A Sparsity Principle for Partially Observable Causal Representation Learning

Danru Xu [1] **Dingling Yao** [2 3] **Sébastien Lachapelle** [4 5] **Perouz Taslakian** [6] **Julius von Kügelgen** [7]
Francesco Locatello [2] **Sara Magliacane** [1]

Abstract

Causal representation learning aims at identifying high-level causal variables from perceptual data. Most methods assume that *all* latent causal variables are captured in the high-dimensional observations. We instead consider a *partially observed* setting, in which each measurement only provides information about a subset of the underlying causal state. Prior work has studied this setting with multiple domains or views, each depending on a *fixed* subset of latents. Here we focus on learning from *unpaired* observations from a dataset with an *instance-dependent* partial observability pattern. Our main contribution is to establish two identifiability results for this setting: one for linear mixing functions without parametric assumptions on the underlying causal model, and one for piecewise linear mixing functions with Gaussian latent causal variables. Based on these insights, we propose two methods for estimating the underlying causal variables by enforcing sparsity in the inferred representation. Experiments on different simulated datasets and established benchmarks highlight the effectiveness of our approach in recovering the ground-truth latents.

1. Introduction

Endowing machine learning models with causal reasoning capabilities is a promising direction for improving their robustness, generalization, and interpretability (Spirtes et al., 2000; Pearl, 2009; Peters et al., 2017). Traditional causal inference methods assume that the causal variables are given a priori, but in many real-world settings, we only have unstructured, high-dimensional observations of a causal system. Motivated by this shortcoming, causal representation learning (CRL; Schölkopf et al., 2021) aims to infer high-level causal variables from low-level data such as images.

A popular approach to identify (i.e., provably recover) high-level latent variables is (nonlinear) independent component analysis (ICA) (Hyvarinen and Morioka, 2016; 2017; Hyvarinen et al., 2019; Khemakhem et al., 2020), which aims to recover independent latent factors from entangled measurements. Several works generalize this setting to the case in which the latent variables can have causal relations (Yao et al., 2022; Brehmer et al., 2022; Lippe et al., 2022; 2023; Ahuja et al., 2023a;b; Lachapelle et al., 2022; 2023; 2024; von Kügelgen et al., 2021; 2023; Wendong et al., 2023; Squires et al., 2023; Buchholz et al., 2023; Zhang et al., 2023), establishing various identifiability results under different assumptions on the available data and the generative process. However, most existing works assume that *all* causal variables are captured in the high-dimensional observations. Notable exceptions include Sturma et al. (2023) and Yao et al. (2023) who study *partially observed* settings with multiple domains (datasets) or views (tuples of observations), respectively, each depending on a *fixed subset* of the latent variables.

In this work, we also focus on learning causal representations in such a *partially observed* setting, where not necessarily all causal variables are captured in any given observation. Our setting differs from prior work in two key aspects: (i) we consider learning from a dataset of *unpaired partial* observations; and (ii) we allow for *instance-dependent* partial observability patterns, meaning that each measurement depends on an unknown, varying (rather than fixed) subset of the underlying causal state.

This setting is motivated by real-world applications in which we cannot at all times observe the complete state of the environment, e.g., because some objects are moving in and out of frame, or are occluded. As a motivating example, consider a stationary camera that takes pictures of a parking lot on different days as shown in Fig. 1a. On different days, different cars are present in the parking lot, and the same car can be parked in different spots. Our task is to recover the position for each car that is present in a certain image.

---

[1]University of Amsterdam [2]Institute of Science and Technology Austria [3]Max Planck Institute for Intelligent Systems, Tübingen, Germany [4]Samsung - SAIT AI Lab, Montreal [5]Mila, Université de Montréal [6]ServiceNow Research [7]Seminar for Statistics, ETH Zürich. Correspondence to: Danru Xu <d.xu3@uva.nl>.
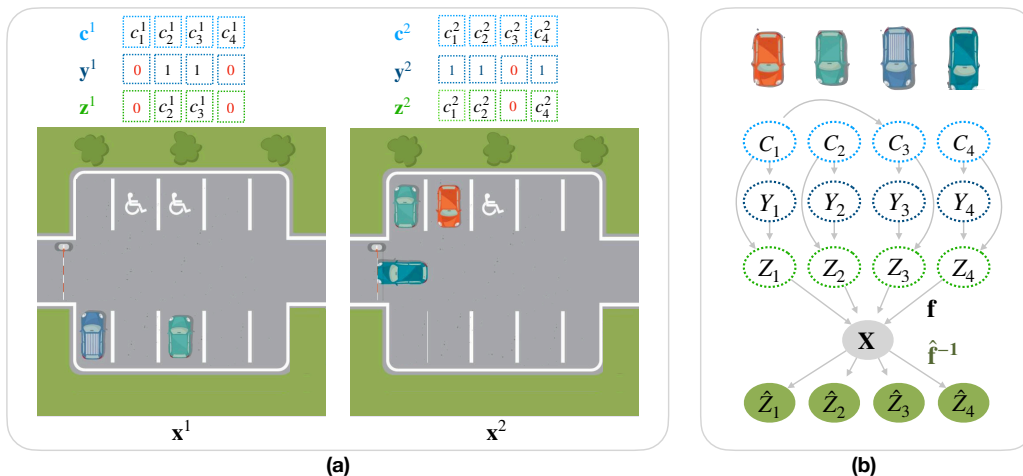
Figure 1: **(a)** Motivating example for the *Unpaired Partial Observation* setting: a stationary camera taking pictures of a car park. We consider $\mathbf{x}^1$ the image on day 1 and $\mathbf{x}^2$ the image on day 2. The latent causal variables $\mathbf{c}^1$ and $\mathbf{c}^2$ represent the positions of four cars on each day. In $\mathbf{x}^1$ only $Car2$ and $Car3$ are visible, while in $\mathbf{x}^2$ all cars except $Car3$ are visible. This is represented by the ones in the binary mask variables $\mathbf{y}^1$ and $\mathbf{y}^2$. The combination of the values of the latent causal variables $\mathbf{c}$ and the masks $\mathbf{y}$ are the *masked causal variables* $\mathbf{z}$, which used by the mixing function $\mathbf{f}$ to generate the images $\mathbf{x}$. **(b)** Causal model of the setting, the dotted line variables are not directly observed, but they are measured only through the observation $\mathbf{X}$. Our goal is to learn a representation $\hat{\mathbf{Z}}$ that identifies $\mathbf{Z}$ up to permutation and element-wise transformation.

In this setting, we only have one observation for a given state of the system (i.e., one image per day), and the subsets of causal variables that are measured in the observation (the parked cars), change dynamically across images. We highlight the following contributions:

- We formalize the *Unpaired Partial Observations* setting for CRL, where each *partial observation* captures only a subset of causal variables and the observations are unpaired, i.e., we do not have simultaneous partial observations of the same state of the system.

- We introduce two theoretical results for identifying causal variables up to permutation and element-wise transformation under partial observability. Both results leverage a sparsity constraint. In particular, Thm. 3.1 proves identifiability for linear mixing function and without parametric assumptions on the underlying causal model. Thm. 3.4 proves identifiability for the *piecewise linear mixing function*, when the causal variables are Gaussian and we can group observations by their partial observability patterns.

- Finally, we propose two methods that implement these theoretical results and validate their effectiveness with experiments on simulated data and image benchmarks, e.g. Causal3DIdent (von Kügelgen et al., 2021), that we modify to test our partial observability setting.

2. Problem setting

In this section, we formalize the *Unpaired Partial Observation* setting, in which we have a set of high-dimensional observations that are functions only of instance-dependent subsets of the true underlying causal variables. This setting consists of four sets of random variables: the *causal variables* $\mathbf{C}$, the binary *mask variables* $\mathbf{Y}$ that represent if a variable is measured in a sample, the *masked causal variables* $\mathbf{Z}$ that combine the information from the causal variables and the masks, and the *observations* $\mathbf{X}$. Our goal is to recover the masked causal variables $\mathbf{Z}$ up to permutation and element-wise transformation, solely from the observations $\mathbf{X}$, despite the instance-dependent partial observability pattern. We show an example of a causal graph of the setting in Fig 1b and discuss each component in the following.

Causal variables $\mathbf{C}$. We define our latent causal variables as a random vector $\mathbf{C} = (C_1, ..., C_n)$ that takes values in $\mathcal{C} = \mathcal{C}_1 \times ... \times \mathcal{C}_n \subseteq \mathbb{R}^n$, which is an open, simply connected latent space. The causal variables follow a distribution with density $p(\mathbf{C})$, which allows for causal relations between them. We assume that $p(\mathbf{c}) \neq 0$ for all $\mathbf{c} \in \mathcal{C}$.

Mask variables $\mathbf{Y}$. We use a binary mask random variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$ with domain $\mathcal{Y} \subseteq \{0,1\}^n$ to represent the dynamic partial observability patterns, i.e., the causal variables that are measured in each of the samples. If $Y_i = 1$ then we consider the variable C_i *measured*, i.e. captured in the observation, otherwise it is considered *unmeasured*. We assume $\mathbf{Y}$ follows $p(\mathbf{Y})$. Further, we define the *support index* random set $\mathbf{S}$ as the index of non-zero components of $\mathbf{Y}$, i.e., $\mathbf{S} := \{i \in [n] : Y_i \neq 0\}$. The support index set has a probability mass function $p(\mathbf{s})$ and

support $\mathcal{S}$ defined as:

$$\mathbb{P}(\mathbf{s}) := \mathbb{P}\left(\bigwedge_{j \in \mathbf{s}}(\mathbf{Y}_j = 1) \wedge \bigwedge_{j \notin \mathbf{s}}(\mathbf{Y}_j = 0)\right)$$

$$\text{and } \mathcal{S} := \{\mathbf{s} \subseteq [n] \mid p(\mathbf{s}) > 0\}.$$

This definition allows us to model also dependences between the components of $\mathbf{Y}$ and masking behavior that depends on the values of the underlying causal variables.

Masked causal variables Z. The *masked causal variables* $\mathbf{Z} = (Z_1, ..., Z_n) \in \mathcal{Z}$ are a combination of the causal variables and the masks, and they are the latent inputs to the mixing function $\mathbf{f}$ that we are trying to recover. In particular, they are the Hadamard product of the causal variables with the binary mask variable, i.e., $\mathbf{Z} = \mathbf{Y} \odot \mathbf{C}$. This means that for sample $i \in [N]$ and any causal variable $j \in [n]$, if the mask value y_j^i is 1, then the causal variable c_j^i is measured and z_j^i is c_j^i. Instead, if $y_j^i = 0$, then the causal variable is unmeasured and z_j^i takes a fixed masked value M_j, which we will consider for simplicity to be 0. Note that this is not equivalent to do-interventions, since masking variables does not influence any downstream variables, as an intervention would, as explained in an example in App. A. Finally, we assume that for all $\mathbf{s} \in \mathcal{S}$, the probability measure $\mathbb{P}_{\mathbf{Z_s}|\mathbf{S}=\mathbf{s}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{|\mathbf{s}|}$.

Observations X. We assume that observations $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ are generated by mixing the masked causal variables $\mathbf{Z}$ with the same *mixing function* $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$, i.e., $\mathbf{X} = \mathbf{f}(\mathbf{Z})$. We refer to partitions of observations with the same unknown *partial observability pattern*, i.e., with the same unknown value of $\mathbf{Y} = \mathbf{y}$, as *groups*, and we assume that each observation $\mathbf{x}^i$ for $i \in [N]$ is part of a group $g^i \in \mathcal{G}$.

Our goal is to identify the masked causal variables $\mathbf{Z}$ from a set of observations $\mathbf{X}$. In CRL we usually cannot recover the exact value of the latent variables, but we can only identify them up to some transformation. Our results guarantee that each ground truth variable is represented by a single estimated variable up to a linear transformation. Similar notions of identifiability were used in previous works (Comon, 1994; Khemakhem et al., 2020; Lachapelle et al., 2023).

Definition 2.1. The ground truth representation vector $\mathbf{Z}$ is said to be identified *up to permutation and element-wise linear transformation* by a learned representation vector $\hat{\mathbf{Z}}$ when there exists a permutation matrix $\mathbf{P}$ and an invertible diagonal matrix $\mathbf{D}$ such that $\hat{\mathbf{Z}} = \mathbf{PDZ}$ almost surely.

To prove our results, we describe the sufficient support index variability, which was originally defined by Lachapelle et al. (2023) for sparse multitask learning for disentanglement.

Assumption 2.2. (Sufficient support index variability (Lachapelle et al., 2023)) For all $i \in [n]$, we assume that the

union of the support indices $\mathbf{s}$ that do not contain i covers all of the other causal variables, or more formally:

$$\bigcup_{\mathbf{s} \in \mathcal{S} \mid i \notin \mathbf{s}} \mathbf{s} = [n] \setminus \{i\}.$$

This assumption avoids cases in which two variables are always missing at the same time, since then we would not be able to disentangle them from the observations. A trivial set $\mathcal{S}$ that satisfies this assumption is $\mathcal{S} = 1, 2, ..., n$, which contains n distinct masks. We conjecture that if this assumption is not satisfied for all i, but only for blocks of variables, we would instead get block identifiability.

3. Identifiability via a Sparsity Principle

In this section, we show how a simple sparsity constraint on the learned representations allows us to learn a disentangled representation of the ground truth variables. We first consider linear mixing functions and prove identifiability without any parametric assumption on the causal variables and while allowing for partial observability patterns that can depend on the value of the causal variables (Thm. 3.1). We then investigate if this *sparsity principle* can also identify variables for nonlinear mixing functions $\mathbf{f}$. This is not the case in general, as we show in Example 3.1.

Our linear result hinges on the existence of a function $\mathbf{g}$ such that the composition of $\mathbf{g}$ and $\mathbf{f}$ is affine. Based on this intuition, we consider *piecewise linear* mixing functions $\mathbf{f}$, since for Gaussian causal variables they can be composed with an appropriate $\mathbf{g}$ to obtain an affine function. In this setting, we prove that we can learn a disentangled representation of the latent variables (Def. 2.1) given that the masks are independent of the causal variables and given that we know the groups of the observations (Thm. 3.4).

3.1. Linear Mixing Function

We show that for linear mixing functions under a perfect reconstruction (Eq. 1), a simple *sparsity constraint* on the learned representation (Eq. 2) allows us to learn a disentangled representation of the ground truth latent variables.

Theorem 3.1 (Element-wise Identifiability for Linear $\mathbf{f}$). *Assume the observation $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ follows the data-generating process in Sec. 2, where $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is an injective linear function, and Ass. 2.2 holds. Let $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ be an invertible linear function onto its image and let $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible continuous function onto its image. If both of the following conditions hold,*

$$\mathbb{E}\left\|\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2 = 0, \quad and \tag{1}$$

$$\mathbb{E}\left\|\mathbf{g}(\mathbf{X})\right\|_0 \leq \mathbb{E}\left\|\mathbf{Z}\right\|_0, \tag{2}$$

then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$ up to a permutation and element-wise linear transformations (Def. 2.1), i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation composed with element-wise invertible linear transformations on $\mathcal{Z}$.

We provide a proof in App. B.1 and now give an intuitive explanation for why it holds. The zero reconstruction loss ensures that no information is lost in the encoding $\mathbf{g}(\mathbf{X})$, which implies that $\mathbf{g}(\mathbf{X})$ is not sparser than $\mathbf{Z}$. Hence, incorporating Eq. (2) as a constraint or regularization term in our methods enables our estimators to match the sparsity of the ground truth variables, which breaks indeterminacies due to rotations of the latent space.

The idea of using a sparsity constraint or regularization is similar to previous work (Lachapelle et al., 2023) in the context of sparse multitask learning. In this paper we leverage these ideas in the distinct setting of partial observability. This result requires that the mixing function $\mathbf{f}$, is injective, but not necessarily bijective. Notably, it does not require any parametric assumptions on the distribution of $\mathbf{Z}$, thus allowing for causal relations or other statistical dependencies among the latent variables. Finally, this result also allows for mask variables that potentially depend on the values of the latent causal variables.

3.2. Is sparsity enough for identification for nonlinear f?

Since linearity of $\mathbf{f}$ is a strong assumption that may not hold in many applications, an obvious question is whether we can extend this result to nonlinear mixing functions. Unfortunately, this is not the case without making further assumptions, as demonstrated by the following example.

Example 3.1. Consider $\mathbf{C} \sim \mathcal{N}(0, \mathbf{I}_2)$, where $\mathbf{I}_2$ is the identity matrix. Assume an independent mask $\mathbf{Y}$ with distribution $p(\mathbf{Y} = \mathbf{y}) = 1/4$ for any $\mathbf{y} \in \{0,1\}^2$, satisfying Ass. 2.2. Let the nonlinear mixing function $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ be

$$\mathbf{f}(\mathbf{z}) := \sinh(\mathbf{R}_{\frac{\pi}{4}} \mathbf{z}) + \sinh(\mathbf{R}_{-\frac{\pi}{4}} \mathbf{z}), \qquad (3)$$

where $\mathbf{R}_\theta$ is a rotation matrix. Consider $\hat{\mathbf{f}}$ and $\mathbf{g}$ to be the identity function, which trivially satisfy Eq. 1, since $\hat{\mathbf{f}} \circ \mathbf{g}(\mathbf{X}) = \mathbf{X}$. We show in Appendix C that $\mathbf{g}$ satisfies the sparsity constraint (Eq. 2). However, despite satisfying all requirements of Thm 3.1 except for the linearity of $\mathbf{f}$, we can show that each component of $\hat{\mathbf{f}}^{-1} \circ \mathbf{f} = \mathbf{f}$ depends on both components of $\mathbf{z}$; or in other words, the learned representation does not identify $\mathbf{Z}$ up to permutation and element-wise transformations. We refer to App. C for details.

3.3. Piecewise Linear Mixing Function

In light of Example 3.1, we consider the role of linearity in Thm. 3.1. We notice that it is a sufficient condition for ensuring that there exists a $\mathbf{g}$ such that $\mathbf{g} \circ \mathbf{f}$ is affine on $\mathcal{Z}$. We then consider the question: even if $\mathbf{f}$ is not affine itself,

can we consider a restricted class of $\mathbf{f}$ and latent variables $\mathbf{Z}$ such that the composition of $\mathbf{f}$ and an appropriate $\mathbf{g}$ is affine? As a first step we consider a piecewise linear $\mathbf{f}$ and assume that the causal variables are Gaussian and that the masks are independent from the causal variables.

Assumption 3.2. We assume $\mathbf{C}$ follows a non-degenerate *multivariate normal distribution*, i.e. $\mathbf{C} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is a positive definite matrix.

Assumption 3.3. We assume $\mathbf{C}$ and $\mathbf{Y}$ are independent from each other, i.e. the partial observability pattern does not depend on the values of the causal variables.

These assumptions represent a classical linear Gaussian Structural Causal Model setting for the causal variables and a missing-at-random assumption in terms of which variables are measured in the observation. On the other hand, in our setting, we do not directly observe the causal variables or the masks, but we only observe them mixed in an observation.

Under these assumptions, the conditional distribution of the masked causal variables $\mathbf{Z}$ given the binary mask vector $\mathbf{Y}$ is defined as a multivariate normal distribution:

$$\mathbf{Z} \mid \mathbf{Y} \sim N(\boldsymbol{\mu}_{\mathbf{Y}}, \boldsymbol{\Sigma}_{\mathbf{Y}})$$
$$\text{where } \boldsymbol{\mu}_{\mathbf{Y}} = (\mu_1 Y_1, \ldots, \mu_n Y_n), \quad \boldsymbol{\Sigma}_{\mathbf{Y}(ij)} = \boldsymbol{\Sigma}_{ij} Y_i Y_j$$

This distribution is a degenerate multivariate normal (De-MVN), i.e., a normal with a singular covariance matrix, if at least one of the causal variables is masked by $\mathbf{Y}$.

Intuitively, we can leverage the Gaussianity of $\mathbf{Z}|\mathbf{Y}$ to enforce that the reconstructed $\hat{\mathbf{Z}}|\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is also Gaussian. We now show that this allows us to identify the latent factors $\mathbf{Z}$ through our *sparsity constraint* on the learned representations (Eq. 5), given that we are able to partition the data according to the unknown value of $\mathbf{Y} = \mathbf{y}$, or in other words, given that we know the group g_i for each observation $\mathbf{x}_i$ for $i \in [N]$. The rationale of this requirement is that we do not need to know the value of the latent mask $\mathbf{Y}$, but we do need to be able to separate observations that are generated by different distributions of $\mathbf{Z}$, so we can effectively enforce the Gaussianity constraint (Eq. 6).

Theorem 3.4 (Element-wise Identifiability for Piecewise Linear **f**). *Assume the observation $\mathbf{X}$ follows the data-generating process in Sec 2, Ass. 2.2, 3.2 and 3.3 hold and $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is an injective continuous piecewise linear function. Let $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ be a continuous invertible piecewise linear function and let $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be a continuous invertible piecewise linear function onto its image. If all following conditions hold:*

$$\mathbb{E} \left\| \mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X})) \right\|_2^2 = 0, \qquad (4)$$

$$\mathbb{E} \left\| \mathbf{g}(\mathbf{X}) \right\|_0 \le \mathbb{E} \left\| \mathbf{Z} \right\|_0 \text{ and} \qquad (5)$$

$$\mathbf{g}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}}) \qquad \forall \mathbf{y} \in \mathcal{Y}, \qquad (6)$$

for some $\boldsymbol{\mu_y} \in \mathbb{R}^n, \boldsymbol{\Sigma_y} \in \mathbb{R}^{n \times n}$, then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$, i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation composed with element-wise invertible linear transformations (Def. 2.1).

