where multi-node or non-atomic interventions are allowed as well. To enable identifiability from multi-node interventional data, we introduced a novel notion of sparsity regarding latent dimensions with nonzero variance terms.

In addition to our theoretical contribution, we presented empirical evidence in the form of a proof-of-concept practical algorithm that recovers the latent variables of an underlying causal model in synthetic data settings where our assumptions are met.

Limitations. While linear mixing is also directly assumed in Squires et al. (2023), it is still a strong assumption. However, we argue that identifying causal variables from linear mixtures can be seen as a crucial module in a larger learning pipeline. Many complementary approaches result in recovering linear mixtures, such as assuming polynomial mixing functions (Ahuja et al., 2023a), considering multi-task prediction problems (Lachapelle et al., 2023), learning of nonlinear causal effects with anchor variables (Saengkyongam et al., 2023), or a large class of problems where deep neural networks are used (Roeder et al., 2021).

Our proposed practical implementation of our theory into a learning algorithm provides empirical evidence that our assumptions suffice for identifying causal variables using multi-node data. However, we only consider one class of interventions that satisfy our assumptions, while there are of course many more families of interventions that do so as well. While our algorithm is likely not the most general implementation of our theory, we stress that it merely stands as an empirical proof-of-concept and generalizing it to all possible cases where our assumptions are fulfilled was beyond the scope of this work.

Outlook. Since we consider hard (or do-) interventions in this work, an interesting avenue for future research would be to explore if this assumption can be relaxed to a more general family of interventions, similarly to how Zhang et al. (2023) generalize the work of Ahuja et al. (2023a). Beyond that, exploring if our sparsity of nonzero variance dimensions principle is also applicable to nonlinear mixtures presents itself as a natural next step.

Acknowledgments

The authors thank Tom Hochsprung for fruitful discussions and comments, as well as the anonymous reviewers for their feedback and suggestions that helped improve the manuscript. This work received funding from the European Research Council (ERC) Starting Grant CausalEarth under the European Union's Horizon 2020 research and innovation program (Grant Agreement No. 948112). S.B. received support from the German Academic Scholarship Foundation.

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional Causal Representation Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 372–407, 2023a.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-Domain Causal Representation Learning via Weak Distributional Invariances. *arXiv preprint arXiv:2310.02854*, 2023b.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 35, pages 38319–38331, 2022.

Simon Buchholz, Goutham Rajendran, Elan Rosenfeld, Bryon Aragam, Bernhard Schölkopf, and Pradeep Ravikumar. Learning Linear Causal Representations from Interventions under General Nonlinear Mixing. *arXiv preprint arXiv:2306.02235*, 2023.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad Constraints for Learning Causal Structure of Latent Variables. In *Advances in Neural Information Processing Systems*, 32, pages 12883–12892, 2019.

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

Paul Erdős and Alfréd Rényi. On Random Graphs I. *Publicationes Mathematicae Debrecen*, 6: 290–297, 1959.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 34, pages 28233–28248, 2021.

Antti Hyttinen, Frederick Eberhardt, and Patrik O. Hoyer. Experiment Selection for Causal Discovery. *Journal of Machine Learning Research*, 14(93):3041–3071, 2013.

A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4):411–430, 2000.

Aapo Hyvärinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, 29, pages 3772–3780, 2016.

Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pages 460–469, 2017.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov Nonlinear ICA: Unsupervised Learning from Nonstationary Time Series. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, pages 939–948, 2020.

Gyula Katona. On separating systems of a finite set. *Journal of Combinatorial Theory*, 1(2):174–194, 1966.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of The 23rd International Conference on Artificial Intelligence and Statistics*, pages 2207–2217, 2020.

Murat Kocaoglu, Alex Dimakis, and Sriram Vishwanath. Cost-Optimal Learning of Causal Graphs. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1875–1884, 2017a.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental Design for Learning Causal Graphs with Latent Variables. In *Advances in Neural Information Processing Systems*, 30, 2017b.