We provide the complete proof in App. B.2. We first provide some results for a weaker notion of identifiability: *identifiability up to affine transformations* (Def. B.4). To this end, we first extend a theorem by Kivva et al. (2022) from the case of non-degenerate to potentially *degenerate* multivariate normal variables (Thm. B.5). Such variables are crucial in our setting because partial observability potentially introduces degenerate cases. The crux of our proof involves handling the case of degenerate variables that do not have probability density. We then show that given the information of the binary mask $\mathbf{Y}$, we can identify the latent factors $\mathbf{Z}$ up to an affine function $\mathbf{h_Y}$ (Lemma B.10). We then show that all of these affine functions can be represented by a single affine function $\mathbf{v} := \hat{\mathbf{f}}^{-1}(\mathbf{f}(\mathbf{Z}))$ defined on $\mathcal{Z}$.

Compared to linear case in Thm. 3.1, the additional constraint in this case is *Gaussianity* on both $\mathbf{Z}|\mathbf{Y}$ and the estimator $\mathbf{g}(\mathbf{X})|\mathbf{Y}$. This ensures that the composition of two piecewise linear functions $\mathbf{g}$ and $\mathbf{f}$ remains affine on $\mathcal{Z}$, extending the results from linear $\mathbf{f}$ to piecewise linear $\mathbf{f}$.

Non-zero mask values. Our theoretical analysis assumes that we mask the unmeasured latent variables with a mask value of 0, i.e. when $\mathbf{y}^i = 0$ for $i \in [N]$, we set $\mathbf{z}_i = 0$. One can wonder whether allowing to set $\mathbf{z}^i = \mathbf{M}$ when $\mathbf{y}_i = 0$ for some potentially nonzero constant vector $\mathbf{M}$ would make the model more expressive. It turns out that this is not the case, since the decoder $\mathbf{f}$ can always shift $\mathbf{Z}$ in arbitrary ways, making the specific value of $\mathbf{M}$ irrelevant.

4. Implementation

We implement our two theoretical results asconstrained optimization problems in Cooper (Gallego-Posada and Ramirez, 2022). We approximate the sparsity constraint, i.e. Eq. 2 in Thm. 3.1 and Eq. 5 in Thm. 3.4, by replacing the L_0 norm with L_1 norm, which is differentiable except at zero. In practice, the L_1 norm of the ground truth variables $\mathbf{Z}$ is unknown, so we instead set a hyperparameter ϵ for the sparsity constraint $\mathbb{E} \|\mathbf{g}(\mathbf{X})\|_1 \le \epsilon$. In our experiments, we use $\epsilon = 0.01$ or 0.001, details are provided in App. D.2.

For linear $\mathbf{f}$, we can reconstruct the latent variables directly from a dataset of observations $\{\mathbf{x}^i\}_{i \in [N]}$ by minimizing the reconstruction error and adding the sparsity constraint (Thm 3.1). For piecewise linear $\mathbf{f}$, Thm. 3.4 requires that we know how to partition the data with the same partial observability *at training time*, i.e., we have information about the group of each observation. At test time, we have already learned $\mathbf{g}$, so we can instead use only the observations without the group information. To encourage the Gaussianity

condition on $\mathbf{g}(\mathbf{X})|\mathbf{Y} = \mathbf{y}$ (Eq. 6) in Thm. 3.4, we add two regularization terms that push the estimated skewness of each latent variable in each group to be 0 and the estimated kurtosis to be 3, which are the values of these moments in the Gaussian distribution. We learn the encoder $\mathbf{g}_\psi$ and the decoder $\hat{\mathbf{f}}_\theta$ by solving the following optimization problem:

$$\min_{(\theta, \psi)} \frac{1}{N} \sum_{i \in [N]} \left\| \mathbf{x}^i - \hat{\mathbf{f}}_\theta(\mathbf{g}_\psi(\mathbf{x}^i)) \right\|_2^2$$
$$+ \sum_{g \in \mathcal{G}} \left(\left| \widehat{\text{skew}}_g(\mathbf{g}_\psi(\mathbf{x})) \right| + \left| \widehat{\text{kurt}}_g(\mathbf{g}_\psi(\mathbf{x})) - 3 \right| \right)$$
$$\text{subject to:} \quad \frac{1}{nN} \sum_{i \in [N]} \left\| \mathbf{g}_\psi(\mathbf{x}^i) \right\|_1 \le \epsilon, \quad (7)$$

where $\widehat{\text{skew}}_g(\mathbf{g}_\psi(\mathbf{x}))$ and $\widehat{\text{kurt}}_g(\mathbf{g}_\psi(\mathbf{x}))$ are the estimated skewness and estimated kurtosis for group $g \in \mathcal{G}$. We solve both problems with the extra-gradient version of Adam (Gidel et al., 2018). We provide all details in App. D.

5. Experimental Results

We perform three sets of experiments, one with numerical data in Sec. 5.1 and two with image datasets, a dataset with multiple balls in Sec. 5.2 and *PartialCausal3DIdent*, a partially observable version of Causal3DIdent (von Kügelgen et al., 2021) in Sec. 5.3. We provide all the code for our method and the experiments at https://github.com/danrux/sparsity-crl.

Following previous work (Hyvarinen et al., 2019; Khemakhem et al., 2020), we report the *mean coefficient of determination (MCC)* to assess that the learned representations match the ground truth up to a permutation and element-wise linear transformations. This metric is based on the Pearson correlation matrix $\text{Corr}^{n \times n}$ between the learned representations $\hat{\mathbf{Z}} = \mathbf{g}(\mathbf{X})$ and ground truth masked causal variables $\mathbf{Z}$. Since our results are up to permutation, we compute the MCC on the permutation π that maximizes the average of $|\text{Corr}_{i, \pi(i)}|$ for each index of a ground truth variable $i \in [n]$, i.e. $\text{MCC} = \frac{1}{n} \max_{\pi \in perm([n])} \sum_{i=1}^{n} |\text{Corr}_{i, \pi(i)}|$. We denote the correlation matrix with the permutation π as $\text{Corr}_\pi^{n \times n}$.

5.1. Numerical Experiments

We generate numerical data following Sec. 2. We investigate varying the partial observability patterns, the underlying causal model and the type of mixing function $\mathbf{f}$. For the encoder $\mathbf{g}$ and the decoder $\hat{\mathbf{f}}$, we use 7-layer MLPs with $[10, 50, 50, 50, 50, 10] \times n$ units per layer, where n is the number of causal variables, with LeakyReLU activations for each layer except the final one in the piecewise linear case. We apply batch normalization to control the norm of $\mathbf{g}(\mathbf{X})$. For simplicity, for most experiments we set the output dimension of $\mathbf{g}$ to n. In an ablation in App. E.1.4

we set the output dimension of $\mathbf{g}$ to an overestimate of n, showing that our method learns to only use n dimensions and provides similar performance to the other experiments. In each setup, we average over 20 random seeds.

We experiment with different ratios of measured variables $\rho = \{1\text{VAR}, 50\%, 75\%\}$, where 1VAR is one measured variable, while 50% and 75% are the percentages of measured variables in each sample. Based on each ρ we predefine a set of K masks $\mathbf{Y}$ that satisfies Ass. 2.2. As discussed in Sec. 3, while the theoretical results assume a mask value of $\mathbf{M} = \mathbf{0}$, the specific value used is inconsequential from the theoretical point of view. On the other hand, the optimization procedure improves for mask values that are out of distribution w.r.t. the unmasked distribution of each variable, since it is then easier to recover which variables are masked in each group. We investigate different masks values and consider $\mathbf{M} = \boldsymbol{\mu} + \delta \cdot \boldsymbol{\sigma}$, where $\boldsymbol{\mu}, \boldsymbol{\sigma} \in \mathbb{R}^n$ are the mean and standard deviation of $\mathbf{Z}$ and $\delta \in \{0, 2, 3, 5, 10\}$.

We consider $n = \{3, 5, 10, 20, 40\}$ causal variables $\mathbf{C}$. In each experiment, we generate a random directed acyclic graph $\mathcal{D}$ from a random graph Erdös-Rényi (ER)-k model, where $k \in \{0, 1, 2, 3\}$ and ER-k is a graph with $n \cdot k$ edges. In particular, ER-0 implies independent $\mathbf{C}$ and therefore $\mathbf{Z}$. Based on $\mathcal{D}$, we consider three types of structure causal models (SCM): i) a *linear Gaussian SCM* where edge weights are sampled uniformly from $[-2, -0.5] \cup [0.5, 2]$ and we use standard Gaussian noises; ii) a *linear exponential SCM*, where we have a similar setup, but with exponential noises with scale 1; iii) a *nonlinear SCM*, where we simulate a nonlinear function with a linear layer, a sigmoid activation and a fully connected layer with 100 hidden units, where edge weights are sampled uniformly from $[-2, -0.5] \cup [0.5, 2]$ and we use standard Gaussian noises.

Results for linear mixing function (Thm. 3.1). We use a fully connected layer to model the linear mixing function $\mathbf{f}$. We show the performance, measured in the average MCC over three random seeds in Table 1. In the first four rows, we investigate how the number of the latent causal variables influences the performance for a linear Gaussian SCM with an average degree $k = 1$ and a ratio of measured variables $\rho = 50\%$. In this case, the method achieves excellent performances for smaller n, but these degrade as n increases. In the second group of rows, we consider the effect of k, the average degree in the causal graph in the linear Gaussian case for $n = 10$ causal variables and $\rho = 50\%$. Also in this case, the performances are excellent for low k, including $k = 0$ which represents independent variables, but they degrade with higher k. The third and fourth group of rows show how the performance varies for different k for the linear exponential and nonlinear SCM, showing a similar performance. Finally, we show the results for varying ratio of measured variables ρ, showing a small degradation when we measure

Table 1: Results for the numerical experiments for linear mixing functions with $\delta = 0$. The bold font indicates which parameters are varying in each block of rows.

n	k	SCM	ρ	**MCC**
5	1	LIN. GAUSS	50 %	0.997±0.002
10	1	LIN. GAUSS	50 %	0.996±0.001
20	1	LIN. GAUSS	50 %	0.987±0.029
40	1	LIN. GAUSS	50 %	0.714±0.153
10	**0**	INDEP. GAUSS	50 %	0.998±0.001
10	**1**	LIN. GAUSS	50 %	0.996±0.001
10	**2**	LIN. GAUSS	50 %	0.904±0.113
10	**3**	LIN. GAUSS	50 %	0.793±0.142
10	**0**	INDEP. EXP	50 %	0.998±0.001
10	**1**	LIN. EXP	50 %	0.998±0.002
10	**2**	LIN. EXP	50 %	0.910±0.108
10	**3**	LIN. EXP	50 %	0.825±0.123
10	**1**	NONLINEAR	50 %	0.997±0.001
10	**2**	NONLINEAR	50 %	0.997±0.001
10	**3**	NONLINEAR	50 %	0.996±0.001
10	1	LIN. GAUSS	**1VAR**	0.998 ±0.002
10	1	LIN. GAUSS	**50 %**	0.996±0.001
10	1	LIN. GAUSS	**75 %**	0.877 ±0.096

more variables at the same time. Intuitively, measuring a smaller number of variables for each sample is the easier setting for disentangling them from the observations.

Results for piecewise linear mixing function (Thm. 3.4). For the piecewise linear $\mathbf{f}$, we use a $m = \{3, 10, 20\}$-hidden-layer MLP with $m-1$ LeakyReLU ($\alpha = 0.2$) activation functions and a final linear layer, to model the piecewise linear mixing function. The number for layers m in the $\mathbf{f}$ mixing function is a proxy for the complexity of the function $\mathbf{f}$, a linear function has $m = 0$, while the higher the m the higher the non-linearity. Following (Lachapelle et al., 2022), the weight matrices are sampled from a standard Gaussian distribution and are orthogonalized by their columns to ensure $\mathbf{f}$ is injective. In this setting, we evaluate with a linear Gaussian SCM, since this satisfies Ass. 3.2 in Thm. 3.4.

We start by showing results in Fig. 2a for the simple case of number of causal variables $n = 3$ and number of layers $m = 3$. Through optimization of Eq. (7), our approach achieves good performances in these simple scenarios. However, in more complex cases, e.g. more latent variables or more complicated $\mathbf{f}$, there is a decline in performance, as shown in Fig. 2b-f. We hypothesize that the reason for this drop is that the estimated skewness and kurtosis cannot guarantee Gaussianity, which is crucial to ensure identifiability in Thm. 3.4. We show that this is empirically the case for $n = 5$ in Fig. 6 in App. D.3. We attribute this issue to the per-group sample variance used to calculate both sample skewness and kurtosis. We test this assumption and potential of our theoretical results, by comparing with an *oracle* that has access to the masks $\mathbf{y}^i$ for each observation $\mathbf{x}^i$. This
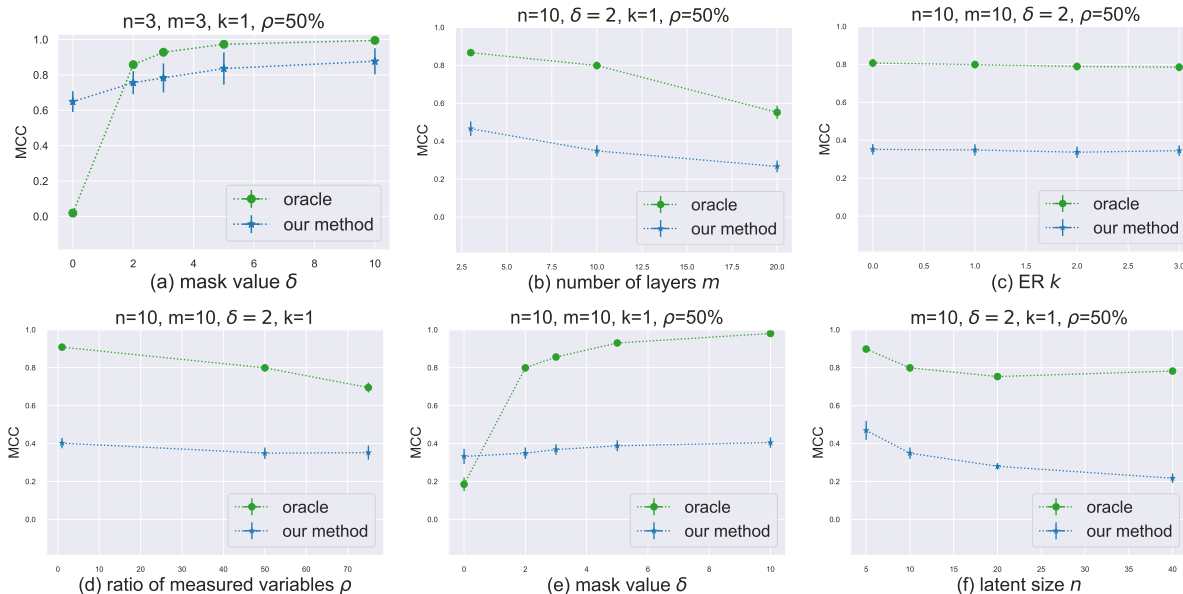
Figure 2: Results for different parameters in the piecewise linear numerical experiments. Our method implements Eq. 7 based on the group information. The *oracle* method implements the same loss, but with as additional information the mask $\mathbf{y}_i$, which it uses to assign a low variance to the masked variables in each sample for the skewness and kurtosis regularization terms. This method showcases the potential of our theoretical results with a stronger Gaussianity constraint.

.

information is used to set the group sample variance to a low value for the masked variables in each group. As shown in Fig. 2b-f this is effective in improving the performances of our sparsity constraint.

We test the performance of our method and the oracle for various parameters in Fig. 2. In particular, we see in Fig. 2a and Fig. 2e that when the distance δ between masked and unmasked variables increases, it enables a more distinct separation of $\mathbf{Z}$, providing better identification results. Similar to the linear case, we see that an increase in complexity, e.g., in the number of causal variable n, as shown in Fig. 2f, or the number of layers m, as shown in Fig. 2b, lowers the performance. Similarly, as shown in Fig. 2d, the performance drops when the ratio of measured variables ρ increases. Interestingly, the average degree of the causal graph k does not have an impact, as shown in Fig. 2c. We provide more results and visualizations in App. E.1. In App. E.1.5 we show the results of applying standard causal discovery methods on the learned representations, assuming the causal Markov and faithfulness assumption, as well as causal sufficiency.Intuitively, the closer the learned representations are to the ground truth, the more accurate are the learned causal relations between them.

5.2. Image dataset: Multiple Balls

We create a new image dataset in which we render in a 2D space b moving balls, as shown in Fig. 30 in App. E.2. Our

latent causal variables are the (x, y) position of each ball $\mathbf{b}$, which we model as Gaussian. We consider two settings: i) a *missing ball* setting in which the balls can only move along the x-axis and they can move out of view, and ii) a *masked position* setting in which the balls can move freely inside the frame and each of their coordinates can be masked by being set to an unknown, but fixed, $\mathbf{M}$ value. Both of these settings showcase possible applications of our approach. The first setting is a simplified version of a setting with a fixed camera capturing a set of objects that might move out of view. The second setting is a simplified version of occlusion, in which an occluded object is not captured in the image, but it still interacts with other objects.

The *missing ball* setting represents the intuitive setting for partial observability, i.e. when an object is out of the frame or occluded. While we model each object with two causal variables, its x and y coordinates, our methods do not allow that masks for two variables are deterministically related (Ass. 2.2). Thus, we constrain the balls to move only on the x-axis, which we then identify from the observations. In order to test also the identifiability of each variable of the same object, we devise the *masked position* setting. In this setting, we can still use our method with a non-zero mask value for each variable, which in this case represents a specific value for one of the x or y coordinates of the ball. We generate datasets for both settings by varying the number of balls $b = 2, 5, 8$. We use a predefined set of

7

Table 2: Results for the multiple balls dataset.

b	MCC Missing	MCC Masked
2	0.963±0.012	0.946±0.005
5	0.950±0.011	0.939±0.003
8	0.928±0.004	0.901±0.002

Table 3: MCC on *PartialCausal3DIdent* over each object class, with masking distance $\delta = 10$ for all latent variables.

OBJECT CLASS ID	**0**	**1**
MCC (MEAN ± STD)	0.842 ± 0.018	0.804 ± 0.023
OBJECT CLASS ID	**2**	**3**
MCC (MEAN ± STD)	0.828 ± 0.014	0.820 ± 0.009
OBJECT CLASS ID	**4**	**5**
MCC (MEAN ± STD)	0.837 ± 0.020	0.821 ± 0.033
OBJECT CLASS ID	**6**	**AVG. MEAN**
MCC (MEAN ± STD)	0.858 ± 0.005	0.832 ± 0.016

K masks $\mathbf{y}$ that satisfies Ass. 2.2. For the missing ball setting, we generate the x-coordinates of each ball from a truncated normal distribution $\mathcal{N}(0.5, 0.1^2)$ with bounds $(0.1, 0.9)$. For the masked position setting, we generate the x and y-coordinates of each ball $i \in [b]$ independently from a truncated 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ with bounds $(0.1, 0.9)^2$, where $\boldsymbol{\mu}_i \sim Unif(0.4, 0.6)^2$, and $\boldsymbol{\Sigma}_i = ((0.01, 0.005)(0.005, 0.01))$ for all groups. For both settings, we generate the masked causal variables as $(\mathbf{Z}_k)_{k=1}^K = \mathbf{y} \cdot \mathbf{C} + (1-\mathbf{y}) \cdot \mathbf{M} \in \mathbb{R}^{K \times n}$, where $\mathbf{M} = \mathbf{0}$ for the missed ball setting, and $\mathbf{M} = \boldsymbol{\mu}_i + \delta \cdot \sigma_i$ for the masked position setting. Instead of a fixed-size training dataset, we generate images online until convergence. We provide more details in App. E.2.

Results. As illustrated in Table 2, while the MCC decreases with an increase in the number of balls b, all MCCs remain above 0.90. The results are consistently higher in the missing ball setting, where there are b variables to reconstruct, while in the masked position setting there are $2b$. Additionally, in the masked position setting, we have dependence between the x and y coordinates, which can make the problem more challenging.

5.3. Image dataset: PartialCausal3DIdent

We explore the capability of our method on a partial observability version of Causal3DIdent (von Kügelgen et al., 2021) that we create. Causal3DIdent collects images of 7 object classes rendered from 10 causal variables, including object color, object positions, spotlight positions, etc. Each latent variable is rescaled into an interval of $[-1, 1]$. The mixing function, i.e., the rendering process, is not piecewise linear (which violates the assumption in Thm. 3.4). We still evaluate our piecewise linear method, following the intuition that non-linear functions can be approximated up to an arbitrary precision by an adequate number of linear pieces.

Dataset generation. Since Causal3DIdent is fully observable, we sample from it to create *PartialCausal3DIdent*, a dataset in which some of the latent variables are masked to a predefined value. For each datum, we sample a latent vector $\mathbf{C} \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ of $n = 10$ independent causal variables. We apply the set of K predefined masks $\mathbf{y} \in \{0, 1\}^{K \times n}$ to get a set of masked latent variables $(\mathbf{Z}_k)_{k=1}^K = \mathbf{y} \cdot \mathbf{C} + (1-\mathbf{y}) \cdot \mathbf{M} \in \mathbb{R}^{K \times n}$. We define the masked value as the maximum of the support area (which

is 1 for all latents). The ratio of the measured variables ρ varies from 10% (only one latent is measured) to 100% (all latents are measured). The average ρ is set to 50%. After obtaining the masked latent variables, we retrieve K corresponding images from the dataset, based on the index searching scheme provided by von Kügelgen et al. (2021).

Results. We evaluate our method separately on each object class in *PartialCausal3DIdent* and show the results in Table 3. Although the performance fluctuates slightly across classes, our method achieves a high MCC over 80% for all classes, which verifies that our approach is empirically applicable even on highly nonlinear high dimensional data. We provide all details and ablation studies on δ in App. E.3.

6. Related Work

Most closely related to our work are recent identifiability studies, which also explicitly learn causal representations in a partially observable setting. Yao et al. (2023) consider learning from tuples of simultaneously observed views, which depend on different fixed (potentially overlapping) subsets of latents with modality-specific mixing functions, and prove identifiability results for different blocks of shared content variables (von Kügelgen et al., 2021). Compared to our setting, such paired multi-view data may be harder to obtain. Sturma et al. (2023) study an unpaired, multi-domain setup, in which observations from each domain depend on a different fixed subset of latents, and show identifiability of the causal representation and graph in the fully linear case. This setting resembles our results for the piecewise linear case in Thm. 3.4, where we assume we have the group information for each observation, which can be seen as a single domain. On the other hand, for the linear case in Thm. 3.1, we do not need the group information, hence we also allow for mixtures of data from multiple domains.

Other works in an i.i.d. setting can be viewed as modelling partial observability *implicitly* by restricting the graph connecting latent and observed variables and establishing identifiability for linear (Adams et al., 2021; Cai et al., 2019;

Silva et al., 2006; Xie et al., 2020; 2022) or discrete (Kivva et al., 2021) settings. In our work, we consider the case in which either the causal model or the mixing are not linear and do not constrain the connectivity between $\mathbf{Z}$ and $\mathbf{X}$.

Not considering partial observability, other works on CRL aim to also learn the causal graph from different types (hard/soft, single/multi-node) of interventions in linear (Squires et al., 2023; Bing et al., 2023), partially (Buchholz et al., 2023; Ahuja et al., 2023a;b; Zhang et al., 2023), or fully nonlinear (von Kügelgen et al., 2023; Varici et al., 2023) settings. We focus instead on only recovering the latent causal variables without access to interventional data.

Other works have also explored the piecewise linear setting for identifiability, including with the assumption of Gaussian causal variables. In particular, Thm. 3.4 resembles one of the identifiability results from Kivva et al. (2022) which assumes $\mathbf{Z}$ is a mixture of *non-degenerate* Gaussians and $\mathbf{f}$ is a piecewise linear function. We note that, in Thm. 3.4, $\mathbf{Z}$ is also a mixture of Gaussians, where the "cluster index" corresponds to $\mathbf{Y}$. However, this mixture contains components which are degenerate, in the sense that their covariances might be singular (this occur when $\mathbf{Y}_i = 0$ for some i). This prevents us from applying the result of Kivva et al. (2022) directly to our setting. On the other hand, although we allow for degenerate components, our result assumes knowledge of the groups g, unlike Kivva et al. (2022).

Prior work has also leveraged sparsity in representation learning, for example via a Spike and Slab prior (Tonolini et al., 2020; Moran et al., 2022), assuming structural sparsity on the support of Jacobian of nonlinear mixing function (Zheng et al., 2022; Zheng and Zhang, 2023), exploiting the sparsity constraint on the linear mixing function (Ng et al., 2023), or by relating the learnt representation to multiple tasks, each depending only on a small subset of latents (Lachapelle et al., 2023; Fumero et al., 2023). Other work that is closely related to ours is work by Lachapelle et al. (2022; 2024), who have proposed a sparsity principle for identifiable CRL in interventional and temporal settings, motivated by the sparse mechanism shift hypothesis (Schölkopf et al., 2021; Perry et al., 2022).

Among these works leveraging sparsity, Lachapelle et al. (2023) introduce a sparsity principle that inspired our work, but for the multi-task setting, where each task depends on a subset of variables. More precisely, they assume that the connections W between tasks Y and representations $f(X)$ are sparse, i.e. $Y = Wf(X)$ and W is a column-sparse matrix (some of the columns are filled with zeros). In our setting, the sparsity assumption is on the representation itself, i.e. each causal variable has a positive probability of being masked. While the proof strategies are similar, our result also had to address the piecewise linear case. Other work that is closely related to ours is work by Lachapelle et al.

(2022; 2024), who have proposed a sparsity principle for identifiable CRL in interventional and temporal settings, motivated by the sparse mechanism shift hypothesis (Schölkopf et al., 2021; Perry et al., 2022).

Our work is also related to *sparse component analysis* (Gribonval and Lesage, 2006) and *sparse dictionary learning* (Mairal et al., 2009). These unsupervised representation learning methods assume linear mixing functions $\mathbf{f}$ and learn a sparse representations of the input $\mathbf{X}$, similar to our $\mathbf{g}(\mathbf{X})$. The identifiability of dictionary learning has been studied, e.g. by Hu and Huang (2023), in the finite sample regime. The main distinction with our work is that we focus on identifiability with nonlinear mixing.

Similar to our piecewise linear theorem, Liu et al. (2022) also assumes that the underlying SCM is linear Gaussian. As opposed to our work, it leverages the fact that the coefficients (or weights) of the causal relations are varying across environments, while in our case we leverage the partial observability patterns and assume that the underlying SCM is the same for all data. Similarly, Liu et al. (2024) extends the results from Liu et al. (2022) to polynomial causal models with exponential family noise variables. These works are related to ours in leveraging changes across environments, but provide a different type of results.

7. Conclusions and Limitations

In this work, we focused on learning causal representations in the *unpaired partial observability* setting, i.e., when only an instance-dependent subset of causal variables are captured in the measurements. We first proved the identifiability with linear mixing functions $\mathbf{f}$ under a sparsity constraint. We then presented an example to illustrate why extending the results to nonlinear $\mathbf{f}$ is not possible without additional assumptions. We proved identifiability when $\mathbf{f}$ is piecewise linear, both the causal variables and the learned representations are Gaussian, and we know the group of each sample.

While our experiments validate our theoretical results, there are still several limitations. From the theoretical point of view, knowing the group of each sample might not always be possible, so extending our results beyond this limitation is an exciting direction. Additionally, piecewise linear functions are a limited class and there might be other classes of nonlinear functions for which our identifiability results could be extended. In particular, our results hinge on the linearity of the composition g and f on $\mathcal{Z}$, which can be implied by other types of constraints on f, g and the causal variables. Finally, our Gaussianity constraint is empirically difficult to satisfy, as shown by the gap between the performances of our method and the oracle. This warrants further investigation in other ways to encourage Gaussianity of the learned representations.

Acknowledgments

Impact Statement

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

References

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

Kartik Ahuja, Jason S Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. *Advances in Neural Information Processing Systems*, 35:15516–15528, 2022.

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023a.

Kartik Ahuja, Amin Mansouri, and Yixin Wang. Multi-domain causal representation learning via weak distributional invariances. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023b.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. *arXiv preprint arXiv:2311.02695*, 2023.

Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco S Cohen. Weakly supervised causal representation learning. *Advances in Neural Information Processing Systems*, 35: 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. In *Advances in Neural Information Processing Systems*, 2023.

Ruichu Cai, Feng Xie, Clark Glymour, Zhifeng Hao, and Kun Zhang. Triad constraints for learning causal structure of latent variables. *Advances in neural information processing systems*, 32, 2019.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 1994.

S.H. Friedberg, A.J. Insel, and L.E. Spence. *Linear Algebra*. Pearson Education, 2014. ISBN 9780321998897.

Marco Fumero, Florian Wenzel, Luca Zancato, Alessandro Achille, Emanuele Rodolà, Stefano Soatto, Bernhard Schölkopf, and Francesco Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. *arXiv preprint arXiv:2304.07939*, 2023.

Jose Gallego-Posada and Juan Ramirez. Cooper: a toolkit for lagrangian-based constrained optimization, 2022.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Rémi Gribonval and Sylvain Lesage. A survey of sparse component analysis for blind source separation: principles, perspectives, and new challenges. In *ESANN'06 proceedings-14th European Symposium on Artificial Neural Networks*, pages 323–330. d-side publi., 2006.

Jingzhou Hu and Kejun Huang. Global identifiability of ℓ_1-based dictionary learning via matrix volume optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

Aapo Hyvärinen and Stephen M Smith. Pairwise likelihood ratios for estimation of non-gaussian structural equation models. *The Journal of Machine Learning Research*, 14 (1):111–152, 2013.

Aapo Hyvarinen, Hiroaki Sasaki, and Richard 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.

Markus Kalisch, Martin Mächler, Diego Colombo, Marloes H Maathuis, and Peter Bühlmann. Causal inference using graphical models with the r package pcalg. *Journal of statistical software*, 47:1–26, 2012.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34:18087–18101, 2021.

Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–15701, 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. PMLR, 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 *International Conference on Machine Learning*, pages 18171–18206. PMLR, 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, 2024.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Stratis Gavves. Citris: Causal identifiability from temporal intervened sequences. In *International Conference on Machine Learning*, pages 13557–13603. PMLR, 2022.

Phillip Lippe, Sara Magliacane, Sindy Löwe, Yuki M Asano, Taco Cohen, and Efstratios Gavves. Biscuit: Causal representation learning from binary interactions. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, 2023.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2022.

Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifiable latent polynomial causal models through the lens of change. In *The Twelfth International Conference on Learning Representations*, 2024.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th annual international conference on machine learning*, pages 689–696, 2009.

G Moran, D Sridhar, Y Wang, and D Blei. Identifiable deep generative models via sparse decoding. *Transactions on machine learning research*, 2022.

Ignavier Ng, Yujia Zheng, Xinshuai Dong, and Kun Zhang. On the identifiability of sparse ica without assuming non-gaussianity. *Advances in Neural Information Processing Systems*, 36, 2023.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

Ronan Perry, Julius Von Kügelgen, and Bernhard Schölkopf. Causal discovery in heterogeneous environments under the sparse mechanism shift hypothesis. *Advances in Neural Information Processing Systems*, 35:10904–10917, 2022.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

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.

Pete Shinners. Pygame, 2011.

Ricardo Silva, Richard Scheines, Clark Glymour, Peter Spirtes, and David Maxwell Chickering. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7(2), 2006.

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge MA, 2nd edition, 2000.

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *40th International Conference on Machine Learning*, 2023.

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Francesco Tonolini, Bjørn Sand Jensen, and Roderick Murray-Smith. Variational sparse coding. In *Uncertainty in Artificial Intelligence*, pages 690–700. PMLR, 2020.

Burak Varici, Emre Acartürk, Karthikeyan Shanmugam, and Ali Tajer. Score-based causal representation learning from interventions: Nonparametric identifiability. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

P Verbyla, NIB Desgranges, and L Wernisch. kpcalg: Kernel pc algorithm for causal structure detection. *URL https://CRAN. R-project. org/package= kpcalg, r package version*, 1(1), 2017.

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. *Advances in neural information processing systems*, 34: 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. In *Advances in Neural Information Processing 36*, 2023.

Liang Wendong, Armin Kekić, Julius von Kügelgen, Simon Buchholz, Michel Besserve, Luigi Gresele, and Bernhard Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, 2023.

Feng Xie, Ruichu Cai, Biwei Huang, Clark Glymour, Zhifeng Hao, and Kun Zhang. Generalized independent noise condition for estimating latent variable causal graphs. *Advances in neural information processing systems*, 33:14891–14902, 2020.

Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022.

Dingling Yao, Danru Xu, Sébastien Lachapelle, Sara Magliacane, Perouz Taslakian, Georg Martius, Julius von Kügelgen, and Francesco Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

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*, 2022.

Jiaqi Zhang, Kristjan Greenewald, Chandler Squires, Akash Srivastava, Karthikeyan Shanmugam, and Caroline Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, 2023.

Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in neural information processing systems*, 31, 2018.

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

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ica: Sparsity and beyond. *Advances in Neural Information Processing Systems*, 35:16411–16422, 2022.

A. Relation between partial observability and do interventions

In our setting we consider two types of related variables: the causal variables $\mathbf{C}$ and the masked causal variables $\mathbf{Z}$. The *causal variables* $\mathbf{C} = (C_1, ..., C_n)$ represent the underlying causal system. The *masked causal variables* $\mathbf{Z} = (Z_1, ..., Z_n) \in \mathcal{Z}$ are a combination of the causal variables and the masks $\mathbf{Y}$, and they are the latent inputs to the mixing function $\mathbf{f}$ that we are trying to recover. In particular, they are the Hadamard product of the causal variables with the binary mask variable, i.e., $\mathbf{Z} = \mathbf{Y} \odot \mathbf{C}$. For sample $i \in [N]$ and any causal variable $j \in [n]$, if the mask value y_j^i is 1, then the causal variable c_j^i is measured and z_j^i is c_j^i. Instead, if $y_j^i = 0$, then the causal variable is unmeasured and z_j^i takes a fixed masked value M_j. We show an example in Fig. 3a, where the binary mask Y_2 for causal variable C_2 is 0.

A superficially similar operation is a *do-intervention* (Pearl, 2009), in which we fix the value of a random variable to a specific value. For example, one could consider an intervention $\mathrm{do}(z_j^i = M_j)$, which forces the value of the variable z_j^i to M_j. As shown in Fig. 3, there are several differences between the masked variables and the intervened variables.
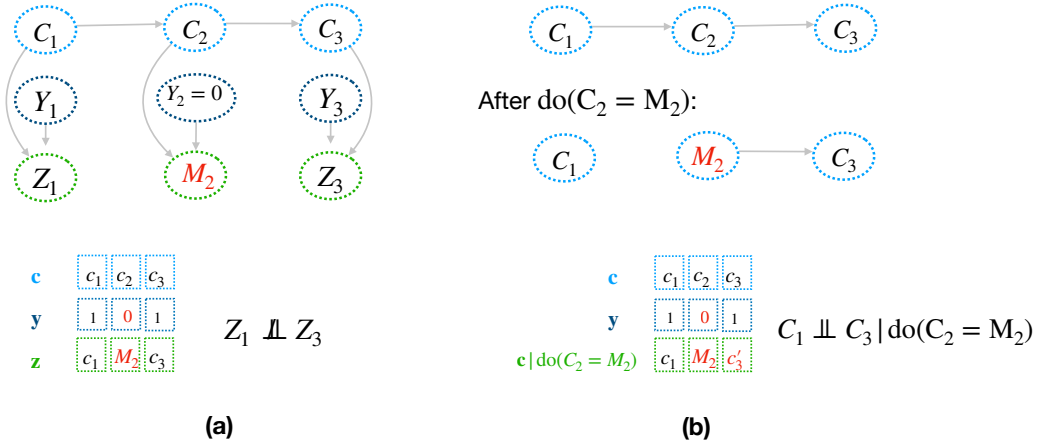

Figure 3: Comparison between (a) masking on C_2, and (b) do intervention on C_2. In the second case, there is an effect on C_3, and the intervention cuts the link and hence the dependence between C_1 and C_3.

Despite having the same fixed value for the masked or intervened variable, these two operations have different effects. In particular, masking variables does not influence any downstream variables, as an intervention would. For example, consider the case of three variables $C_1 \to C_2 \to C_3$ in Fig. 3a. In this case, if we mask C_2, then there will be no effect on the value of C_3, and C_1 and C_3 will still be dependent. If we performed an intervention on C_2, as in Fig. 3b, then there would be a change in the value of C_3, and C_1 and C_3 would be independent.

B. Proofs

B.1. Proof of the results for linear mixing functions (Theorem 3.1)

We first introduce the definition of *dependent inputs*, which intuitively is the set of variables on which a reconstruction of a given variable depends.

Definition B.1. Let $\mathbf{v} : \mathcal{Z} \to \mathbb{R}^n$ be a function with variables $\mathbf{z} = (z_1, ..., z_n) \in \mathcal{Z}$. For all $i \in [n]$ consider N_i to be the set of all variables on which $\mathbf{v}_i$ depends, which we will call *dependent inputs*. Formally, we define the set of dependent inputs $N_i \subseteq [n]$ as

$$N_i := \{ j \in [n] \mid \exists (z_j, \mathbf{z}_{-j}^0), (z_j', \mathbf{z}_{-j}^0) \in \mathcal{Z} \text{ where } z_j \neq z_j' \text{ s.t. } \mathbf{v}_i(\mathbf{z}_{-j}^0, z_j) \neq \mathbf{v}_i(\mathbf{z}_{-j}^0, z_j') \}, \tag{8}$$

where $\mathbf{z}_{-j}^0$ is any $n - 1$ dimensional vector that represents all of the components of a vector $\mathbf{z} \in \mathcal{Z}$ except for the index j.

This definition intuitively represents the indices of $\mathbf{z}$ for which the function $\mathbf{v}_i$ is not constant, or in other words, the inputs on which $\mathbf{v}_i$ depends. Note that if there is only one z_j such that $(z_j, \mathbf{z}_{-j}^0) \in \mathcal{Z}$ for an appropriate $\mathbf{z}_{-j}^0$, then any function defined on $\mathcal{Z}$ is constant for that index, and hence N_i does not contain j.

We now show a lemma that we will use to prove the theorem showing that for any diffeomorphism and any variable, there always exists a permutation that ensures that a variable is in the set of its dependent inputs. Intuitively, this ensures that for all variables, there always exists a permutation, such that the reconstruction of a given variable depends on the variable itself.

Lemma B.2 (Existence of permutation π s.t. $i \in N_{\pi(i)}$). *Let $\mathbf{v} : \mathcal{Z} \to \mathbb{R}^n$ be a diffeomorphism onto its image with variables $\mathbf{z} = (z_1, ..., z_n) \in \mathcal{Z}$. Assume there exists a point $\mathbf{z}^0 \in \mathcal{Z}$ and a value $\epsilon > 0$, such that $\forall i \in [n]$, $\mathbf{e}_i \odot [-\epsilon, \epsilon)^n + \mathbf{z}^0 \subset \mathcal{Z}$, where $\mathbf{e}_i$ is the standard basis for space $\mathbb{R}^n$ for the i-th dimension, i.e. a n-dimensional vector in which all dimensions except i are 0, and the i-th dimension is 1. Then there exists a permutation $\pi : [n] \to [n]$ such that $i \in N_{\pi(i)}$ for all i, where N_i is defined as in Def. B.1.*

Proof. Since $\mathbf{v}$ is a diffeomorphism, its Jacobian $D\mathbf{v} = \{\frac{\partial \mathbf{v}_i}{\partial z_j}\}_{i,j \in [n]}$ is invertible everywhere, so it is invertible at $\mathbf{z}^0 \in \mathcal{Z}$. Since $D\mathbf{v}(\mathbf{z}^0)$ is invertible, we have that its determinant is non-zero, i.e.

$$\det(D\mathbf{v}(\mathbf{z}^0)) := \sum_{\pi \in \mathfrak{S}_n} \operatorname{sign}(\pi) \prod_{i=1}^{n} D\mathbf{v}(\mathbf{z}^0)_{\pi(i),i} \neq 0 \,, \tag{9}$$

where $\mathfrak{S}_n$ is the set of n-permutations. This equation implies that at least one term of the sum is non-zero, and that for that term, all of the terms in the product are non-zero, meaning:

$$\exists \pi \in \mathfrak{S}_n, \forall i \in [n], D\mathbf{v}(\mathbf{z}^0)_{\pi(i),i} \neq 0 \,. \tag{10}$$

This means that, for all $i \in [n]$, $\frac{\partial \mathbf{v}_{\pi(i)}}{\partial z_i}(\mathbf{z}^0) \neq 0$, which implies that $\mathbf{v}_{\pi(i)}$ is not constant for z_i in $\mathbf{z}^0$. Then by definition of N_i in Def. B.1, $i \in N_{\pi(i)}$.

$\square$

As an intermediate step, we first prove an important lemma that shows that by enforcing sparsity of the transformed variables, the corresponding transformation is an element-wise linear function. Intuitively, these transformed variables will be the reconstructed masked latent variables $\mathbf{Z}$.

Lemma B.3 (Element-wise Identifiability for Linear Transformation). *Assume that the masked latent variables $\mathbf{Z}$ with support $\mathcal{Z}$ follow the data generating process in Sec. 2 and Ass. 2.2 holds. Let the function $\mathbf{v} : \mathcal{Z} \to \mathbb{R}^n$ be invertible and linear on $\mathcal{Z}$, and*

$$\mathbb{E} \left\| \mathbf{v}(\mathbf{Z}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 \,. \tag{11}$$

Then $\mathbf{v}$ is a permutation composed with an element-wise invertible linear transformation on $\mathcal{Z}$.

Proof. We reuse the definition of the support indices $\mathbf{S} := \{i \in [n] : Z_i \neq 0\}$ and analyze each side of inequality (11), starting with its right-hand side.

$$\mathbb{E}\|\mathbf{Z}\|_0 = \mathbb{E} \sum_{i=1}^{n} \mathbb{1}(\mathbf{Z}_i \neq 0) \tag{12}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \mathbb{E} \left[\sum_{i=1}^{n} \mathbb{1}(\mathbf{Z}_i \neq 0) \mid \mathbf{S} = \mathbf{s} \right] \tag{13}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(\mathbf{Z}_i \neq 0) \mid \mathbf{S} = \mathbf{s}] \tag{14}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{1}(i \in \mathbf{s}) \tag{15}$$

14

Now analyzing the left hand side of (11), starting with similar steps as previously we get

$$\mathbb{E}||\mathbf{v}(\mathbf{Z})||_0 = \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(\mathbf{v}_i(\mathbf{Z}) \neq 0) \mid \mathbf{S} = \mathbf{s}] \tag{16}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) \neq 0] \tag{17}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \left(1 - \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0]\right) . \tag{18}$$

For $\mathbf{v}$ to be a permutation composed with an element-wise invertible linear transformation on $\mathcal{Z}$, it is enough to show there exists a permutation $\pi : [n] \rightarrow [n]$ such that, for every i, $N_i = \{\pi(i)\}$. To achieve this, we are going to first show that

$$\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \alpha_i \mathbb{1}(N_i \cap \mathbf{s} = \varnothing) , \tag{19}$$

where $\alpha_i \in \{0, 1\}$. Since $\mathbf{v}_i$ is linear on $\mathcal{Z}$, we have that $\mathbf{v}_i(\mathbf{Z}) = \mathbf{w}^i \cdot \mathbf{Z} + \mathbf{c}_i$ for some $\mathbf{w}^i \in \mathbb{R}^n$. Furthermore, we show that in this case $N_i = \{j \in [n] \mid \mathbf{w}_j^i \neq 0\}$. In one direction, if $\mathbf{w}_j^i = 0$, then $\mathbf{v}_i$ is constant in dimension j, so $\mathbf{w}_j^i$ should be non-zero for j to be included in N_i. For the other direction, if $\mathbf{w}_j^i \neq 0$ there are two cases in which we could have that $j \notin N_i$: (i) if $\mathbf{w}_j^i \mathbf{Z}_j$ always cancels out with another $\mathbf{w}_k^i \mathbf{Z}_k$, where both $\mathbf{w}_j^i, \mathbf{w}_k^i \neq 0$, (ii) if the variable $\mathbf{Z}_j$ can only take a single value. We show that neither of these cases can happen because of Ass. 2.2. In particular, the first case cannot happen because two masks for two different variables cannot be the same and satisfy this assumption. The second case also cannot happen because of Ass. 2.2, since there needs to be a support index set in which i is masked and one in which it is not masked. Thus,

$$\mathbf{v}_i(\mathbf{Z}) = \mathbf{w}^i \cdot \mathbf{Z} + \mathbf{c}_i = \mathbf{w}_{N_i}^i \cdot \mathbf{Z}_{N_i} + \mathbf{c}_i .$$

Case 1: Suppose $N_i \cap \mathbf{s} = \varnothing$. Then,

$$\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}_{N_i}^i \cdot \mathbf{Z}_{N_i} + \mathbf{c}_i = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}_{N_i}^i \cdot \mathbf{0} + \mathbf{c}_i = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{c}_i = 0] = \alpha_i .$$

Note that the event $\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{c}_i = 0]$ is deterministically either true of false, hence $\alpha_i \in \{0, 1\}$.