Sébastien Lachapelle and Simon Lacoste-Julien. Partial Disentanglement via Mechanism Sparsity. *arXiv preprint arXiv:2207.07732*, 2022.

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

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between Disentanglement and Sparsity: Generalization and Identifiability in Multi-Task Learning. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18171–18206, 2023.

Sébastien Lachapelle, Pau Rodríguez López, Yash Sharma, Katie Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Actions, Interventions and Sparse Temporal Dependencies. *arXiv preprint arXiv:2401.04890*, 2024.

Wendong Liang, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal Component Analysis. *arXiv preprint arXiv:2305.17225*, 2023.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In *International Conference on Learning Representations*, 2022a.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. CITRIS: Causal Identifiability from Temporal Intervened Sequences. In *Proceedings of the 39th International Conference on Machine Learning*, pages 13557–13603, 2022b.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. Intervention Design for Causal Representation Learning. *UAI 2022 Workshop on Causal Representation Learning*, 2022c.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M. Asano, Taco Cohen, and Efstratios Gavves. BISCUIT: Causal Representation Learning from Binary Interactions. In *Proceedings of the 39th Conference on Uncertainty in Artificial Intelligence*, pages 1263–1273, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124, 2019.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2018.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *The Journal of Machine Learning Research*, 21(1):99:3919–99:4026, 2020.

Hiroshi Morioka and Aapo Hyvärinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, pages 3399–3426, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 32, pages 8026–8037, 2019.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Ronan Perry, Julius von Kügelgen, and Bernhard Schölkopf. Causal Discovery in Heterogeneous Environments Under the Sparse Mechanism Shift Hypothesis. In *Advances in Neural Information Processing Systems*, 35, pages 10904–10917, 2022.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On Linear Identifiability of Learned Representations. In *Proceedings of the 38th International Conference on Machine Learning*, pages 9030–9039, 2021.

Sorawit Saengkyongam, Elan Rosenfeld, Pradeep Ravikumar, Niklas Pfister, and Jonas Peters. Identifying Representations for Intervention Extrapolation. *arXiv preprint arXiv:2310.04295*, 2023.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Karthikeyan Shanmugam, Murat Kocaoglu, Alexandros G. Dimakis, and Sriram Vishwanath. Learning Causal Graphs with Small Interventions. In *Advances in Neural Information Processing Systems*, 28, pages 3195–3203, 2015.

Ricardo Silva, Richard Scheine, Clark Glymour, and Peter Spirtes. Learning the Structure of Linear Latent Variable Models. *Journal of Machine Learning Research*, 7(8):191–246, 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. The MIT Press, 2001.

Chandler Squires, Anna Seigal, Salil S. Bhate, and Caroline Uhler. Linear Causal Disentanglement via Interventions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 32540–32560, 2023.

Burak Varıcı, Emre Acartürk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based Causal Representation Learning with Interventions. *arXiv preprint arXiv:2301.08230*, 2023.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *Advances in Neural Information Processing Systems*, 34, pages 16451–16467, 2021.

Julius von Kügelgen, Michel Besserve, Wendong Liang, Luigi Gresele, Armin Kekić, Elias Bareinboim, David M. Blei, and Bernhard Schölkopf. Nonparametric Identifiability of Causal Representations from Unknown Interventions. *arXiv preprint arXiv:2306.00542*, 2023.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. In *Advances in Neural Information Processing Systems*, 33, pages 14891–14902, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of Linear Non-Gaussian Latent Hierarchical Structure. In *Proceedings of the 39th International Conference on Machine Learning*, pages 24370–24387, 2022.

Weiran Yao, Yuewen Sun, Alex Ho, Changyin Sun, and Kun Zhang. Learning Temporally Causal Latent Processes from General Temporal Data. In *International Conference on Learning Representations*, 2021.