Case 2: Suppose $N_i \cap \mathbf{s} \neq \varnothing$. Thus, $\mathbf{w}_\mathbf{s}^i \neq \mathbf{0}$. Thus,

$$\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}^i \cdot \mathbf{Z} + \mathbf{c}_i = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}_\mathbf{s}^i \cdot \mathbf{Z}_\mathbf{s} + \mathbf{c}_i = 0] .$$

Note that the event $\{\mathbf{Z}_\mathbf{s} \mid \mathbf{w}_\mathbf{s}^i \cdot \mathbf{Z}_\mathbf{s} + \mathbf{c}_i = 0\}$ corresponds to the kernel of the linear map $\mathbf{w}_\mathbf{s}^i$. We can thus infer its dimensionality via the rank-nullity theorem (Friedberg et al., 2014) which states that $\text{rank}(\mathbf{w}_\mathbf{s}^i) + \dim(\text{Ker}(\mathbf{w}_\mathbf{s}^i)) = \dim(\text{Dom}(\mathbf{w}_\mathbf{s}^i))$, where Ker() is nullity and Dom() is domain, which here implies that $1 + \dim(\text{Ker}(\mathbf{w}_\mathbf{s}^i)) = |\mathbf{s}|$. We thus have $\dim(\{\mathbf{Z}_\mathbf{s} \mid \mathbf{w}_\mathbf{s}^i \cdot \mathbf{Z}_\mathbf{s} + \mathbf{c}_i = 0\}) = |\mathbf{s}| - 1$. Since $\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}$ has a density w.r.t. to the Lebesgue measure, we have that $\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}_\mathbf{s}^i \cdot \mathbf{Z}_\mathbf{s} + \mathbf{c}_i = 0] = 0$ (since a density w.r.t. to Lebesgue cannot concentrate mass on a lower-dimensional linear subspace).

We thus have proved that indeed, $\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \alpha_i \mathbb{1}(N_i \cap \mathbf{s} = \varnothing)$.

Putting (11), (15), (18) and (19) together, we obtain

$$\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [(1 - \alpha_i) + \alpha_i \mathbb{1}(N_i \cap \mathbf{s} \neq \varnothing)] \leq \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{1}(i \in \mathbf{s}) \tag{20}$$

We can now use Lemma B.2, because Ass. 2.2 implies the existence of the $\mathbf{z}_0$ point, which in this case is $(0, ..., 0)$.

By using this lemma, we can show that there exists a permutation π such that, for all $i \in [n], i \in N_{\pi(i)}$. We now permute the terms on the l.h.s. according to π and reorganize the terms as:

$$\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [(1 - \alpha_{\pi(i)}) + \alpha_{\pi(i)} \mathbb{1}(N_{\pi(i)} \cap \mathbf{s} \neq \varnothing)] \leq \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{1}(i \in \mathbf{s})$$

$$\sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [(1 - \alpha_{\pi(i)}) + \alpha_{\pi(i)} \mathbb{1}(N_{\pi(i)} \cap \mathbf{s} \neq \varnothing) - \mathbb{1}(i \in \mathbf{s})] \leq 0 \tag{21}$$

15

Note how, for all i, $(1-\alpha_{\pi(i)})+\alpha_{\pi(i)}\mathbb{1}(N_{\pi(i)}\cap\mathbf{s}\neq\varnothing)-\mathbb{1}(i\in\mathbf{s})\geq 0$, since whenever $i\in\mathbf{s}$, we must have $N_{\pi(i)}\cap\mathbf{s}\neq\varnothing$, because we chose a permutation such that $i\in N_{\pi(i)}$. Note also that this is true irrespective of the value of $\alpha_{\pi(i)}\in\{0,1\}$.

We then can notice that if $i\notin\mathbf{s}$, then $\alpha_{\pi(i)}=0$. The function can have either value 0 or 1, but in any case not negative. Hence the inequality in (21) is actually an equality and hence for all $\mathbf{s}\in\mathcal{S}$ and all $i\in[n]$,

$$(1-\alpha_{\pi(i)})+\alpha_{\pi(i)}\mathbb{1}(N_{\pi(i)}\cap\mathbf{s}\neq\varnothing)-\mathbb{1}(i\in\mathbf{s})=0\,. \tag{22}$$

The first thing we conclude is that if $i\notin\mathbf{s}$, then $\alpha_{\pi(i)}=1$, since otherwise Equ.(22) is violated. Under Ass. 2.2, we have that, for all $i\in[n]$, there exists an $\mathbf{s}\in\mathcal{S}$ such that $i\notin\mathbf{s}$. We thus conclude that $\alpha_i=1$ for all $i\in[n]$, which allows us to write

$$\mathbb{1}(N_{\pi(i)}\cap\mathbf{s}\neq\varnothing)=\mathbb{1}(i\in\mathbf{s}) \tag{23}$$
$$\mathbb{1}(N_{\pi(i)}\cap\mathbf{s}=\varnothing)=\mathbb{1}(i\notin\mathbf{s})\,. \tag{24}$$

Importantly, this means

$$\forall i\in[n],\forall\mathbf{s}\in\mathcal{S},\ i\notin\mathbf{s}\implies N_{\pi(i)}\cap\mathbf{s}=\varnothing\implies N_{\pi(i)}\subseteq\mathbf{s}^c\,, \tag{25}$$

which can be rewritten as

$$\forall i\in[n], N_{\pi(i)}\subseteq\bigcap_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}}\mathbf{s}^c\,. \tag{26}$$

We now rewrite Assumption 2.2 below and take the complement on both sides:

$$\forall i\in[n],\ \bigcup_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}}\mathbf{s}=[n]\setminus\{i\} \tag{27}$$
$$\bigcap_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}}\mathbf{s}^c=\{i\} \tag{28}$$

Combining (26) with (28) implies that $N_{\pi(i)}=\{i\}$ for all i, which concludes the proof. $\qquad\square$

We now prove identifiability up to permutation and element-wise linear transformations for the case of a linear mixing function, given the assumption of sufficient support index variability.

Theorem 3.1 (Element-wise Identifiability for Linear $\mathbf{f}$). *Assume the observation $\mathbf{X}=\mathbf{f}(\mathbf{Z})$ follows the data-generating process in Sec. 2, where $\mathbf{f}:\mathcal{Z}\to\mathcal{X}$ is an injective linear function, and Ass. 2.2 holds. Let $\mathbf{g}:\mathcal{X}\to\mathbb{R}^n$ be an invertible linear function onto its image and let $\hat{\mathbf{f}}:\mathbb{R}^n\to\mathbb{R}^d$ be an invertible continuous function onto its image. If both of the following conditions hold,*

$$\mathbb{E}\left\|\mathbf{X}-\hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2=0\,,\quad and \tag{1}$$
$$\mathbb{E}\left\|\mathbf{g}(\mathbf{X})\right\|_0\leq\mathbb{E}\left\|\mathbf{Z}\right\|_0\,, \tag{2}$$

then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$ up to a permutation and element-wise linear transformations (Def. 2.1), i.e., $\hat{\mathbf{f}}^{-1}\circ\mathbf{f}$ is a permutation composed with element-wise invertible linear transformations on $\mathcal{Z}$.

Proof. Since $\mathbf{X}=\mathbf{f}(\mathbf{Z})$, we can rewrite Equation (49) (perfect reconstruction) as

$$\mathbb{E}||\mathbf{f}(\mathbf{Z})-\hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2=0\,. \tag{29}$$

This means $\mathbf{f}$ and $\hat{\mathbf{f}}\circ\mathbf{g}\circ\mathbf{f}$ are equal $\mathbb{P}_{\mathbf{Z}}$-almost everywhere. Both of these functions are continuous, $\mathbf{f}$ by assumption and $\hat{\mathbf{f}}\circ\mathbf{g}\circ\mathbf{f}$ because $\hat{\mathbf{f}}$ is continuous, and $\mathbf{f},\mathbf{g}$ are linear. Since they are continuous and equal $\mathbb{P}_{\mathbf{Z}}$-almost everywhere, this means that they must be equal over the support of $\mathbf{Z}$, $\mathcal{Z}$, i.e.,

$$\mathbf{f}(\mathbf{z})=\hat{\mathbf{f}}\circ\mathbf{g}\circ\mathbf{f}(\mathbf{z})\,,\forall\mathbf{z}\in\mathcal{Z}\,. \tag{30}$$

This can be easily shown by considering any point $\mathbf{z}' \in \mathcal{Z}$ on which $\mathbf{f}$ and $\hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}$ are different, i.e. $\hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}(\mathbf{z}') \neq \mathbf{f}(\mathbf{z}')$, This would imply that $(\mathbf{f} - \hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f})$, which is also a continuous function, is non-zero in $\mathbf{z}'$, and in its neighbourhood. This would contradict the assumption that $\mathbf{f}$ and $\hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}$ are the same almost everywhere. We can now apply the inverseon both sides to obtain

$$\hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}) = \underbrace{\mathbf{g} \circ \mathbf{f}(\mathbf{z})}_{\mathbf{v}:=}, \forall \mathbf{z} \in \mathcal{Z}. \tag{31}$$

Since both $\mathbf{f}$ is an injective linear function from $\mathcal{Z}$ to $\mathcal{X}$ and $\mathbf{g}$ is an invertible linear function from $\mathcal{X}$ onto its image, then, $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is an invertible linear function on $\mathcal{Z}$. As $\mathbf{v}$ is equal to $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ on $\mathcal{Z}$, then, we have $\mathbf{v}$ is also an invertible linear function on $\mathcal{Z}$. By Lemma B.3, we can derive $\mathbf{v}$ is a permutation composed with an element-wise linear transformation on $\mathcal{Z}$. $\qquad\square$

B.2. Proof of the results for the piecewise linear functions(Theorem 3.4)

For piecewise linear mixing functions, we first prove an intermediate result (Thm B.5) for a weaker form of identifiability that does not imply a disentanglement, but an affine correspondence between the ground truth and the learned variables.

Definition B.4 (Identifiability up to affine transformation (Khemakhem et al., 2020; Lachapelle et al., 2022)). The ground truth representation vector $\mathbf{Z}$ (n-dimensional random vector) is *identified up to affine transformations* by a learned representation vector $\hat{\mathbf{Z}}$ (also n-dimensional random vector) when there exists an invertible *linear transformation* $\mathbf{h}$ such that, $\hat{\mathbf{Z}} = \mathbf{h}(\mathbf{Z})$ almost surely.

We then introduce Lemma B.10, which proves that, when $\mathbf{Y}$ is provided, we can identify the distribution of latent variables by linear transformations that may vary across $\mathbf{Y}$. It is crucial to emphasize that in this work, having $\mathbf{Y} = \mathbf{y}$ does not imply knowledge of the exact value of $\mathbf{y}$; instead, we simply need the information on *grouping*, i.e. the partitioning of the dataset based on mask values. Finally, we use these intermediate results to conclude the prove of the theorem.

B.2.1. LINEAR IDENTIFIABILITY FOR (DE)-MVNS WITH PIECEWISE AFFINE $\mathbf{f}$

In what follows, we let $\equiv$ denote equality in distribution.

Theorem B.5 (Linear Identifiability for (De)-MVNs with Piecewise Affine $\mathbf{f}$). *Assume $\mathbf{f}, \hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ are injective and piecewise affine. We assume $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ follow a (degenerate) multivariate normal distribution. If $\mathbf{f}(\mathbf{Z}) \equiv \hat{\mathbf{f}}(\hat{\mathbf{Z}})$, then there exists an invertible affine transformation $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{h}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$ (Def. B.4).*

Intuitively, Thm B.5 states that if we can align two (De)- MVN-distributed random vectors to the same distribution through piecewise affine transformations, then their distributions can be interchanged via an affine transformation. This finding is initially inspired by Kivva et al. (2022), who established the linear identifiability for latent variables following *non-denegerate* Multivariate Normal (MVN) distributions. Our primary focus revolves around addressing two key issues to extend the findings from MVN to de-MVN: i) determining the identifiability of de-MVN distributions, which do not have a probability density function, and ii) constructing an affine function when the domain changes from $\mathbb{R}^n$ to $\mathcal{Z}$.

We start by proving an useful lemma about (degenerate) multivariate normal distributions.

Lemma B.6. *(Degenerate) Multivariate Normals, or (De)MVNs, are close under affine transformation. More formally, if $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\mu \in \mathbb{R}$ and $|\boldsymbol{\Sigma}| \geq 0$, is a potentially degenerate multivariate normal variable, then $\mathbf{A}\mathbf{Z}$, where $\mathbf{A} \in \mathbb{R}^{n \times n}$ is also a potentially degenerate multivariate normal variable.*

Proof. Let $\hat{\mathbf{Z}} = \mathbf{A}\mathbf{Z}$, then $P_{\hat{\mathbf{Z}}} = \mathbf{A}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$ where the determinant of the covariance, $|\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T| \geq 0$. Therefore, $\hat{\mathbf{Z}}$ is a potentially degenerate multivariate normal variable. $\qquad\square$

We now summarize the results on the identifiability of non-degenerate multivariate normal variables by Kivva et al. (2022). We report an adapted version of Theorem C.3 by Kivva et al. (2022).

Theorem B.7 (Identifiability of non-degenerate MVNs (Kivva et al., 2022)). *Consider a pair of non-degenerate MVNs in $\mathbb{R}^n$. If*

$$P = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad and \quad P' = N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{32}$$

and there exists a ball $B(\mathbf{x}_0, \delta)$, where $\mathbf{x}_0 \in \mathbb{R}^n$ and $\delta > 0$, such that P and P' induce the same measure on $B(\mathbf{x}_0, \delta)$, then $P \equiv P'$.

The original proof follows from the identity theorem for real analytic functions. We extend this result to the case of potentially degenerate multivariate normal variables, which we call *(De)-MVNs*. We first propose an intermediate result for the case in which only one of the variables is a (degenerate) multivariate normal, while the other variable is a non-degenerate multivariate normal. We then use this result to prove the general case in which both variables are potentially degenerate MVNs.

Lemma B.8 (Identifiability of a (De)-MVNs and a non-degenerate MVN). *Consider a pair of random vectors $\mathbf{X}$, $\mathbf{X}'$ in $\mathbb{R}^n$ distributed as*

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad and \quad \mathbf{X}' \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{33}$$

for appropriate values of $\boldsymbol{\mu}, \boldsymbol{\mu}'$ and where the determinant $|\boldsymbol{\Sigma}| \geq 0$ and the determinant $|\boldsymbol{\Sigma}'| > 0$. In other words, $\mathbf{X}$ is a potentially degenerate MVN, while $\mathbf{X}'$ is a non-degenerate MVN.

If there exists a ball $B(\mathbf{x}_0, \delta) \subseteq \mathbb{R}^n$, where $\mathbf{x}_0 \in \mathbb{R}^n$ and $\delta > 0$, such that $\mathbf{X}$ and $\mathbf{X}'$ follow the same distribution on $B(\mathbf{x}_0, \delta)$, then $\mathbf{X} \equiv \mathbf{X}'$, i.e., $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}', \boldsymbol{\Sigma}')$.

Proof. Let the rank of $\boldsymbol{\Sigma}$ be $k \leq n$ and consider the spectral decomposition of $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = QDQ^T, \tag{34}$$

where Q is an orthogonal $n \times n$ matrix and D a the diagonal matrix. If $n = k$ we consider D to have k diagonal entries $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2)$ where σ_i for $i \in [k]$ are the eigenvalues. Otherwise, if $k < n$, then D has n diagonal entries $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2, 0, \ldots, 0)$ where σ_i for $i \in [k]$ are the eigenvalues.

Let $\mathbf{Y} = Q^T \mathbf{X}$ and $\mathbf{Y}' = Q^T \mathbf{X}'$. Since Q is an orthogonal matrix, this means that

$$\mathbf{Y} \sim N(Q^T \boldsymbol{\mu}, Q^T Q D Q^T Q) = N(Q^T \mu, D) \tag{35}$$
$$\mathbf{Y}' \sim N(Q^T \boldsymbol{\mu}', Q^T \Sigma' Q) \tag{36}$$

Since we know $\mathbf{X} \equiv \mathbf{X}'$ in $B(\mathbf{x}_0, \delta)$, then we can derive that $\mathbf{Y} \equiv \mathbf{Y}'$ in $B(Q^T \mathbf{x}_0, \tilde{\delta})$ for an appropriate $\tilde{\delta} > 0$. We project $B(Q^T \mathbf{x}_0, \tilde{\delta})$ into two subspaces, $B(Q^T \mathbf{x}_0, \tilde{\delta})_{1:k}$ and $B(Q^T \mathbf{x}_0, \tilde{\delta})_{k+1:n}$. The first captures the first k dimensions of the ball, and the second the last $(n - k)$ dimensions.

We can pick the first k dimensions of $\mathbf{Y}$ and $\mathbf{Y}'$, and denote them as $\mathbf{Y}_{1:k}$ and $\mathbf{Y}'_{1:k}$ respectively. The first k dimensions of both variables are still the same, i.e., $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$ in $B(Q^T \mathbf{x}_0, \tilde{\delta})_{1:k}$. We can show that $\mathbf{Y}_{1:k}$ is a non-degenerate multivariate normal, because its covariance matrix $D_{1:k,1:k}$ is full rank. Since both $\mathbf{Y}_{1:k}$ and $\mathbf{Y}'_{1:k}$ are non-degenerate multivariate normals, by Theorem B.7 by (Kivva et al., 2022) we have $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$.

We will now prove by contradiction that $\mathbf{Y}$ is also a non-degenerate MVN, i.e., that $k = n$. We consider the other $(n - k)$ dimensions of $\mathbf{Y}$ and $\mathbf{Y}'$. The covariance matrix of $\mathbf{Y}_{k+1:n}$ is $D_{k+1:n,k+1:n}$, which is a zero matrix. However, since determinant $|\boldsymbol{\Sigma}'| > 0$, the variance of any component of $\mathbf{Y}'_{k+1:n}$ cannot be 0. Since $\mathbf{Y}_{k+1:n} \equiv \mathbf{Y}'_{k+1:n}$ in the ball $B(Q^T \mathbf{x}_0, \tilde{\delta})_{k+1:n}$, their covariance matrices should be the same. We now come to a contradiction, because one is supposed to be a zero matrix, while the other one is supposed to be full rank. We therefore derive that $k = n$, and hence $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$ implies $\mathbf{Y} \equiv \mathbf{Y}'$. We can now exploit that $\mathbf{X} = Q\mathbf{Y}$ and $\mathbf{X}' = Q\mathbf{Y}'$, due to the orthogonality of Q, and conclude that $\mathbf{X} \equiv \mathbf{X}'$. $\square$

Lemma B.9 (Identifiability of (De)-MVNs). *Consider a pair of random vectors $\mathbf{X}$, $\mathbf{X}'$ in $\mathbb{R}^n$ distributed as*

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad and \quad \mathbf{X}' \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{37}$$

for appropriate values of $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\mu}', \boldsymbol{\Sigma}'$, including also singular $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$. If there exists a ball $B(\mathbf{x}_0, \delta) \subseteq \mathbb{R}^n$, where $\mathbf{x}_0 \in \mathcal{X}$, $\delta > 0$ and $\mathcal{X}$ is support of $\mathbf{X}$, such that $\mathbf{X}$ and $\mathbf{X}'$ follow the same distribution on $B(\mathbf{x}_0, \delta)$, then $\mathbf{X} \equiv \mathbf{X}'$, i.e., $(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{\mu}', \boldsymbol{\Sigma}')$.

Proof. Let the rank of $\mathbf{\Sigma}$ be $k \leq n$ and consider the spectral decomposition of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = QDQ^T, \tag{38}$$

where Q is an orthogonal $n \times n$ matrix and D a the diagonal matrix. If the rank $k < n$, i.e. $\mathbf{X}$ is a degenerate multivariate normal, we consider D to have n diagonal entries $\sigma_1^2, \sigma_2^2, ..., \sigma_k^2, 0, ..., 0$, where σ_i for $i \in [k]$ are the eigenvalues.