Weiran Yao, Guangyi Chen, and Kun Zhang. Temporally Disentangled Representation Learning. In *Advances in Neural Information Processing Systems*, 35, pages 26492–26503, 2022.

Jiaqi Zhang, Chandler Squires, Kristjan Greenewald, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability Guarantees for Causal Disentanglement from Soft Interventions. *arXiv preprint arXiv:2307.06250*, 2023.

Appendix A. Proofs

A.1. Auxiliary Lemmata

In order to prove our main result in Theorem 5, we require some additional lemmata, which we state and prove here.

Lemma 3 (Non-vanishing variance under mixing.) *For all invertible* $\mathbf{L} \in \mathbb{R}^{d\times d}$,

$$\forall j \in [d], \mathrm{Var}((\mathbf{ZL})_j) \neq 0.$$

Proof

We allow for arbitrary SCMs with jointly independent noise terms, which are assumed to be non-degenerate in the sense that no variable Z_i is independent of its own noise term η_i. In other words, no variable is deterministically determined by its parents.

Let us label the Z_i's such that they are topologically ordered, i.e., Z_i is causally prior to Z_j for $i < j$ and Z_d is at bottom of the topological order. Each Z_i can be expressed as a function of its parents and respective noise term, i.e.,

$$Z_i = f_i(Z_1, \dots Z_{i-1}, \eta_i) \,, \tag{4}$$

where in this formulation we do not require functions f_i to be minimal, that is f_i is not required to depend on all of its input arguments. Additionally the noises η_i are jointly independent and have non-vanishing variance. Solving the latent variables in terms of the independent noises η yields,

$$Z_i = g_i(\eta_1, \dots, \eta_i) \,, \tag{5}$$

where the g_i's can be calculated by successively substituting structural equations of parents into those of children.

Consider an invertible matrix $\mathbf{L} \in \mathbb{R}^{d\times d}$ and an arbitrary index $j \in [d]$. Since $\mathbf{L}$ is invertible, its j-th column must have nonzero entries which we will denote by $\alpha_1 = (\mathbf{L})_{i_1,j}, \dots, \alpha_s = (\mathbf{L})_{i_s,j} \in \mathbb{R}\backslash\{0\}$ where $i_1 < i_2 < \cdots < i_s$. Then, we can write

$$\tilde{Z} := (\mathbf{ZL})_j = \sum_{k=1}^{s} \alpha_k Z_{i_k}. \tag{6}$$

Now, assume $\mathrm{Var}(\tilde{Z}) = 0$ from which we will derive a contradiction. Since the square root of the variance is a norm (the L^2-norm to be precise), $\mathrm{Var}(\tilde{Z}) = 0$ implies that

$$\sum_{k=1}^{s} \alpha_k Z_{i_k} = \sum_{k=1}^{s} \alpha_k g_k(\eta_{i_1}, \dots, \eta_{i_k}) = 0 \tag{7}$$

almost surely with respect to the noise distribution. Since all α_k's are nonzero we have

$$Z_{i_s} = -\frac{1}{\alpha_s} \sum_{k=1}^{s-1} \alpha_k g_k(\eta_{i_1}, \dots, \eta_{i_k}). \tag{8}$$

Hence, we have expressed Z_{i_s} as a function of the noise terms up to i_{s-1}, which by the joint independence of noise terms implies $Z_{i_s} \perp \eta_{i_s}$. This contradicts our non-degeneracy assumption stated at the beginning of the proof. ■

Before presenting the proof of our main theorem, we require an additional technical lemma, adapted from Lachapelle et al. (2023, Lemma B.1). We restate it here.

Lemma 4 (Invertible matrices contain a permutation, (Lachapelle et al., 2023)) *Let $\mathbf{L} \in \mathbb{R}^{d \times d}$ be an invertible matrix. There, there exists a permutation $\sigma : [d] \to [d]$ such that $\mathbf{L}_{i,\sigma(i)} \neq 0$ for all $i \in [d]$.*

Proof Since $\mathbf{L}$ is invertible, its determinant is nonzero, i.e.,

$$\det(\mathbf{L}) := \sum_{\sigma \in \mathfrak{S}_d} \operatorname{sign}(\sigma) \prod_{i=1}^{d} \mathbf{L}_{i,\sigma(i)} \neq 0,$$

where $\mathfrak{S}_d$ is the set of d-permutations. This equation implies that at least one term of the sum is nonzero, meaning there exists $\sigma \in \mathfrak{S}_d$ such that for all $i \in [d], \mathbf{L}_{i,\sigma(i)} \neq 0$. ■

A.2. Proof of Theorem 5

Our main proof is inspired by that of Lachapelle et al. (2023, Theorem B.5). While they prove that enforcing a certain notion of sparsity in the context of multi-task prediction problems allows for learning disentangled representations, we can adapt their argument to the setting where we have data from multiple environments, by considering our specific definition of sparsity w.r.t. nonzero variance terms in interventional data (cf. Section 3.2).

Theorem 5 (Disentanglement via intervention sparsity.) *Assume the data generating process described in Section 3.1. Let $\mathbf{L} \in \mathbb{R}^{d \times m}$ be an injective matrix and $\hat{\mathbf{Z}} = \mathbf{Z}\mathbf{L}$. Suppose Assumption 4 holds. Then, if*

$$\mathbb{E}_{P_E} \|\hat{\mathbf{Z}}^E\|_{\mathrm{Var}} \leq \mathbb{E}_{P_E} \|\mathbf{Z}^E\|_{\mathrm{Var}},$$

$\hat{\mathbf{Z}}$ is causally disentangled up to redundancies w.r.t. $\mathbf{Z}$ (cf. Definition 1), where $\|\mathbf{Z}^{\cdot}\|_{\mathrm{Var}}$ is defined as the function $\|\mathbf{Z}^{\cdot}\|_{\mathrm{Var}} : (\mathcal{E}, P_E) \to \mathbb{R},\ e \mapsto \|\mathbf{Z}^e\|_{\mathrm{Var}}$.

In particular, if $m = d$, it holds that $\mathbf{L} = \mathbf{D}\mathbf{P}$, where $\mathbf{D}$ is a diagonal invertible matrix and $\mathbf{P}$ is a permutation matrix.

Proof We begin here with the case where $m = d$ from which we will then derive the more general case with $m > d$.

Since $\hat{\mathbf{Z}} = \mathbf{Z}\mathbf{L}$ we can write $\mathbb{E}\|\mathbf{Z}^E\mathbf{L}\|_{\mathrm{Var}} \leq \mathbb{E}\|\mathbf{Z}^E\|_{\mathrm{Var}}$, where we drop the subscript of the expectation for brevity.

We write

$$\begin{aligned}\mathbb{E}\|\mathbf{Z}^E\|_{\mathrm{Var}} &= \mathbb{E}_{p(S)}\mathbb{E}_{P_{E|S}}\left[\sum_{j=1}^{d}\mathbb{1}(\mathrm{Var}(Z_j^E)\neq 0)\mid S\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}\mathbb{E}_{P_{E|S}}\left[\mathbb{1}(\mathrm{Var}(Z_j^E)\neq 0)\mid S\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}P_{E|S}\left[\mathrm{Var}(Z_j^E)\neq 0\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}\mathbb{1}(j\in S),\end{aligned}$$

where the last step uses the definition of S.