Let $\mathbf{Y} = Q^T\mathbf{X}$ and $\mathbf{Y}' = Q^T\mathbf{X}'$. This means that

$$\mathbf{Y} \sim N(Q^T\boldsymbol{\mu}, Q^TQDQ^TQ) = N(Q^T\boldsymbol{\mu}, D) \tag{39}$$

$$\mathbf{Y}' \sim N(Q^T\boldsymbol{\mu}', Q^T\Sigma'Q) \tag{40}$$

Since we know $\mathbf{X} \equiv \mathbf{X}'$ in $B(\mathbf{x}_0, \delta)$, then we can derive that $\mathbf{Y} \equiv \mathbf{Y}'$ in $B(Q^T\mathbf{x}_0, \tilde{\delta})$ for an appropriate $\tilde{\delta} > 0$. We project $B(Q^T\mathbf{x}_0, \tilde{\delta})$ into two subspaces, $B(Q^T\mathbf{x}_0, \tilde{\delta})_{1:k}$ and $B(Q^T\mathbf{x}_0, \tilde{\delta})_{k+1:n}$. The first captures the first k dimensions of the ball, and the second the last $(n - k)$ dimensions.

We can pick the first k dimensions of $\mathbf{Y}$ and $\mathbf{Y}'$, and denote them as $\mathbf{Y}_{1:k}$ and $\mathbf{Y}'_{1:k}$ respectively. The first k dimensions of both variables are still the same, i.e., $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$ in $B(Q^T\mathbf{x}_0, \tilde{\delta})_{1:k}$. We can show that $\mathbf{Y}_{1:k}$ is a non-degenerate multivariate normal, because its covariance matrix $D_{1:k,1:k}$ is full rank. So by Lemma B.8 we have $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$, i.e. $(Q^T\Sigma'Q)_{1:k,1:k} = D_{1:k,1:k}$.

For the other $(n - k)$ dimensions of $\mathbf{Y}$ and $\mathbf{Y}'$, i.e., $\mathbf{Y}_{k+1:n}$ and $\mathbf{Y}'_{k+1:n}$, we can also show that $\mathbf{Y}_{k+1:n} \equiv \mathbf{Y}'_{k+1:n}$ in $B(Q^T\mathbf{x}_0, \tilde{\delta})_{k+1:n}$. For $\mathbf{Y}_{k+1:n}$, since $\mathbf{x}_0$ is contained in $B(\mathbf{x}_0, \delta)$, we can derive that $Q^T\mathbf{x}_0$ is contained in $B(Q^T\mathbf{x}_0, \tilde{\delta})$. Since the covariance matrix of $\mathbf{Y}_{k+1:n}$ is $D_{k+1:n,k+1:n}$, which is a zero matrix, the distribution of $\mathbf{Y}_{k+1:n}$ is a point mass with all of the probability on a single value $(Q^T\mathbf{x}_0)_{k+1:n}$. From (39), we know that $\mathbf{Y}_{k+1:n} \sim N((Q^T\boldsymbol{\mu})_{k+1:n}, D_{k+1:n,k+1:n}) = N((Q^T\boldsymbol{\mu})_{k+1:n}, \mathbf{0})$, so we can derive $(Q^T\boldsymbol{\mu})_{k+1:n} = (Q^T\mathbf{x}_0)_{k+1:n}$.

Since $\mathbf{Y}_{k+1:n} \equiv \mathbf{Y}'_{k+1:n}$ in $B(Q^T\mathbf{x}_0, \tilde{\delta})_{k+1:n}$, and $\mathbf{Y}_{k+1:n}$ is a point mass on $(Q^T\boldsymbol{\mu})_{k+1:n} \in B(Q^T\mathbf{x}_0, \tilde{\delta})_{k+1:n}$, then we can derive that $\mathbf{Y}'_{k+1:n}$ should be a point mass on the same point $(Q^T\boldsymbol{\mu})_{k+1:n}$. Therefore, $(Q^T\Sigma'Q)_{k+1:n,k+1:n}$ is a zero matrix, which is equal to $D_{k+1:n,k+1:n}$, and $(Q^T\boldsymbol{\mu})_{k+1:n} = (Q^T\boldsymbol{\mu}')_{k+1:n}$. This means that $\mathbf{Y}_{k+1:n} \equiv \mathbf{Y}'_{k+1:n}$.

We can now exploit that $\mathbf{X} = Q\mathbf{Y}$ and $\mathbf{X}' = Q\mathbf{Y}'$ due to the orthogonality of Q, and we can write:

$$\mathbf{X} = Q\mathbf{Y} = Q(\mathbf{Y}_{1:k}{}^T, \mathbf{Y}_{k+1:n}{}^T) \tag{41}$$

$$\mathbf{X}' = Q\mathbf{Y}' = Q(\mathbf{Y}'_{1:k}{}^T, \mathbf{Y}'_{k+1:n}{}^T), \tag{42}$$

which together with $\mathbf{Y}_{1:k} \equiv \mathbf{Y}'_{1:k}$ and $\mathbf{Y}_{k+1:n} \equiv \mathbf{Y}'_{k+1:n}$ implies that $\mathbf{X} \equiv \mathbf{X}'$. $\qquad\square$

Theorem B.5 (Linear Identifiability for (De)-MVNs with Piecewise Affine $\mathbf{f}$). *Assume $\mathbf{f}, \hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ are injective and piecewise affine. We assume $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ follow a (degenerate) multivariate normal distribution. If $\mathbf{f}(\mathbf{Z}) \equiv \hat{\mathbf{f}}(\hat{\mathbf{Z}})$, then there exists an invertible affine transformation $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{h}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$ (Def. B.4).*

Proof. By the definition of (De) MVN, we can write $\mathbf{Z} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$ with dimension $d_\varepsilon \leq n$ and support $\mathbb{R}^{d_\varepsilon}$, and $\mathbf{A} \in \mathbb{R}^{n \times d_\varepsilon}$ has full column rank. The support of $\mathbf{Z}$, $\mathcal{Z} \subseteq \mathbb{R}^n$, is $\mathcal{Z} = colsp(\mathbf{A}) + \mathbf{b}$, where $colsp(\mathbf{A})$ is the column space of matrix $\mathbf{A}$.

Since $f(\mathbf{Z})$ and $\hat{f}(\hat{\mathbf{Z}})$ are equally distributed, they have the same image space. Therefore, we can define $\mathbf{h}_0 := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ on $\mathcal{Z}$, which is a piecewise affine function. Since we assumed that $\mathbf{f}(\mathbf{Z})$ is equally distributed to $\hat{\mathbf{f}}(\hat{\mathbf{Z}})$, then $\hat{\mathbf{Z}} \equiv \hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{Z}) = \mathbf{h}_0(\mathbf{Z})$. We can now define another function that takes as domain the whole space $\mathbb{R}^n$:

$$\widetilde{\mathbf{h}}(\mathbf{x}) := \mathbf{h}_0(\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{x} - \mathbf{b}) + \mathbf{b}), \tag{43}$$

where $\mathbf{x} \in \mathbb{R}^n$. This function is the same as $\mathbf{h}_0$ on $\mathcal{Z}$ We first proof that $\widetilde{\mathbf{h}}$ and $\mathbf{h}_0$ agree on $\mathcal{Z}$. In particular, $\forall \mathbf{z}_0 \in \mathcal{Z}$,

$\exists \boldsymbol{\varepsilon}_0 \in \mathbb{R}^{d_\varepsilon}$, s.t. $\mathbf{z}_0 = \mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}$. Thus, by definition of $\widetilde{\mathbf{h}}$, we have $\forall \mathbf{z}_0 \in \mathcal{Z}$,

$$\widetilde{\mathbf{h}}(\mathbf{z}_0) = \mathbf{h}_0(\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{z}_0 - \mathbf{b}) + \mathbf{b}) \tag{44}$$

$$= \mathbf{h}_0(\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b} - \mathbf{b}) + \mathbf{b}) \tag{45}$$

$$= \mathbf{h}_0(\mathbf{A}\underbrace{(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{A}}_{=\mathbf{I}}\boldsymbol{\varepsilon}_0 + \mathbf{b}) \tag{46}$$

$$= \mathbf{h}_0(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}) \tag{47}$$

$$= \mathbf{h}_0(\mathbf{z}_0). \tag{48}$$

So we have $\mathbf{h}_0$ and $\widetilde{\mathbf{h}}$ agree on $\mathcal{Z}$. This implies that $\hat{\mathbf{Z}} \equiv \widetilde{\mathbf{h}}(\mathbf{Z})$.

Then, we will complete the proof by two steps:

  i There exists a $\mathbf{z}_0 \in \mathcal{Z}$ and $\delta > 0$ s.t. $\widetilde{\mathbf{h}}$ is affine on the ball $\mathbf{B}(\mathbf{z}_0, \delta) \subset \mathbb{R}^n$ (potentially not completely contained in $\mathcal{Z}$).

  ii We can define an affine function $\mathbf{h}$ on $\mathcal{Z}$ such that $\widetilde{\mathbf{h}}$ and $\mathbf{h}$ agree on the ball B. Then we show $\mathbf{h}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$.

(i). We now first show that $\exists \mathbf{z}_0 \in \mathcal{Z}$, s.t. $\widetilde{\mathbf{h}}$ is differentiable at point $\mathbf{z}_0$.

As shown in Eq. 47, $\widetilde{\mathbf{h}}(\mathbf{z}_0)$ can be written as the composition of the functions $\mathbf{h}_0(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b})$ and $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{z}_0 - \mathbf{b})$. The function $\mathbf{h}_0(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b})$ is piecewise affine, because it is a composition of (piecewise) affine maps. Hence, there must exists a point $\boldsymbol{\epsilon}_0 \in \mathbb{R}^{d_\epsilon}$ s.t. $\mathbf{h}_0(\mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b})$ is differentiable at that point. Moreover, we have that $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{x} - \mathbf{b})$ is differentiable at $\mathbf{z}_0 := \mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}$, because it is an affine function, and $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{x} - \mathbf{b})$ evaluated at $\mathbf{z}_0$ yields $\boldsymbol{\varepsilon}_0$, so the composition $\widetilde{\mathbf{h}}$ is differentiable at $\mathbf{z}_0$. Since $\widetilde{\mathbf{h}}$ is piecewise affine and differentiable in $\mathbf{z}_0$, then we can construct a ball $\mathbf{B}(\mathbf{z}_0, \delta)$ (not necessarily completely contained in $\mathcal{Z}$) which contains one single linear piece of $\widetilde{\mathbf{h}}$.

(ii) Let $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ be an invertible affine function such that $\mathbf{h}$ coincides with $\widetilde{\mathbf{h}}$ on $\mathbf{B}(\mathbf{z}_0, \delta)$. This means $\mathbf{h}(\mathbf{Z})$ and $\widetilde{\mathbf{h}}(\mathbf{Z})$ coincide on $\mathbf{B}(\mathbf{z}_0, \delta) \cap \mathcal{Z}$. Since we have shown that $\widetilde{\mathbf{h}}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$ are equal in distribution on $\mathcal{Z}$, then $\mathbf{h}(\mathbf{Z}) = \widetilde{\mathbf{h}}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$ on $\mathbf{B}(\mathbf{z}_0, \delta) \cap \mathcal{Z}$.

Since we assume $\mathbf{Z}$ is a (De-)MVN, by Lemma B.6, $\mathbf{h}(\mathbf{Z})$ is a (De-)MVN as well, because $\mathbf{h}$ is affine. Moreover, we know $\hat{\mathbf{Z}}$ is a (De-)MVN by assumption. We leverage the fact that they are equal on the intersection of the ball $\mathbf{B}(\mathbf{z}_0, \delta)$ and the support of $\mathcal{Z}$ and use Lemma B.9 to prove that $\mathbf{h}(\mathbf{Z}) \equiv \hat{\mathbf{Z}}$. $\qquad\square$

B.2.2. LINEAR IDENTIFIABILITY GIVEN $\mathbf{Y} = \mathbf{y}$ FOR PIECEWISE LINEAR $\mathbf{f}$

We now show that given the information of the binary mask $\mathbf{Y}$, we can identify the latent factors $\mathbf{Z}$ up to an affine transformation (Def. B.4). It is crucial to emphasize that in this paper, having $\mathbf{Y} = \mathbf{y}$ does not imply knowledge of the exact value of $\mathbf{y}$; instead, it simply needs the information on *grouping*, i.e. the partitioning of the dataset based on mask values.

Lemma B.10 (Linear Identifiability given $\mathbf{Y} = \mathbf{y}$ for Piecewise Linear $\mathbf{f}$). *Assume the observation $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ follows the data-generating process in Sec. 2, Ass. 3.2 and 3.3 hold, and $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is an injective piecewise linear function. Let $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ be a continuous piecewise linear function and $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an injective piecewise linear function. If both following conditions hold,*

$$\mathbb{E}\left\|\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2 = 0, \, and \tag{49}$$

$$\mathbf{g}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y}), \tag{50}$$

for some $\boldsymbol{\mu}_\mathbf{y} \in \mathbb{R}^n, \boldsymbol{\Sigma}_\mathbf{y} \in \mathbb{R}^{n \times n}$, then $\mathbf{Z} \mid (\mathbf{Y} = \mathbf{y})$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y})$ up to affine transformation, i.e., there exists an invertible affine function $\mathbf{h}_\mathbf{Y} : \mathcal{Z}_\mathbf{Y} \to \mathbb{R}^n$, such that $\mathbf{h}_\mathbf{y}(\mathbf{Z}) \mid (\mathbf{Y} = \mathbf{y}) \equiv \mathbf{g}(\mathbf{f}(\mathbf{Z})) \mid (\mathbf{Y} = \mathbf{y})$.

Proof. We start by introducing some additional notation.

Definition B.11. Let $\mathbf{v} : \mathcal{Z} \to \mathbb{R}^n$ be a function with variables $\mathbf{z} = (z_1, ..., z_n) \in \mathcal{Z}$. Let $\mathbf{y} \in [0, 1]^n$ be a n-dimensional binary mask. We define the *sub-support space* as

$$\mathcal{Z}_\mathbf{y} = \{\mathbf{z} \in \mathcal{Z} : \mathbb{1}(z_i = 0) \geq \mathbb{1}(\mathbf{y}_i = 0), \quad \forall i \in [n]\}. \tag{51}$$

Intuitively, this means if $\mathbf{y}_i = 0$, this must cause $\mathbf{z}_i = 0$.

Alternatively, let $\mathbf{s} \subseteq [n]$ be the index set of nonzero elements in $\mathbf{y}$, we can define

$$\mathcal{Z}_{\mathbf{s}} = \{\mathbf{z} \in \mathcal{Z} : \mathbb{1}(z_i = 0) \geq \mathbb{1}(i \notin \mathbf{s}), \quad \forall i \in [n]\}. \tag{52}$$

From the perfect reconstruction constraint (4), we can derive

$$\mathbb{E}||\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))||_2^2 = 0 \tag{53}$$

$$\mathbb{E}||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 = 0 \tag{54}$$

$$\mathbb{E}\left\{\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 \mid \mathbf{Y}\right]\right\} = 0 \tag{55}$$

$$\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 \mid \mathbf{Y}\right] = 0 \quad \mathbb{P}_{\mathbf{Y}}-\text{a.e.} \tag{56}$$

by first substituting $\mathbf{X} = \mathbf{f}(\mathbf{Z})$, then applying the law of total expectation and finally using the fact that the sum of squares is a positive function. Finally $\mathbf{Y}$ is a discrete random variable, in this case $\mathbb{P}_{\mathbf{Y}}$-almost everywhere means everywhere on its support. We now denote $\mathbf{v} := \mathbf{g} \circ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$. Then, following (56), we have for any value $\mathbf{y} \in \mathcal{Y}$ we have

$$\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{v}(\mathbf{Z}))||_2^2 \mid \mathbf{Y} = \mathbf{y}\right] = 0, \tag{57}$$

This means that for the data that satisfy $\mathbf{Y} = \mathbf{y}$, $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\mathbf{v}(\mathbf{Z}))$ are equal $\mathbb{P}_{\mathbf{Z}|\mathbf{Y}}$-almost everywhere, which implies $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\mathbf{v}(\mathbf{Z}))$ are equally distributed. Since by assumption $\mathbf{Z}$ and $\mathbf{v}(\mathbf{Z}) = \mathbf{g}(\mathbf{X})$ are potentially degenerate MVNs, by Theorem B.5, there exists an invertible affine transformation $\mathbf{h}_{\mathbf{y}} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathbf{h}_{\mathbf{y}}(\mathbf{Z}) \equiv \mathbf{v}(\mathbf{Z}). \tag{58}$$

This proves that for data coming from the same mask $\mathbf{Y} = \mathbf{y}$ (potentially unknown), we can identify the masked causal variables mixed through a piecewise linear function up to a linear transformation. $\quad\square$

B.2.3. COMBINING RESULTS AND CONCLUDING THE PROOF OF THM 3.4

Before we prove Theorem 3.4, we first report an adapted version of Lemma D.2 from Kivva et al. (2022), which shows the linearity of the function $\mathbf{f}$ if both $\mathbf{Z}$ and $\mathbf{f}(\mathbf{Z})$ follow the same MVN distribution.

Lemma B.12. *(Linearity of $\mathbf{f}$ for non-degenerate case (Kivva et al., 2022)) Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $|\boldsymbol{\Sigma}| > 0$, i.e. $\mathbf{Z}$ is a non-degenerate MVN. Assume that $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous piecewise affine function such that $\mathbf{f}(\mathbf{Z}) \equiv \mathbf{Z}$, i.e. $\mathbf{f}(\mathbf{Z})$ is equal in distribution to $\mathbf{Z}$. Then $\mathbf{f}$ is affine.*

The original Lemma is proved by contradiction. We extend this result to the case of degenerate multivariate normal variables.

Lemma B.13 (Linearity of $\mathbf{f}$ for (De) MVNs). *Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $|\boldsymbol{\Sigma}| \geq 0$, i.e. $\mathbf{Z}$ is potentially a degenerate multivariate normal. Assume that $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ is a continuous piecewise affine function such that $\mathbf{f}(\mathbf{Z}) \equiv \mathbf{Z}$. Then $\mathbf{f}$ is affine over $\mathcal{Z}$, the support of $\mathbf{Z}$.*

Proof. First, for the simplest case, if $rank(\boldsymbol{\Sigma}) = 0$, then, $\mathcal{Z}$ is a single point $\{\boldsymbol{\mu}\}$, and $\mathbf{f}$ is affine over $\mathcal{Z}$.

If $rank(\boldsymbol{\Sigma}) > 0$, by the definition of (De) MVN, we know that $\mathbf{Z} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}$, where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I})$ with dimension $d_{\boldsymbol{\varepsilon}}$, and $\mathbf{A} \in \mathbb{R}^{n \times d_{\boldsymbol{\varepsilon}}}$ has full column rank. This implies that $\mathcal{Z} = colsp(A) + \mathbf{b}$.

If we substitute $\mathbf{Z} = \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}$ into $\mathbf{f}(\mathbf{Z}) \equiv \mathbf{Z}$, we get

$$\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}) \equiv \mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b} \tag{59}$$

$$\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}) - \mathbf{b} \equiv \mathbf{A}\boldsymbol{\varepsilon} \tag{60}$$

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\left[\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}) - \mathbf{b}\right] \equiv \boldsymbol{\varepsilon} \tag{61}$$

The $\equiv$ symbol denotes that $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\left[\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon} + \mathbf{b}) - \mathbf{b}\right]$ is equal to $\boldsymbol{\varepsilon}$ in distribution, which means it is not necessarily the same in each point of the support.

By Lemma B.12, we can derive the function $(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T[\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}+\mathbf{b})-\mathbf{b}]$ is affine over $\mathbb{R}^{d_\varepsilon}$. Therefore, there exists $\mathbf{M} \in \mathbb{R}^{d_\varepsilon \times d_\varepsilon}$, and $\mathbf{c} \in \mathbb{R}^{d_\varepsilon}$ such that for every $\boldsymbol{\varepsilon}_0 \in \mathbb{R}^{d_\varepsilon}$:

$$(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T[\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}_0+\mathbf{b})-\mathbf{b}] = \mathbf{M}\boldsymbol{\varepsilon}_0 + \mathbf{c} \tag{62}$$

$$\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T[\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}_0+\mathbf{b})-\mathbf{b}] = \mathbf{A}(\mathbf{M}\boldsymbol{\varepsilon}_0 + \mathbf{c}) \tag{63}$$

Since $\mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$ is a projection and its image space is $colsp(A)$, then, it is the identity operator $\mathbf{I}$ on $colsp(A)$.

Since $\mathbf{f}(\mathbf{Z}) \equiv \mathbf{Z}$, then, the support of $\mathbf{f}(\mathbf{Z})$ should be the same as the support of $\mathbf{Z}$, which is $colsp(A) + \mathbf{b}$. This implies $\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}) - \mathbf{b} \in colsp(A)$. Then, we have

$$\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}) - \mathbf{b} = \mathbf{A}(\mathbf{M}\boldsymbol{\varepsilon}_0 + \mathbf{c}) \tag{64}$$

$$\mathbf{f}(\mathbf{A}\boldsymbol{\varepsilon}_0 + \mathbf{b}) = \mathbf{A}(\mathbf{M}\boldsymbol{\varepsilon}_0 + \mathbf{c}) + \mathbf{b}. \tag{65}$$

By definition, $\boldsymbol{\varepsilon}_0 = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{z}_0 - \mathbf{b})$ for every $\mathbf{z}_0 \in \mathcal{Z}$. We substitute this into Equation 65, and we get

$$\mathbf{f}(\mathbf{z}_0) = \mathbf{A}[\mathbf{M}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T(\mathbf{z}_0 - \mathbf{b}) + \mathbf{c}] + \mathbf{b}. \tag{66}$$

Hence, we can conclude $\mathbf{f}$ is a linear function over $\mathcal{Z}$. $\qquad\square$

With these results, we can now prove the following Theorem 3.4.

Theorem 3.4 (Element-wise Identifiability for Piecewise Linear $\mathbf{f}$). *Assume the observation $\mathbf{X}$ follows the data-generating process in Sec 2, Ass. 2.2, 3.2 and 3.3 hold and $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is an injective continuous piecewise linear function. Let $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ be a continuous invertible piecewise linear function and let $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be a continuous invertible piecewise linear function onto its image. If all following conditions hold:*

$$\mathbb{E}\left\|\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2 = 0\,, \tag{4}$$

$$\mathbb{E}\left\|\mathbf{g}(\mathbf{X})\right\|_0 \leq \mathbb{E}\left\|\mathbf{Z}\right\|_0 \ and \tag{5}$$

$$\mathbf{g}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}}) \qquad \forall \mathbf{y} \in \mathcal{Y}, \tag{6}$$

for some $\boldsymbol{\mu}_{\mathbf{y}} \in \mathbb{R}^n, \boldsymbol{\Sigma}_{\mathbf{y}} \in \mathbb{R}^{n \times n}$, then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$, i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation composed with element-wise invertible linear transformations (Def. 2.1).

Proof. Since $\mathbf{X} = \mathbf{f}(\mathbf{Z})$, we can rewrite Equation (4) (perfect reconstruction) as

$$\mathbb{E}||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 = 0\,. \tag{67}$$

This means $\mathbf{f}$ and $\hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}$ are equal $\mathbb{P}_{\mathbf{Z}}$-almost everywhere.

Since $\mathbf{f}$, $\mathbf{g}$ and $\hat{\mathbf{f}}$ are continuous, we can derive $\mathbf{f}$ and $\hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}$ must be equal over the support of $\mathbf{Z}$, $\mathcal{Z}$, i.e.,

$$\mathbf{f}(\mathbf{z}) = \hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}(\mathbf{z})\,, \forall \mathbf{z} \in \mathcal{Z}\,. \tag{68}$$

We denote $\mathbf{v} := \mathbf{g} \circ \mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$, an invertible continuous piecewise affine function. Take the left inverse of $\hat{\mathbf{f}}$ on both sides, then we have,

$$\hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}) = \mathbf{v}(\mathbf{z})\,, \forall \mathbf{z} \in \mathcal{Z}\,. \tag{69}$$

From Lemma B.10, we know that for all $\mathbf{y} \in \mathcal{Y}$, given the mask $\mathbf{Y} = \mathbf{y}$, there exists an invertible affine transformation $\mathbf{h}_{\mathbf{y}} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{v}(\mathbf{Z}) \equiv \mathbf{h}_{\mathbf{y}}(\mathbf{Z})$ over $\mathcal{Z}_{\mathbf{y}}$.

Since we know $\mathbf{Z}|\mathbf{Y} = \mathbf{y}$ follows a (De)MVN and $\mathbf{h}_{\mathbf{y}}$ is affine, by Lemma B.6, we can derive that $\mathbf{h}_{\mathbf{y}}(\mathbf{Z})$ follows a (De-)MVN distribution as well. Then, we can rewrite $\mathbf{v}(\mathbf{Z}) \equiv \mathbf{h}_{\mathbf{y}}(\mathbf{Z})$ in distribution as

$$\mathbf{h}_{\mathbf{y}}(\mathbf{Z}) \equiv \mathbf{v} \circ \mathbf{h}_{\mathbf{y}}^{-1} \circ \mathbf{h}_{\mathbf{y}}(\mathbf{Z}). \tag{70}$$

Since $\mathbf{h_y}$ is affine, its inverse is continuous. In addition, $\mathbf{v}$ is continuous as it is the composition of continuous functions $\mathbf{f}$ and $\mathbf{g}$. Therefore, $\mathbf{v} \circ \mathbf{h_y}^{-1}$ is continuous piecewise affine. By using Lemma B.13, we get that $\mathbf{v} \circ \mathbf{h_y}^{-1}$ is affine over $\mathbf{h_y}(\mathcal{Z_y})$, i.e. there exists an invertible affine map $\mathbf{h}_1$ such that

$$\mathbf{v} \circ \mathbf{h_y}^{-1}(\mathbf{z}) = \mathbf{h}_1(\mathbf{z}) \quad \Longleftrightarrow \quad \mathbf{v}(\mathbf{z}) = \mathbf{h}_1 \circ \mathbf{h_y}(\mathbf{z}) \quad \forall \mathbf{z} \in \mathcal{Z_y}. \tag{71}$$

Since $\mathbf{v}$ is a composition of affine maps, it is also affine on $\mathcal{Z_y}, \forall \mathbf{y} \in \mathcal{Y}$.

We define $\mathcal{Z_s} := \prod_{i=1}^{n}(\mathbb{R} \cdot \mathbb{1}\{i \in \mathbf{s}\})$, which is the space of the supports for each dimension given a value of the support index $\mathbf{s}$. For the dimensions which are included in $\mathbf{s}_i$, this is $\mathbb{R}$, while for the others it is 0. By this definition and by the sufficient variability assumption Ass. 2.2, $\mathcal{Z} = \bigcup_{\mathbf{s} \in \mathcal{S}} \mathcal{Z_s}$.

Since $\mathcal{S}$ is a finite set, which implies countable, we can find a way to order the elements in $\mathcal{S}$, denoted as $\{\mathbf{s}_1, ..., \mathbf{s}_{|\mathcal{S}|}\}$. Thus, $\mathcal{Z} = \bigcup_{i=1}^{|\mathcal{S}|} \mathcal{Z}_{\mathbf{s}_i}$.

While we have already proven that $\mathbf{v}$ is affine over each subspace $\mathcal{Z}_{\mathbf{s}_i}$, we now show that $\mathbf{v}$ is a linear function on $\mathcal{Z}$, i.e. $\mathbf{v}(\mathbf{z}) = \mathbf{w}\mathbf{z}, \forall \mathbf{z} \in \mathcal{Z}$, where $\mathbf{w} \in \mathbb{R}^{n \times n}$.

i We first consider two index sets $\mathbf{s}_1$ and $\mathbf{s}_2$. Without loss of generality, we assume the index in $\mathbf{s}_1$ is from 1 to $|\mathbf{s}_1|$, and the index in $\mathbf{s}_2$ is from m to $m + |\mathbf{s}_2|$, where $m \in \{1, 2, ..., n - |\mathbf{s}_2| + 1\}$. Since $\mathbf{v}$ is affine on $\mathcal{Z}_{\mathbf{s}_1}$ and $\mathcal{Z}_{\mathbf{s}_2}$ individually, we have

$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_1 \mathbf{a}_1 + ... + \mathbf{z}_{|\mathbf{s}_1|}\mathbf{a}_{|\mathbf{s}_1|} + \mathbf{c}_1 \quad \forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_1} \tag{72}$$
$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_m \mathbf{b}_m + ... + \mathbf{z}_{|\mathbf{s}_2|+m}\mathbf{b}_{|\mathbf{s}_2|+m} + \mathbf{c}_2 \quad \forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_2}. \tag{73}$$

Since $\mathcal{Z}_{\mathbf{s}_1} \cap \mathcal{Z}_{\mathbf{s}_2} = \{\mathbf{0}\}$, then we can get $\mathbf{c}_1 = \mathbf{c}_2$.

Case 1. $\mathbf{s}_1 \cap \mathbf{s}_2 = \varnothing$. Then, we have $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_1} \cup \mathcal{Z}_{\mathbf{s}_2}$

$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_1 \mathbf{a}_1 + ... + \mathbf{z}_{|\mathbf{s}_1|}\mathbf{a}_{|\mathbf{s}_1|} + \mathbf{z}_m \mathbf{b}_m + ... + \mathbf{z}_{|\mathbf{s}_2|+m}\mathbf{b}_{|\mathbf{s}_2|+m} + \mathbf{c}_1 \tag{74}$$

Case 2. $\mathbf{s}_1 \cap \mathbf{s}_2 \neq \varnothing$. Without loss of generality, we assume $|\mathbf{s}_2| \leq |\mathbf{s}_1|$.

(a). $\mathbf{s}_2 \subseteq \mathbf{s}_1$. Then, we can directly get $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_1} \cup \mathcal{Z}_{\mathbf{s}_2}$

$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_1 \mathbf{a}_1 + ... + \mathbf{z}_{|\mathbf{s}_1|}\mathbf{a}_{|\mathbf{s}_1|} + \mathbf{c}_1 \tag{75}$$

(b). $\mathbf{s}_2 \nsubseteq \mathbf{s}_1$. Without loss of generality, we assume $\mathbf{s}_1 \cap \mathbf{s}_2 = \{m, ..., m + t\}$, where $t \in \{1, ..., |\mathbf{s}_2|\}$. Then, $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_1} \cap \mathcal{Z}_{\mathbf{s}_2}$, we have

$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_m \mathbf{a}_m + ... + \mathbf{z}_{m+t}\mathbf{a}_{m+t} + \mathbf{c}_1 \tag{76}$$
$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_m \mathbf{b}_m + ... + \mathbf{z}_{m+t}\mathbf{b}_{m+t} + \mathbf{c}_1. \tag{77}$$

This implies $\mathbf{a}_i = \mathbf{b}_i, i = m, ..., m + t$. Therefore, we can derive $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{s}_1} \cup \mathcal{Z}_{\mathbf{s}_2}$,

$$\mathbf{v}(\mathbf{z}) = \mathbf{z}_1 \mathbf{a}_1 + ... + \mathbf{z}_{|\mathbf{s}_1|}\mathbf{a}_{|\mathbf{s}_1|} + \mathbf{z}_{m+t+1}\mathbf{b}_{m+t+1} + ... + \mathbf{z}_{|\mathbf{s}_2|+m}\mathbf{b}_{|\mathbf{s}_2|+m} + \mathbf{c}_1. \tag{78}$$

By considering both cases, we can now proof that $\mathbf{v}$ is a linear function on $\mathcal{Z}_{\mathbf{s}_1} \cup \mathcal{Z}_{\mathbf{s}_2}$.

ii We can iterate this strategy by iteratively adding new $\mathcal{Z}_{\mathbf{s}_i}$ to the union, until $\mathcal{Z}_{|\mathcal{S}|}$. Finally, we have that $\mathbf{v}$ is a linear function on $\bigcup_{\mathbf{s} \in \mathcal{S}} \mathcal{Z_s} = \mathcal{Z}$.

Then, the rest proof immediately follows from Lemma B.3, where we have proven element-wise identifiability for the linear transformation.

$\square$

C. Example: sparsity is not enough for identifiability in the non-linear case.

In Theorem 3.1 we prove that we can achieve element-wise identifiability for an invertible linear mixing function $\mathbf{f}$ in the Partially Observable Causal Representation Learning setting, assuming sufficient support index variability (Ass. 2.2) and under the condition of perfect reconstruction and a sparsity principle.
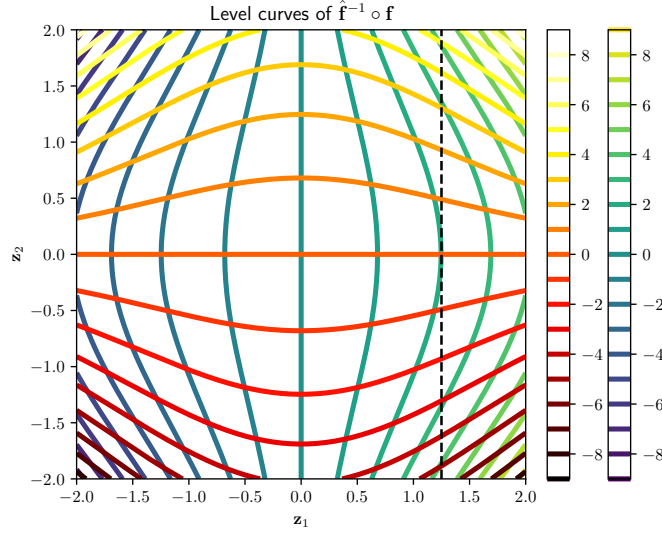
Figure 4: Level curves of the function $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ of Example C.1. The cold color scheme corresponds to the level curves of $(\hat{\mathbf{f}}^{-1} \circ \mathbf{f})_1(\mathbf{z})$ while the warm color scheme corresponds to $(\hat{\mathbf{f}}^{-1} \circ \mathbf{f})_2(\mathbf{z})$. The example gives a concrete case where all assumptions of Theorem 3.1 hold except for the linearity of $\mathbf{f}$. We can see that $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is not a permutation composed with an element-wise invertible transformation, since along the vertical dashed line, we can see that both components of $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ change.

An obvious extension of this result might be considering the identifiability for non-linear mixing functions. In this section we show a counter-example that describes why this is not possible in general without any further assumption (e.g. assuming both a piecewise linear mixing function and a Gaussian causal model, as shown by our results in Thm. 3.4).

Example C.1. Assume we have 2 latent and 2 observed variables, i.e. $n = 2$ and $d = 2$. Furthermore, assume that the domain of the causal variables $\mathcal{C} = \mathbb{R}^2$ and that Ass. 3.2 and 3.3 hold. For example, consider $\mathbf{C} \sim \mathcal{N}(0, \mathbf{I}_2)$ with an independent mask $\mathbf{Y}$ with distribution $p(\mathbf{Y} = \mathbf{y}) = 1/4$ for any $\mathbf{y} \in \{0, 1\}^2$, which trivially satisfies Ass. 3.2 and 3.3. In this case, the support of $\mathbf{Z} = \mathbf{Y} \odot \mathbf{C}$ is $\mathcal{Z} = \mathbb{R}^2$. Assume further that $\mathbf{f} : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by

$$\mathbf{f}(\mathbf{z}) := \sinh(\mathbf{R}_{\frac{\pi}{4}} \mathbf{z}) + \sinh(\mathbf{R}_{-\frac{\pi}{4}} \mathbf{z}), \tag{79}$$

where $\sinh(x) := \frac{e^x - e^{-x}}{2}$ (applied element-wise above) and $\mathbf{R}_\theta$ is a rotation matrix defined by

$$\mathbf{R}_\theta := \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}. \tag{80}$$

Consider $\hat{\mathbf{f}}$ and $\mathbf{g}$ to be the identity function. We now show that all the assumptions of Theorem 3.1 except the linearity of $\mathbf{f}$ are satisfied. First, notice how

$$\mathbb{E}||\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))||^2 = \mathbb{E}||\mathbf{X} - \mathbf{X}||^2 = 0. \tag{81}$$

Since $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ by definition and $\mathbf{g}(\mathbf{X}) = \mathbf{X}$, since $\mathbf{g}$ is the identity function, then $\mathbf{f}(\mathbf{Z}) = \mathbf{g}(\mathbf{X})$ and thus $\mathbb{E}||\mathbf{g}(\mathbf{X})||_0 = \mathbb{E}||\mathbf{f}(\mathbf{Z})||_0$.

We now show that $\mathbb{E}||\mathbf{f}(\mathbf{Z})||_0 \le \mathbb{E}||\mathbf{Z}||_0$, which we will then use to prove the sparsity condition on $\mathbb{E}||\mathbf{g}(\mathbf{X})||_0$.

We can see that

$$\mathbf{f}(\begin{bmatrix} z_1 \\ 0 \end{bmatrix}) = \sinh(z_1 \begin{bmatrix} \cos\frac{\pi}{4} \\ \sin\frac{\pi}{4} \end{bmatrix}) + \sinh(z_1 \begin{bmatrix} \cos -\frac{\pi}{4} \\ \sin -\frac{\pi}{4} \end{bmatrix}) \tag{82}$$

$$= \sinh(z_1 \begin{bmatrix} \cos\frac{\pi}{4} \\ \sin\frac{\pi}{4} \end{bmatrix}) + \sinh(z_1 \begin{bmatrix} \cos\frac{\pi}{4} \\ -\sin\frac{\pi}{4} \end{bmatrix}) \tag{83}$$

$$= \begin{bmatrix} 2\sinh(z_1 \cos\frac{\pi}{4}) \\ \sinh(z_1 \sin\frac{\pi}{4}) + \sinh(-z_1 \sin\frac{\pi}{4}) \end{bmatrix} \tag{84}$$

$$= \begin{bmatrix} 2\sinh(z_1 \cos\frac{\pi}{4}) \\ \sinh(z_1 \sin\frac{\pi}{4}) - \sinh(z_1 \sin\frac{\pi}{4}) \end{bmatrix} = \begin{bmatrix} 2\sinh(z_1 \cos\frac{\pi}{4}) \\ 0 \end{bmatrix}, \tag{85}$$

where we use the fact that $\sin$ and $\sinh$ are odd, while $\cos$ is even. An analogous argument shows that

$$\mathbf{f}(\begin{bmatrix} 0 \\ z_2 \end{bmatrix}) = \begin{bmatrix} 0 \\ 2\sinh(z_2 \cos\frac{\pi}{4}) \end{bmatrix}. \tag{86}$$

We can also easily show that $\mathbf{f}(0) = 0$. The above shows that, for all $\mathbf{z} \in \mathbb{R}^2$, $||\mathbf{f}(\mathbf{z})||_0 \leq ||\mathbf{z}||_0$. By taking the expectation on both sides we get the desired result: $\mathbb{E}||\mathbf{f}(\mathbf{Z})||_0 \leq \mathbb{E}||\mathbf{Z}||_0$ and thus $\mathbb{E}||\mathbf{g}(\mathbf{X})||_0 \leq \mathbb{E}||\mathbf{Z}||_0$.

However, we can see in in Figure 4 and in the computation below, $\hat{\mathbf{f}}^{-1} \circ \mathbf{f} = \mathbf{f}$ is not a permutation composed with an element-wise invertible transformation on $\mathbb{R}^2$.

$$\mathbf{f}(\begin{bmatrix} z_1 \\ z_2 \end{bmatrix}) = \begin{bmatrix} \sinh(\cos\frac{\pi}{4}z_1 - \sin\frac{\pi}{4}z_2) + \sinh(\cos\frac{\pi}{4}z_1 + \sin\frac{\pi}{4}z_2) \\ \sinh(\sin\frac{\pi}{4}z_1 + \cos\frac{\pi}{4}z_2) - \sinh(\sin\frac{\pi}{4}z_1 - \cos\frac{\pi}{4}z_2) \end{bmatrix} \tag{87}$$

This counter-example shows that the results of Theorem 3.1 do not apply in general if $\mathbf{f}$ is nonlinear, but only with additional assumptions, as shown in Theorem 3.4.

D. Implementation details

This section provides further details about the experiment implementation in Section 4. The implementation is built upon the code open-sourced by Lachapelle et al. (2022) released under Apache 2.0 License; Ahuja et al. (2022); von Kügelgen et al. (2021) released under MIT License; Zheng et al. (2018) released under Apache 2.0 License.

D.1. Hyperparameters

	Numerical linear $\mathbf{f}$ Section 5.1	Numerical p.w. linear $\mathbf{f}$ Section 5.1	Multiple balls Section 5.2	PartialCausal3DIdent Section 5.3
ϵ	0.001	0.01	0.01	0.01
Optimizer	ExtraAdam	ExtraAdam	ExtraAdam	ExtraAdam
Primal optimizer learning rate	1e-4	5e-5	1e-5	1e-4
Dual optimizer learning rate	1e-4/2	5e-5/2	1e-5/2	1e-4/2
Batch size	6144	10000	$50 \times K$	$13 \times K$
Group number K	2^n	latent size $5n$	latent size n	10
# Seeds	20	20	3	3
# Iterations	30000	20000	10000	10000

Table 4: Parameters for experiments results in Sec. 5 and App. E.

D.2. Sensitivity analysis on ϵ

We conduct a sensitivity analysis on ϵ. As shown in Figure 5, we observe that the linear case exhibits heightened sensitivity ϵ, whereas the piecewise linear case does not display such sensitivity. We attribute this difference to the additional group and mask information given in the training phase for piecewise linear case, enhancing the method's robustness to variations in this hyperparameter.
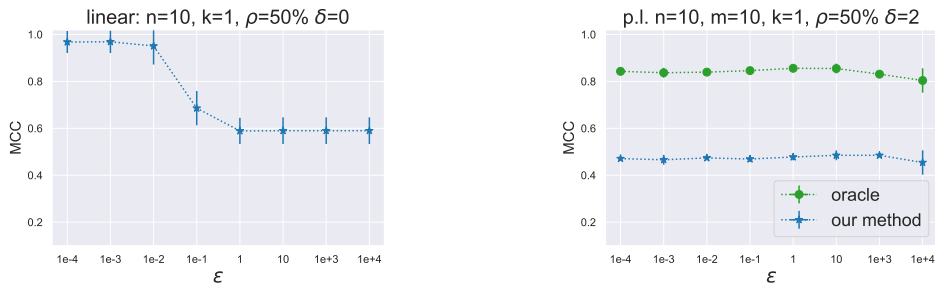
Figure 5: Sensitivity analysis on ϵ varies from $1e-4$ to $1e4$. The left graph is for the linear case, and right-hand side is for the piecewise linear case.

D.3. Oracle method

In Sec. 4, to encourage Gaussianity of learned representations, we add a regularization term to force the sample skewness and kurtosis to match the Gaussian distribution. However, as we mention in Sec. 4, estimated skewness and kurtosis cannot guarantee Gaussianity. In Figure 6, we empirically show even by adding the two penalty terms into the loss function, the estimator we obtain is still highly non-gaussian.