Next, we apply a similar treatment to $\mathbb{E}\|\mathbf{Z}^E\mathbf{L}'\|_{\mathrm{Var}}$ and write

$$\begin{aligned}\mathbb{E}\|\mathbf{Z}^E\mathbf{L}\|_{\mathrm{Var}} &= \mathbb{E}_{p(S)}\mathbb{E}_{P_E}\left[\sum_{j=1}^{d}\mathbb{1}(\mathrm{Var}(\mathbf{Z}^E\mathbf{L}_{:j})\neq 0)\mid S\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}\mathbb{E}_{P_E}\left[\mathbb{1}(\mathrm{Var}(\mathbf{Z}^E\mathbf{L}_{:j})\neq 0)\mid S\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}P_{E|S}\left[\mathrm{Var}(\mathbf{Z}^E\mathbf{L}_{:j})\neq 0\right]\\ &= \mathbb{E}_{p(S)}\sum_{j=1}^{d}P_{E|S}\left[\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})\neq 0\right].\end{aligned}$$

Let N_j denote the nonzero entries of $\mathbf{L}$ in column j, i.e., $N_j := \{i\in[d]\mid \mathbf{L}_{i,j}\neq 0\}$. Now, we take the intersection of S and N_j, $S\cap N_j$, and consider two possible cases.

First, assume $S\cap N_j=\varnothing$. Then, we can see that $\mathbf{L}_{S,j}=\mathbf{0}$ and it follows that

$$P_{E|S}\left[\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})=0\right]=1.$$

For the second case, consider $S\cap N_j\neq\varnothing$. Now, we can see that $\mathbf{L}_{S,j}\neq\mathbf{0}$. By Lemma 3 we have that

$$\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})\neq 0,$$

and can therefore directly conclude that

$$P_{E|S}\left[\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})=0\right]=0.$$

Taking both considered cases and recalling that

$$P_{E|S}\left[\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})\neq 0\right]=1-P_{E|S}\left[\mathrm{Var}(\mathbf{Z}_S^E\mathbf{L}_{S,j})=0\right],$$

we can write

$$P_{E|S}\big[\mathrm{Var}(\mathbf{Z}_S^E \mathbf{L}_{S,j}) \neq 0\big] = 1 - \mathbb{1}(S \cap N_j = \varnothing) \\ = \mathbb{1}(S \cap N_j \neq \varnothing),$$

and consequently

$$\mathbb{E}\|\hat{\mathbf{Z}}^E \mathbf{L}\|_{\mathrm{Var}} = \mathbb{E}_{p(S)} \sum_{j=1}^{d} \mathbb{1}(S \cap N_j \neq \varnothing).$$

Recalling our original constraint, we can now write

$$\mathbb{E}_{p(S)} \sum_{j=1}^{d} \mathbb{1}(S \cap N_j \neq \varnothing) \leq \mathbb{E}_{p(S)} \sum_{j=1}^{d} \mathbb{1}(j \in S). \tag{9}$$

Since $\mathbf{L}$ is an invertible matrix, by Lemma 4, we know that there exists a permutation $\sigma : [d] \to [d]$, such that $\mathbf{L}_{j,\sigma(j)} \neq 0, \forall j \in [d]$. Since the sum over all dimensions d on the LHS of Eq. (9) is invariant under σ, we can apply this permutation to the LHS and by collecting terms write

$$\mathbb{E}_{p(S)} \sum_{j=1}^{d} \mathbb{1}(S \cap N_{\sigma(j)} \neq \varnothing) - \mathbb{1}(j \in S) \leq 0. \tag{10}$$

Now, notice $\forall j \in [d]$

$$\mathbb{1}(S \cap N_{\sigma(j)} \neq \varnothing) - \mathbb{1}(j \in S) \geq 0,$$

which holds since whenever $j \in S$ it also holds that $j \in S \cap N_{\sigma(j)}$, since σ by definition permutes the columns of $\mathbf{L}$ such that $\forall j$ there is a nonzero entry $\mathbf{L}_{j,\sigma(j)}$, and thus $j \in N_{\sigma(j)}$.