In the Oracle method, we adopt the assumption of knowing masks in train phases. Instead of directly estimating the unmeasured part as a constant value, we provide less information by replacing $\mathbf{g}_\psi(\mathbf{x}_i)$ with a low log standard deviation -10 in Eq. 7; for the measured latent variables, we assign 1 to $\mathbf{g}_\psi(\mathbf{x}_i)$. After training, we obtain the encoder function $\mathbf{g}$. In the test phase, data in distinct groups can be mixed together. There is no longer a requirement for mask-related information. In addition, we find empirically that if we replace $\hat{\boldsymbol{\mu}}_g$ with a constant value, e.g., 2, it also enhances the performance for most setups except for the case when $\delta = 0$. When $\delta = 0$, set the constant value to be 0 can obtain better results.
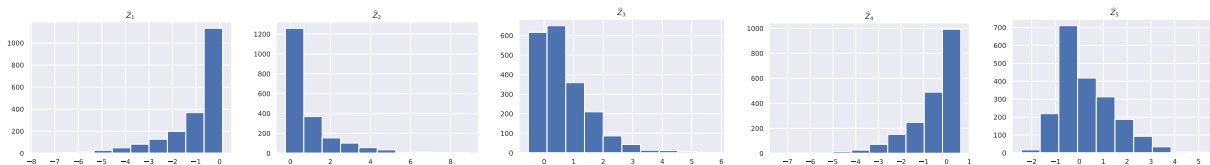

Figure 6: Histogram of estimator for our method for $n = 5$, $m = 10$, independent $\mathbf{Z}$, ρ=50%,

D.4. Predefined masks

In order to satisfy Ass. 2.2, i.e. for each latent variable C_i $i \in [n]$, all the other latent variables should be measured in at least one of the groups where C_i is unmeasured, the minimum number of distinct masks is equal to the latent size n, and the maximum number is up to 2^n. In this paper, we consider 3 different strategies to generate masks.

i Consider all 2^n possible masks as a mask set $\mathbf{Y}$. For each data $\mathbf{c}^i$, $i \in [N]$, we uniformly sample one mask $\mathbf{y}^i$ from the mask set $\mathbf{Y}$. This strategy is used in numerical experiments with linear $\mathbf{f}$ since we do not need group-wise masks.

ii Define kn possible masks $\{\mathbf{y}_1, .., \mathbf{y}_n\}$ with the same ratio of measured variables ρ. Define $\mathbf{y}_g$, $g = [n]$ by this way: randomly set ρn elements as 1, the rest are 0s. This strategy is used in numerical experiments with piecewise linear $\mathbf{f}$ ($k = 5$) and multiple balls experiment ($k = 1$) due to the necessity of having a specific number of samples from the same group in each batch for the computation of sample skewness and kurtosis—something strategy i) is incapable of achieving.

iii Define n possible masks $\{\mathbf{y}_1, .., \mathbf{y}_n\}$ with different ratio of measured variables ρ_g, $g = [n]$ from 1 var to 100%. Define $\mathbf{y}_g$ this way: set 1-th to $(g + \rho_g n)$-th element as 1, the rest are 0s. This is employed in the PartialCausal3DIdent experiment.

E. Full experimental results

E.1. Numerical experiments

E.1.1. LINEAR MIXING FUNCTION

We first show the Pearson Correlation matrix $\text{Corr}_{\pi}^{n \times n}$ with the permutation π between ground truth latent variables $\mathbf{Z}$ and the estimator $\hat{\mathbf{Z}} = \mathbf{g}(\mathbf{X})$. One figure represents one ablation study in Table 1. We choose one random seed to plot for each setup. Ground truth Pear. Corr. matrix on the left shows the original linear correlation inside $\mathbf{Z}$, compared with the estimator on the right-hand side.
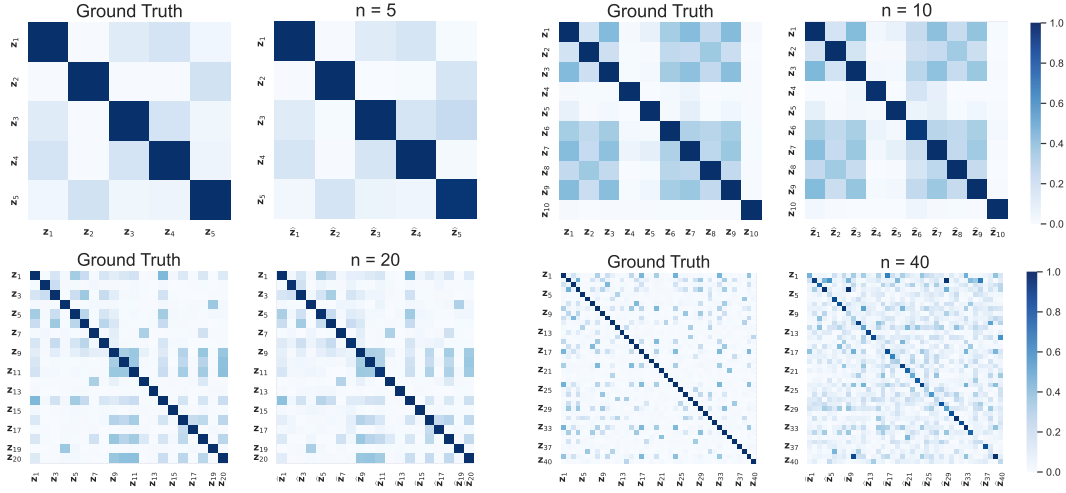
Figure 7: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the latent dimension n from 5 to 40 and fixing $\delta = 0\sigma$, $\rho = 50$, $k = 1$. For all methods it it harder to learn an increasing number of latent variables.
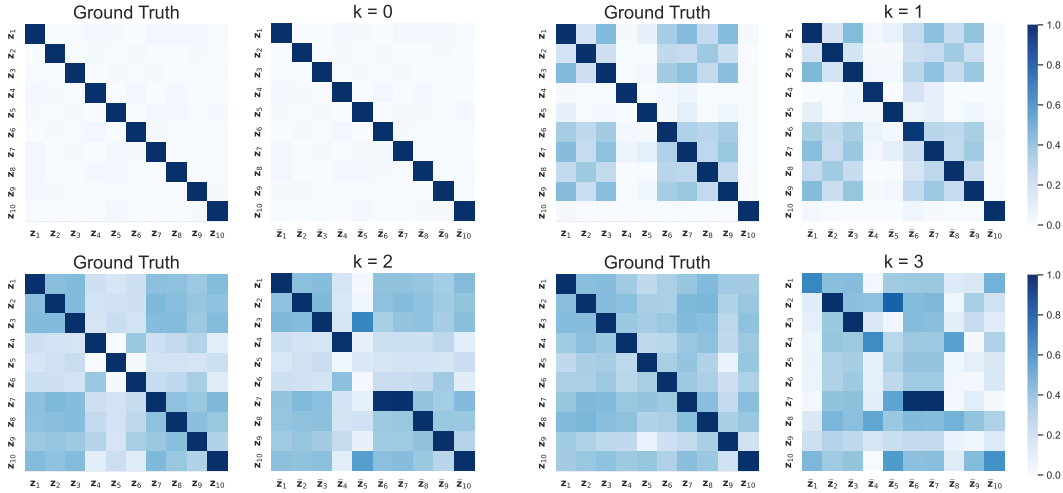

Figure 8: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$. For all methods it is harder to learn the latent variables in denser graphs.
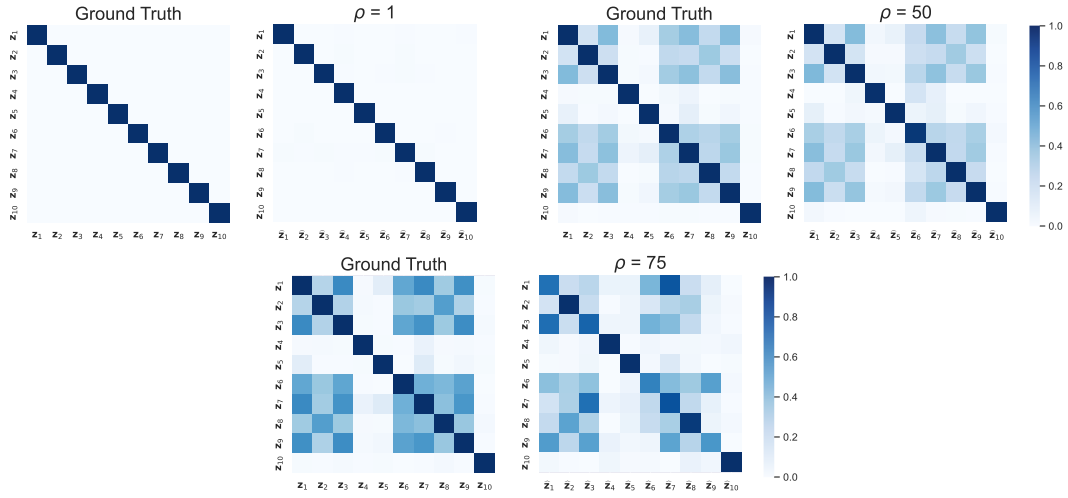
Figure 9: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the ratio of active (unmasked) variables ρ from 1 variable only to 75% and fixing $n = 10$, $\delta = 0\sigma$, $k = 1$. Learning the latent variables is harder when there are many active variables.
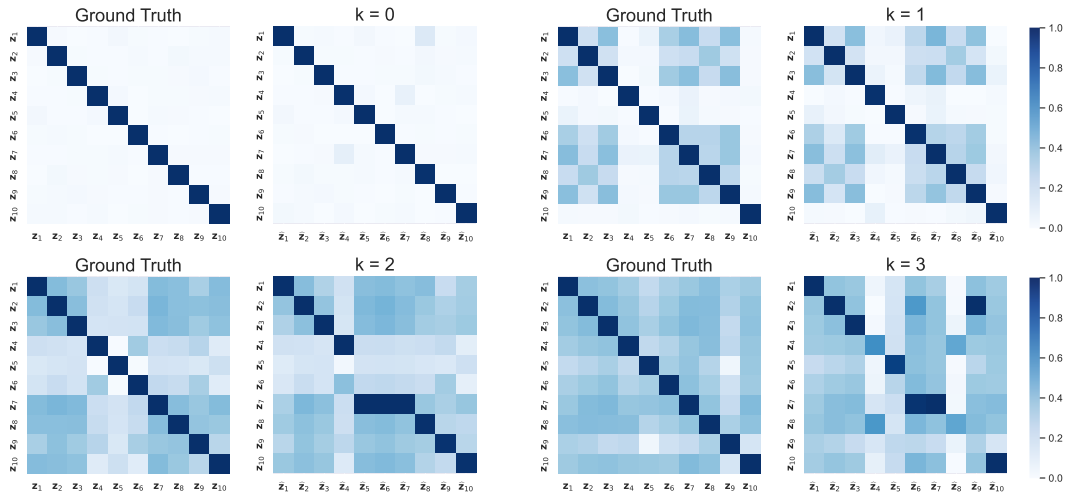

Figure 10: Linear mixing function with linear causal relation and Exponential noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$. For all methods it is harder to learn the latent variables for denser graphs.
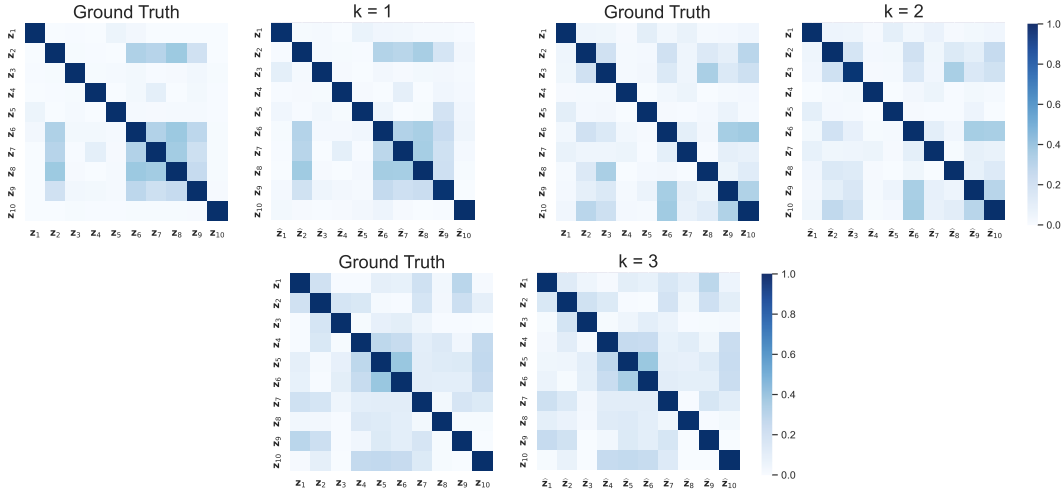
Figure 11: Linear mixing function with nonlinear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$.

To exclude the natural dependency between latent variables that potentially enhance the value of metric MCC, we provide additional experiment results with *independent* latent variables using 20 random seeds in Table 5.

Table 5: Results for the numerical experiments in the linear case when $\mathbf{Z}$ are independent.

n	CAUSAL GRAPH	ρ	MCC
5	INDEPENDENT	50 %	0.998 ± 0.002
10	INDEPENDENT	50 %	0.998 ± 0.001
20	INDEPENDENT	50 %	0.997 ± 0.002
40	INDEPENDENT	50 %	0.748 ± 0.145
10	INDEPENDENT	1 VAR	0.998 ± 0.001
10	INDEPENDENT	50 %	0.998 ± 0.001
10	INDEPENDENT	75 %	0.998 ± 0.002

E.1.2. PIECEWISE LINEAR MIXING FUNCTION

In Table 6 and Table 7, we provide the numerical results of averaged MCC and standard deviation over 20 random seeds as shown in Figure 2. In the second block of Table 6, we provide two additional results of our method (training without mask information) when $\mathbf{Z}$ are independent. In Table 7, the second right column is for training without mask information, and the most right column is the oracle method, i.e., providing mask information in the training phase. Additionally, to exclude the natural dependency between latent variables that potentially enhance the value of metric MCC, we provide additional experiment results with *independent* latent variables in Table 8.

Table 6: Results for the numerical experiments in the piecewise linear case for simple settings.

n	m	k	ρ	δ	**MCC** (MEAN $\pm$ STD)
3	3	0	50 %	0	0.630$\pm$0.064
3	3	0	50 %	2	0.755$\pm$0.068
3	3	0	50 %	3	0.769$\pm$0.075
3	3	0	50 %	5	0.813$\pm$0.070
3	3	0	50 %	10	0.872$\pm$0.071
3	3	1	50 %	0	0.649$\pm$0.059
3	3	1	50 %	2	0.756$\pm$0.065
3	3	1	50 %	3	0.783$\pm$0.082
3	3	1	50 %	5	0.836$\pm$0.091
3	3	1	50 %	10	0.877$\pm$0.074

Table 7: Results for the numerical experiments in the piecewise linear case for the *dependent* variable case.

n	m	k	ρ	δ	**MCC** (MEAN $\pm$ STD)	**MCC** (MEAN $\pm$ STD) ORACLE
5	10	1	50 %	2	0.469$\pm$0.050	0.898$\pm$0.018
10	10	1	50 %	2	0.349$\pm$0.030	0.799$\pm$0.020
20	10	1	50 %	2	0.280$\pm$0.017	0.753$\pm$0.015
40	10	1	50 %	2	0.217$\pm$0.025	0.782$\pm$0.012
10	10	1	50 %	2	0.349$\pm$0.030	0.799$\pm$0.020
10	10	2	50 %	2	0.337$\pm$0.030	0.789$\pm$0.021
10	10	3	50 %	2	0.345$\pm$0.028	0.785$\pm$0.022
10	3	1	50 %	2	0.466$\pm$0.038	0.867$\pm$0.007
10	10	1	50 %	2	0.349$\pm$0.030	0.799$\pm$0.020
10	20	1	50 %	2	0.267$\pm$0.031	0.552$\pm$0.035
10	10	1	1VAR	2	0.402$\pm$0.027	0.908$\pm$0.013
10	10	1	50 %	2	0.349$\pm$0.030	0.799$\pm$0.020
10	10	1	75 %	2	0.352$\pm$0.037	0.695$\pm$0.026
10	10	1	50 %	0	0.332$\pm$0.040	0.185$\pm$0.036
10	10	1	50 %	3	0.368$\pm$0.028	0.856$\pm$0.015
10	10	1	50 %	5	0.388$\pm$0.029	0.930$\pm$0.010
10	10	1	50 %	10	0.405$\pm$0.027	0.980$\pm$0.007

Table 8: Results for the numerical experiments in the piecewise linear case for the *independent* variable case.

n	m	k	ρ	δ	**MCC** (MEAN $\pm$ STD)	**MCC** (MEAN $\pm$ STD) ORACLE
5	10	0	50 %	2	0.462$\pm$0.065	0.910$\pm$0.014
10	10	0	50 %	2	0.352$\pm$0.028	0.807$\pm$0.022
20	10	0	50 %	2	0.296$\pm$0.018	0.732$\pm$0.011
40	10	0	50 %	2	0.246$\pm$0.013	0.746$\pm$0.015
10	3	0	50 %	2	0.523$\pm$0.026	0.865$\pm$0.008
10	10	0	50 %	2	0.352$\pm$0.028	0.807$\pm$0.022
10	20	0	50 %	2	0.269$\pm$0.023	0.519$\pm$0.028
10	10	0	1VAR	2	0.412$\pm$0.027	0.906$\pm$0.016
10	10	0	50 %	2	0.352$\pm$0.028	0.807$\pm$0.022
10	10	0	75 %	2	0.340$\pm$0.028	0.719$\pm$0.017
10	10	0	50 %	0	0.318$\pm$0.026	0.236$\pm$0.040
10	10	0	50 %	3	0.366$\pm$0.029	0.852$\pm$0.020
10	10	0	50 %	5	0.388$\pm$0.027	0.926$\pm$0.014

For each setup in Table 7, we choose one random seed to plot the heatmap of the Pearson Correlation matrix $\mathrm{Corr}_\pi^{n \times n}$ with the permutation π. One figure represents one ablation study of one parameter. Ground truth Pear. Corr. matrix on the left shows the original linear correlation inside Z, compared with the estimator on the right-hand side.
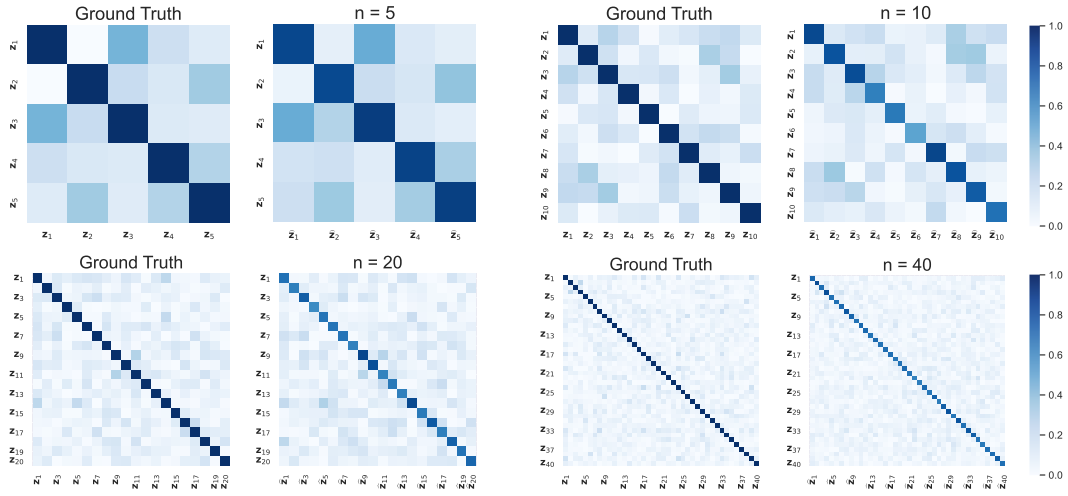
Figure 12: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the latent dimension n from 5 to 40 and fixing $\delta = 2.0\sigma$, $m = 10$, $\rho = 50$, $k = 1$. The case with higher n is more complicated to learn the latent variables.
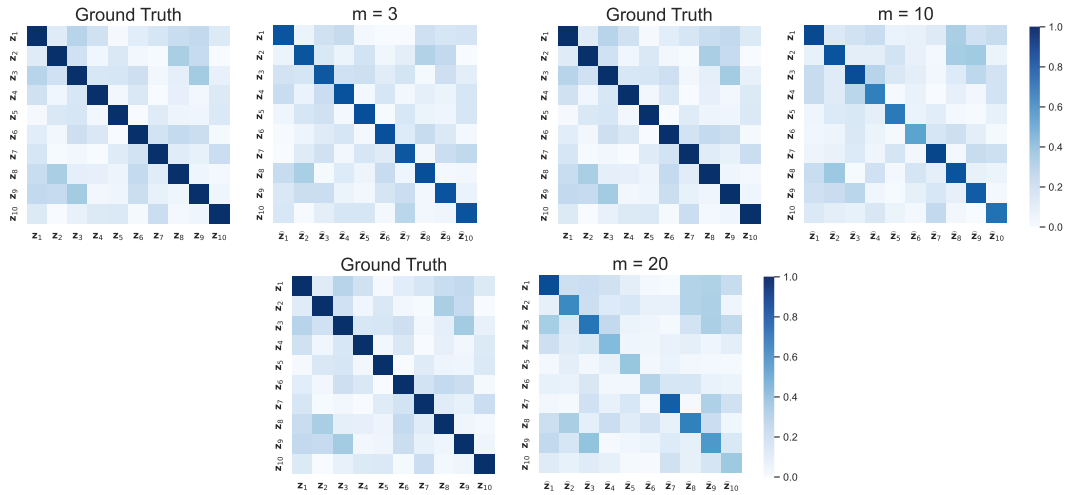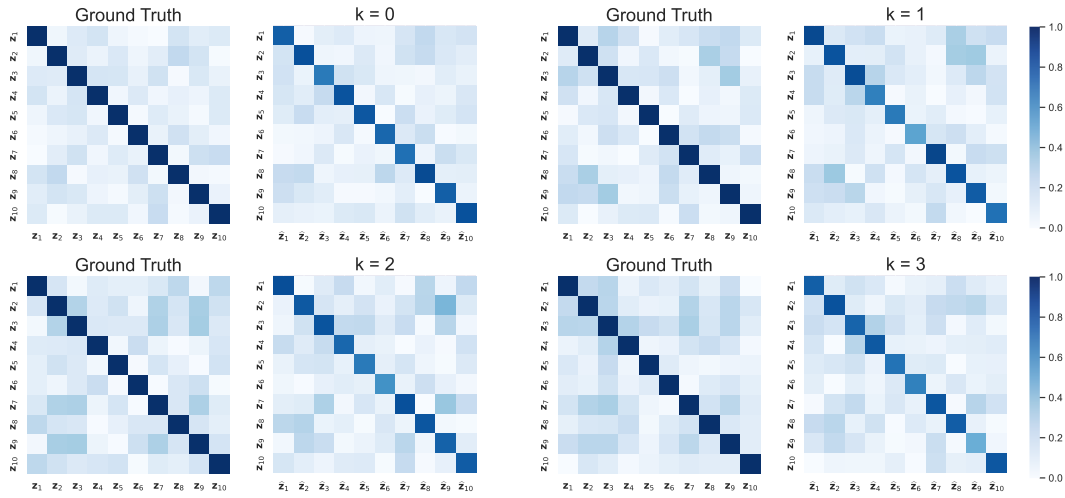

Figure 13: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the number of Leaky-ReLU layers $(m - 1)$ from 3 to 20 and fixing $n = 10$, $\delta = 2.0\sigma$, $\rho = 50$, $k = 1$. The case with larger m is more complicated to learn the latent variables due to a greater extent of nonlinearity.