The overall expression in Eq. (10) is non-positive, while all elements of the sum are non-negative, from which we conclude that all elements of the sum must be equal to zero. We can thus write

$$\forall S \in \mathcal{S}, \forall j \in [d], \mathbb{1}(S \cap N_{\sigma(j)} \neq \varnothing) = \mathbb{1}(j \in S).$$

Consequently,

$$\forall S \in \mathcal{S}, \forall j \in [d], j \notin S \implies S \cap N_{\sigma(j)} = \varnothing.$$

Since $S \cap N_{\sigma(j)} = \varnothing \iff N_{\sigma(j)} \subseteq S^c$, where S^c denotes the complement of S, we can write

$$\forall S \in \mathcal{S}, \forall j \in [d], j \notin S \implies N_{\sigma(j)} \subseteq S^c \tag{11}$$

$$\forall j \in [d], N_{\sigma(j)} \subseteq \bigcap_{S \in \mathcal{S} | j \notin S} S^c. \tag{12}$$

Recall Assumption 4, which states that

$$\bigcup_{S \in \mathcal{S} | j \notin S} S = [d] \setminus \{j\}. \tag{13}$$

If we take the complement of both sides of Eq. (13) and apply De Morgan's law, we get

$$\bigcap_{S \in \mathcal{S} | j \notin S} S^c = \{j\},$$

which we insert into Eq. (12) and finally see that $N_{\sigma(j)} = \{j\}$. From this we conclude that $\mathbf{L} = \mathbf{DP}$, where $\mathbf{D}$ is a diagonal invertible matrix and $\mathbf{P}$ is a permutation matrix.

Now, consider the case where $m > d$. Let $\mathbf{L}'$ be an invertible $d \times d$ submatrix of $\mathbf{L}$. Note that if $\mathbf{L}'$ is a submatrix of $\mathbf{L}$, the original constraint in the theorem implies that $\mathbb{E}\|\mathbf{Z}^E\mathbf{L}'\|_{\text{Var}} \leq \mathbb{E}\|\mathbf{Z}^E\|_{\text{Var}}$.

For any column j of $\mathbf{L}$ with nonzero entries, there exists a $d \times d$ invertible submatrix $\mathbf{L}'(j)$ that contains the column j and whose columns are all linearly independent. This holds by the following argument: there is a submatrix $\mathbf{L}''$ of $\mathbf{L}$ with d independent columns, since $\mathbf{L}$ has rank d. If j is not already a column of $\mathbf{L}''$, by the exchange theorem from linear algebra we can replace some column of $\mathbf{L}''$ with j to obtain the submatrix $\mathbf{L}'(j)$ with the desired properties.

Now, for each such $\mathbf{L}'(j)$ we can apply the argument of the proof for the case $m = d$ and deduce that any (nonzero) column j of $\mathbf{L}$ consists of a single nonzero entry and thus each $\hat{Z}_j$ must be zero, or a scalar multiple of a latent Z_i. ■

The last step of our proof for the case $m = d$ underlines why we make Assumption 4: after applying De Morgan's law, this assumption leaves us with a set that contains a single element, which in turn directly allows us to conclude that the mixing matrix $\mathbf{L}$ consists of a diagonal matrix $\mathbf{D}$ and a permutation $\mathbf{P}$. Just as Lachapelle et al. (2023) state, if we did not make Assumption 4 and this final step resulted not in a singleton, but a set with multiple elements, this would not allow us to directly conclude that $\mathbf{L} = \mathbf{DP}$. If this were the case, we conjecture that we would only be able to achieve partial disentanglement of blocks of variables (Lachapelle and Lacoste-Julien, 2022).

Appendix B. Experimental Details

In this section, we present additional details to reproduce all of our experimental results. We begin by outlining the synthetic data generating process and what exactly is subject to randomization across runs, continue with implementation and training details of our proposed method, provide details on the nonlinear SCMs we use in our experiments, introduce the metric we use to quantify our results and finally present additional experimental results not reported in the main text.

B.1. Synthetic Data Generation

Across all experimental settings, we use a train/test split of 75/25, meaning if we have $n = 100000$ samples per environment, we use $n_{\text{train}} = 75000$ samples for training and $n_{\text{test}} = 25000$ samples to evaluate our metric.

For each random seed, we resample the underlying SCM, resample n samples from the resulting SCM and reinitialize our algorithm with new initial weights for $\hat{\mathbf{L}}$. The ground truth matrix $\mathbf{L}$ is sampled once and kept fixed across random seeds. We show that our results do not rely on a cherry picked mixing matrix by reporting additional runs over random seeds for randomly resampled $\mathbf{L}$ (cf. Appendix B.5).

B.2. Implementation and Training Details

We use `pytorch` (Paszke et al., 2019) to implement our algorithm in practice. For optimization, we use the `adamW` algorithm (Loshchilov and Hutter, 2018) with learning rate $lr = 2 * 10^{-3}$ and otherwise default parameters provided by `pytorch`. For all experiments, we use a batch size of 4096 and train for 50 epochs.

We refrain from performing hyperparameter optimization in this work. All hyperparameters are chosen to roughly result in equal magnitudes of each respective loss term and shared across all experiments. The values of the hyperparameters are chosen as

$$\begin{aligned}\lambda_e &= 1,\\ \lambda_m &= 1,\\ \lambda_{\text{diag}} &= 10,\\ \lambda_{\text{norm}} &= 5.\end{aligned}$$

For the implementation of ICA, we use the FastICA algorithm (Hyvärinen and Oja, 2000).

B.3. Nonlinear SCMs

We sample the adjacency matrix of the graph of both nonlinear SCMs according to the procedure detailed in Section 4.2 with $p = 0.75$. The resulting graph that is shared by both models is shown in Fig. 3.

The structural equations of SCM 1 are:

$$\begin{aligned}Z_1 &:= \eta_1,\\ Z_2 &:= Z_1^2 + \eta_2,\\ Z_3 &:= Z_1^2 + Z_2^2 + \eta_3,\\ Z_4 &:= Z_1^2 + Z_2^2 + Z_3^2 + \eta_4,\\ Z_5 &:= Z_1^2 + Z_3^2 + Z_4^2 + \eta_5,\\ Z_6 &:= Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2 + \eta_6,\end{aligned}$$

and those of SCM 2 are defined as:

$$\begin{aligned}Z_1 &:= \eta_1,\\ Z_2 &:= \sin(Z_1) + \eta_2,\\ Z_3 &:= \sqrt{Z_1 + Z_2} + \eta_3,\\ Z_4 &:= \log(Z_1^2 + Z_2) + Z_3^2 + \eta_4,\\ Z_5 &:= Z_3 \cos(Z_1) + \arctan(Z_4) + \eta_5,\\ Z_6 &:= Z_2 Z_3 e^{\frac{Z_4^2}{Z_5}} + \eta_6.\end{aligned}$$

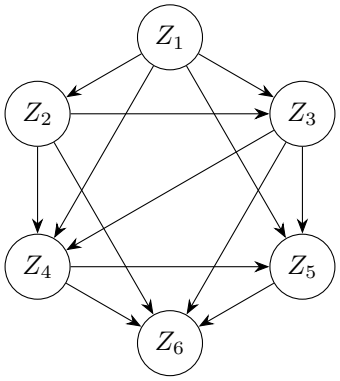


Figure 3: Causal graph of SCM 1 and SCM 2.

B.4. Metric

Here, we present the formal definition of our used metric, the mean correlation coeffiecient (MCC) (Hyvärinen and Morioka, 2016; Khemakhem et al., 2020).

Let $\mathbf{C}$ be the Pearson correlation matrix between the ground truth variables $\mathbf{Z}$ and the learned variables $\hat{\mathbf{Z}}$. Then, the MCC score is defined as

$$\text{MCC} := \max_{\pi \in \mathfrak{S}_d} \frac{1}{d} \sum_{j=1}^{d} |\mathbf{C}_{j,\pi(j)}|,$$

where $\mathfrak{S}_d$ is the set of d-permutations and $|\cdot|$ denotes the absolute value.

B.5. Additional Experiments

Here, we present repetitions of the experiment in Section 4.3 where we vary the number of ground truth variables d under additional random samples of the mixing matrix $\mathbf{L}$. The results are reported in Fig. 4.

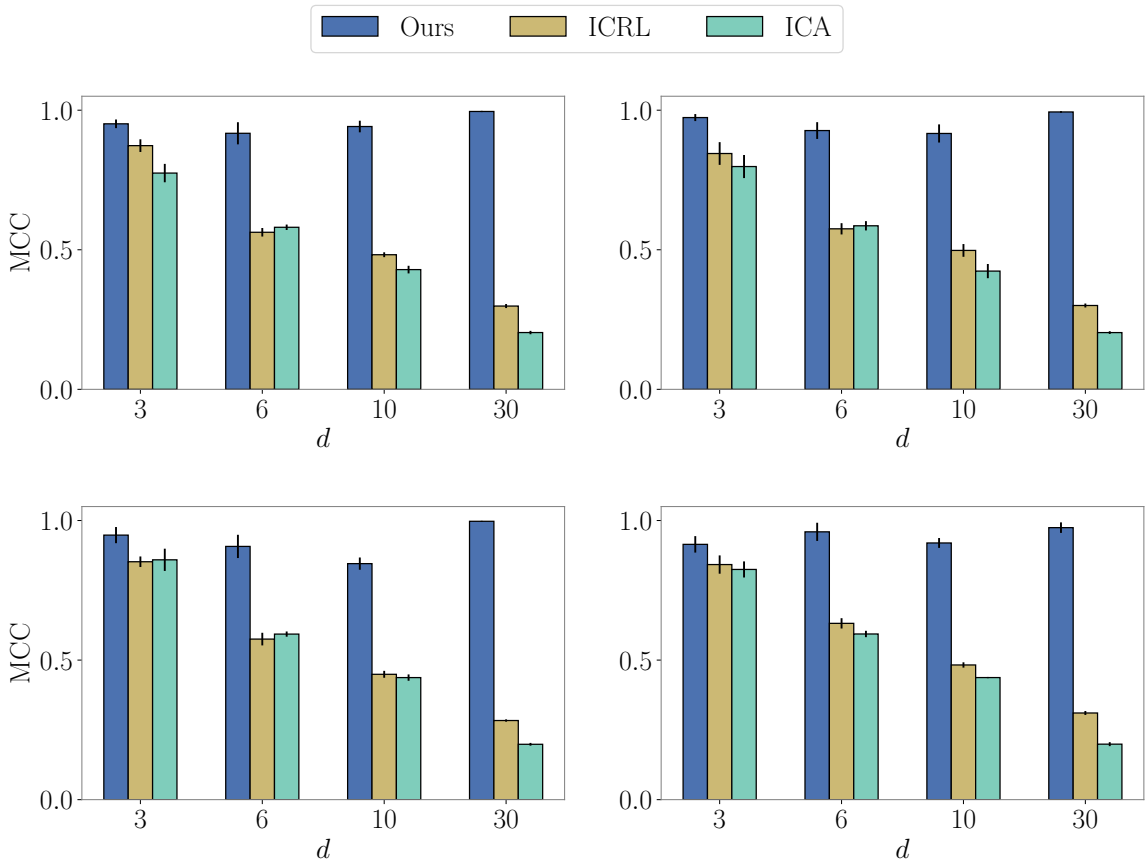


Figure 4: Experiments for changing number of variables d for additional mixing matrices $\mathbf{L}$. Each subfigure corresponds to a different $\mathbf{L}$. We report the mean MCC score for five random seeds with error bars indicating the standard error.