Figure 14: Piecewise linear mixing function with nonlinear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $m = 10$, $\delta = 2.0\sigma$, $\rho = 50$. In the case with a denser graph, it is more complicated to learn the latent variables.
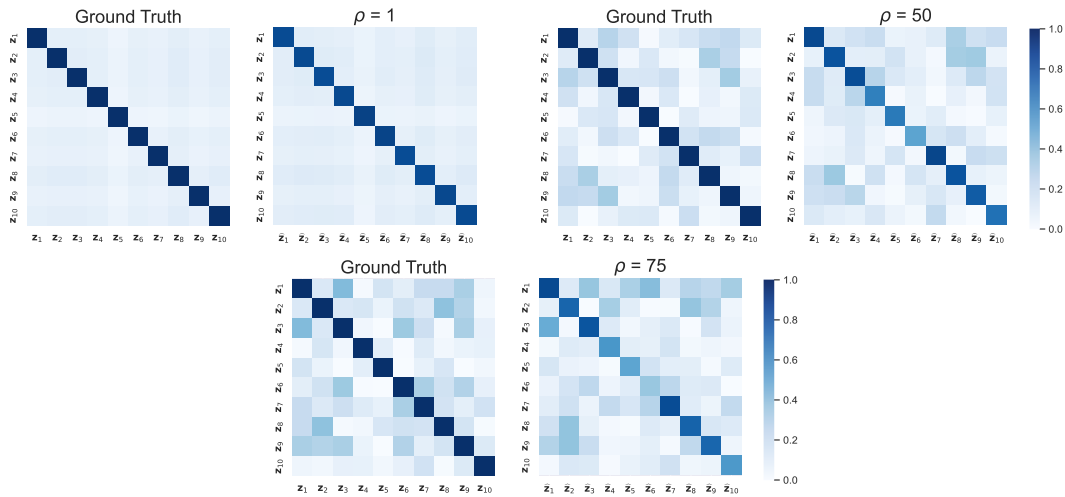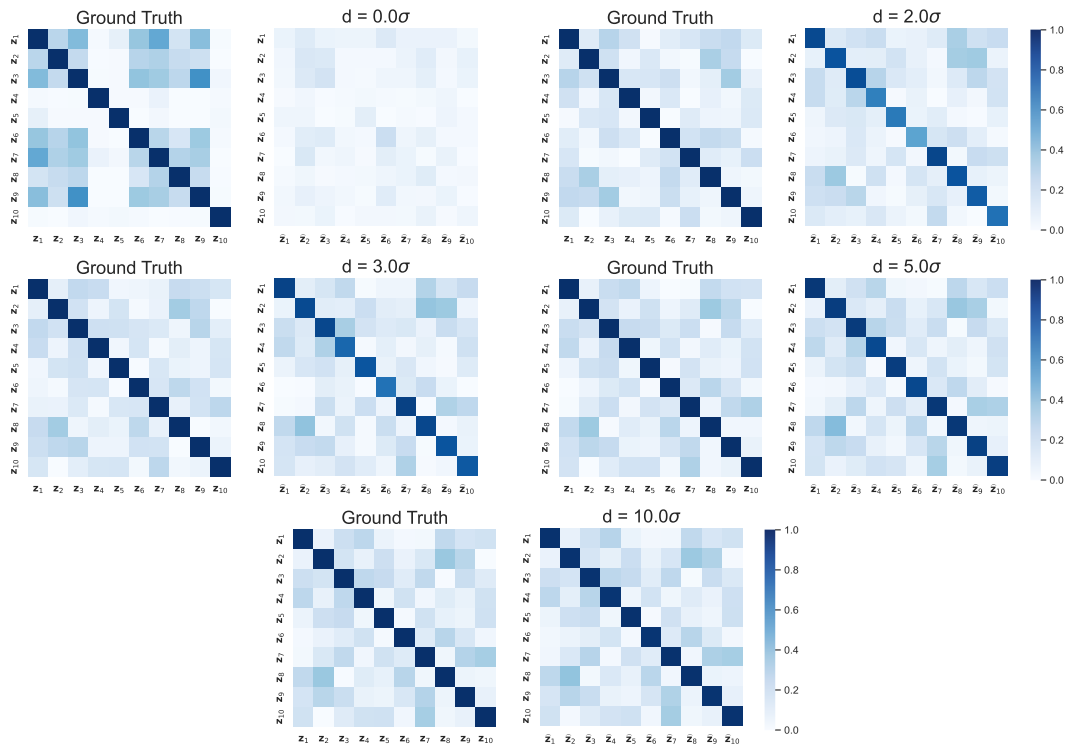

Figure 15: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the ratio of active (unmasked) variables ρ from 1 variable only to 75% and fixing $n = 10$, $m = 10$, $\delta = 2.0\sigma$, $k = 1$. Learning the latent variables is more complicated in the case with a larger portion of active variables.

Figure 16: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the distance between masked value and mean of latents from 0.0σ to 10.0σ, where σ is the standard deviation of latents, and fixing $n = 10$, $m = 10$, $\rho = 50$, $k = 1$. Learning the latent variables is more complicated in the case with a smaller distance.

E.1.3. TEST ON INDEPENDENT LATENT VARIABLES

To exclude the effect of the inherent causal relation, which might increase the value of metric MCC, we additionally test our estimated encoder **g** when the unmasked causal variables are independent with each other, and implement the same ablation study as Table 1 for linear mixing function and Table 7 for piecewise linear mixing function. As we can see, in Table 9, comparing the right two columns, there is no significant difference, which provides evidence that the MCC obtained by our method does not come from the inherent causal relation among latents.

Table 9: Results of testing on *independent* latents for linear mixing functions with $\delta = 0$. The bold font indicates which parameters are varying in each block of rows.

| n	k	SCM	ρ	MCC (TABLE 1)	MCC (TEST ON INDEPENDENT LATENTS)
5	1	LIN. GAUSS	50 %	0.997±0.002	0.991±0.010
10	1	LIN. GAUSS	50 %	0.996±0.001	0.992±0.004
20	1	LIN. GAUSS	50 %	0.987±0.029	0.985±0.031
40	1	LIN. GAUSS	50 %	0.714±0.153	0.718±0.142
10	**0**	INDEP. GAUSS	50 %	0.998±0.001	0.998±0.010
10	**1**	LIN. GAUSS	50 %	0.996±0.001	0.992±0.004
10	**2**	LIN. GAUSS	50 %	0.904±0.113	0.893±0.122
10	**3**	LIN. GAUSS	50 %	0.793±0.142	0.744±0.179
10	**0**	INDEP. EXP	50 %	0.998±0.001	0.998±0.001
10	**1**	LIN. EXP	50 %	0.998±0.002	0.996±0.002
10	**2**	LIN. EXP	50 %	0.910±0.108	0.895±0.123
10	**3**	LIN. EXP	50 %	0.825±0.123	0.776±0.156
10	**1**	NONLINEAR	50 %	0.997±0.001	0.994±0.005
10	**2**	NONLINEAR	50 %	0.997±0.001	0.996±0.003
10	**3**	NONLINEAR	50 %	0.996±0.001	0.994±0.004
10	1	LIN. GAUSS	**1VAR**	0.998 ±0.002	0.993±0.005
10	1	LIN. GAUSS	**50 %**	0.996±0.001	0.992±0.004
10	1	LIN. GAUSS	**75 %**	0.877 ±0.096	0.925±0.068
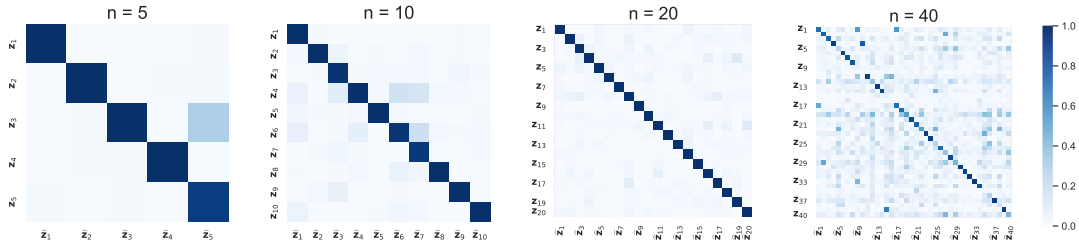
Figure 17: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the latent dimension n from 5 to 40 and fixing $\delta = 0\sigma$, $\rho = 50$, $k = 1$. The case with higher n is more complicated to learn the latent variables.
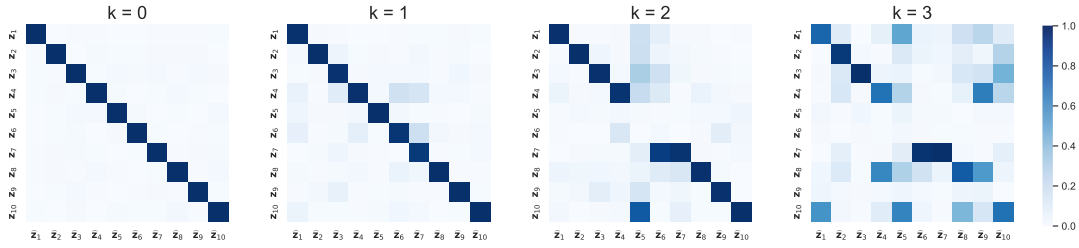
Figure 18: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$. The case with a denser graph is more complicated to learn the latent variables.
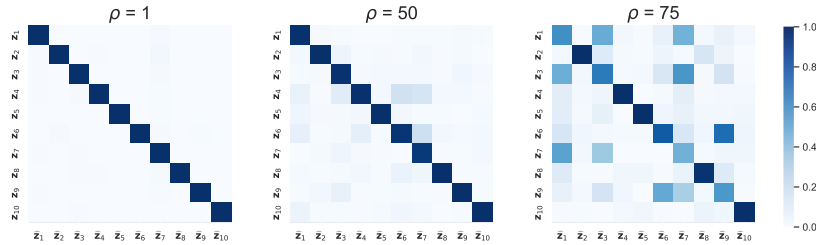

Figure 19: Linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the ratio of active (unmasked) variables ρ from 1 variable only to 75% and fixing $n = 10$, $\delta = 0\sigma$, $k = 1$. Learning the latent variables is more complicated in the case with a larger portion of active variables.
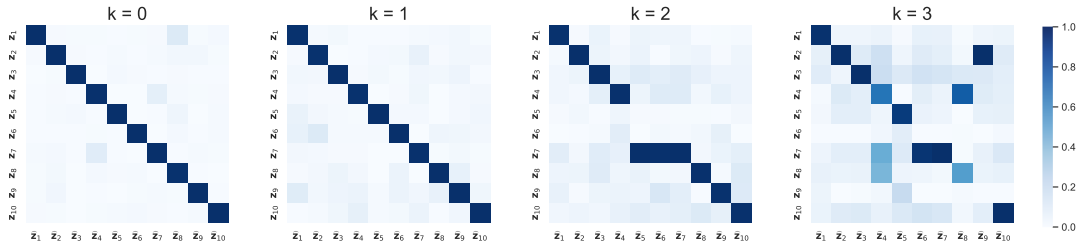

Figure 20: Linear mixing function with linear causal relation and Exponential noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$. The case with a denser graph is more complicated to learn the latent variables.
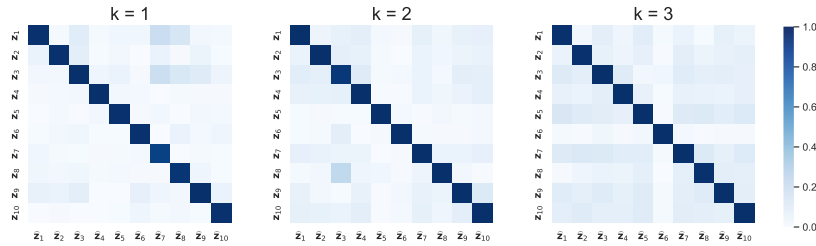

Figure 21: Linear mixing function with nonlinear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $\delta = 0\sigma$, $\rho = 50$. The case with a denser graph is more complicated to learn the latent variables.

Table 10: Results for the numerical experiments in the piecewise linear case for training on *dependent* variable but testing on *independent* variable case.

| n	m	k	ρ	δ	MCC(TABLE 7)	MCC(TEST ON INDEPENDENT LATENTS)
5	10	1	50 %	2	0.898±0.018	0.923±0.021
10	10	1	50 %	2	0.799±0.020	0.807±0.023
20	10	1	50 %	2	0.753±0.015	0.800±0.022
40	10	1	50 %	2	0.782±0.012	0.869±0.013
10	10	1	50 %	2	0.799±0.020	0.807±0.023
10	10	2	50 %	2	0.789±0.021	0.836±0.036
10	10	3	50 %	2	0.785±0.022	0.847±0.036
10	3	1	50 %	2	0.867±0.007	0.923±0.034
10	10	1	50 %	2	0.799±0.020	0.807±0.023
10	20	1	50 %	2	0.552±0.035	0.537±0.035
10	10	1	1VAR	2	0.908±0.013	0.946±0.011
10	10	1	50 %	2	0.799±0.020	0.807±0.023
10	10	1	75 %	2	0.695±0.026	0.678±0.030
10	10	1	50 %	0	0.185±0.036	0.232±0.038
10	10	1	50 %	3	0.856±0.015	0.881±0.018
10	10	1	50 %	5	0.930±0.010	0.949±0.009
10	10	1	50 %	10	0.980±0.007	0.986±0.006
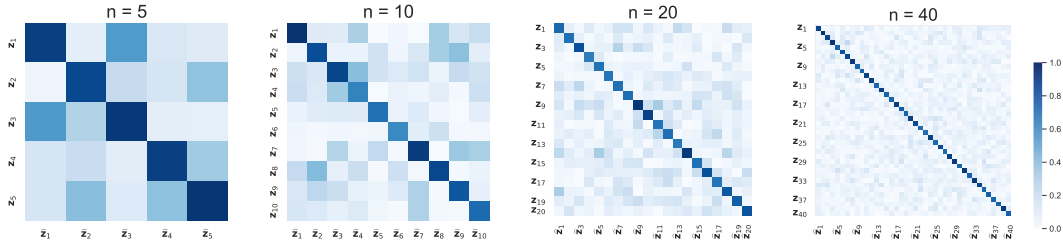
Figure 22: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the latent dimension n from 5 to 40 and fixing $\delta = 2.0\sigma$, $m = 10$, $\rho = 50$, $k = 1$. The case with higher n is more complicated to learn the latent variables.
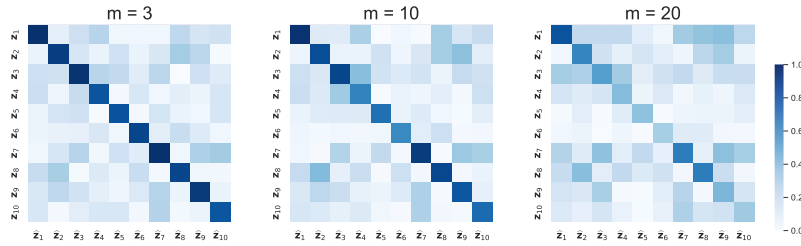

Figure 23: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the number of Leaky-ReLU layers $(m-1)$ from 3 to 20 and fixing $n = 10$, $\delta = 2.0\sigma$, $\rho = 50$, $k = 1$. The case with larger m is more complicated to learn the latent variables due to a greater extent of nonlinearity.
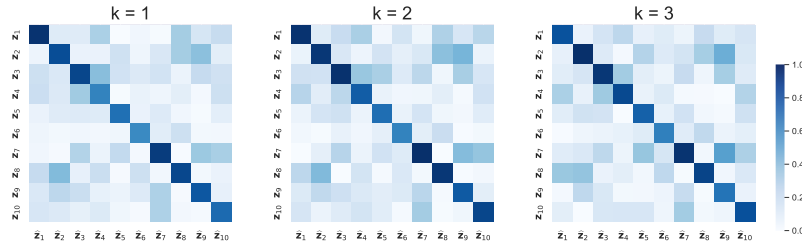
Figure 24: Piecewise linear mixing function with nonlinear causal relation and Gaussian noise: ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $m = 10$, $\delta = 2.0\sigma$, $\rho = 50$. In the case with a denser graph, it is more complicated to learn the latent variables.
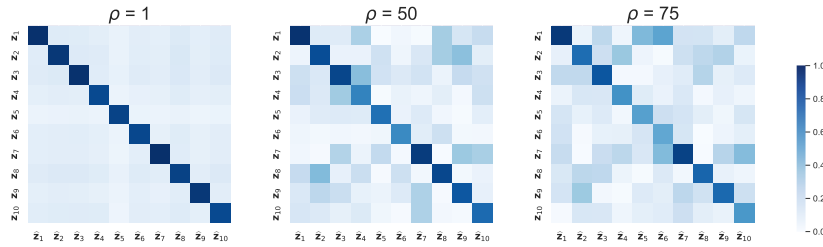

Figure 25: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the ratio of active (unmasked) variables ρ from 1 variable only to 75% and fixing $n = 10$, $m = 10$, $\delta = 2.0\sigma$, $k = 1$. Learning the latent variables is more complicated in the case with a larger portion of active variables.
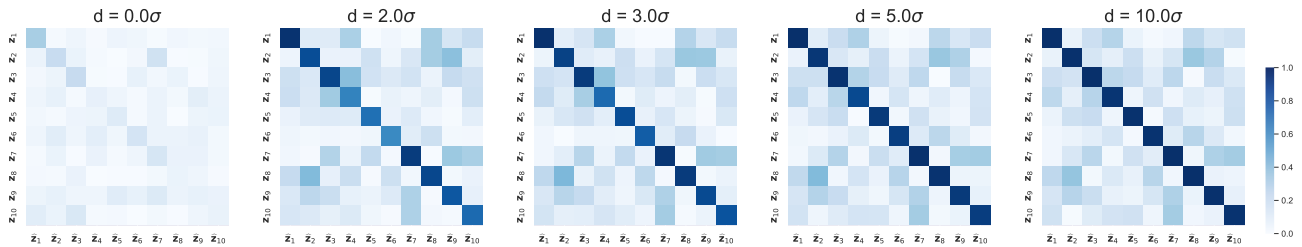

Figure 26: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the distance between masked value and mean of latents from 0.0σ to 10.0σ, where σ is the standard deviation of latents, and fixing $n = 10$, $m = 10$, $\rho = 50$, $k = 1$. Learning the latent variables is more complicated in the case with a smaller distance.

E.1.4. OVER-PARAMETERIZATION OF THE NUMBER OF CAUSAL REPRESENTATIONS

Our theoretical result assume that we know the number of causal variables in advance, but in practice (and as often also the case in other CRL methods), we can show empirically that if we overestimate the true number of latent variables, we will still be able to identify the true causal variables, while some of them will be unused. We provide ablation study below, which seem to confirm this empirically.

In the experiments below we consider $n = 10$ ground truth causal variables, and nn latent variables that we estimate, where $nn \geq n$. For all experiments, we consider an average over 3 random seeds. We use a causal graph ER-k, where ER-k is a graph with $n \cdot k$ edges, δ is the distance between mask value and mean of causal variables and ρ denotes the ratio of measured variables. For the piecewise linear case m denotes the number of hidden layers in the MLP. We choose one representative setting for the linear case and one representative setting for the piecewise linear case, and report how the MCC varies when we increase nn from the original size $nn = n$ to double the size $nn = 2n$. We report the average MCC over 3 random seeds and a heatmap of the correlation matrix for each of the settings below.

$nn(n = 10)$	$10(n+0)$	$12(n+2)$	$14(n+4)$	$16(n+6)$	$18(n+8)$	$20(n+10)$
MCC	0.998±0.001	0.931±0.056	0.965±0.051	0.991±0.001	0.996±0.001	0.978±0.031

Table 11: Average MCCs over 3 random seeds for linear $\mathbf{f}$ with $n = 10$, $\delta = 0.0\sigma$, ρ=50, $k = 1$, Linear Gaussian SCM

$nn(n = 10)$	$10(n+0)$	$12(n+2)$	$14(n+4)$	$16(n+6)$	$18(n+8)$	$20(n+10)$
MCC	0.796±0.026	0.798±0.012	0.804±0.015	0.809±0.012	0.814±0.014	0.816±0.012

Table 12: Average MCCs over 3 random seeds for piecewise $\mathbf{f}$ with $n = 10$, $\delta = 2.0\sigma$, ρ=50, $k = 1$, $m = 10$, Linear Gaussian SCM

As shown in both Table 11 and Table 12, our method seems to perform similarly well, even if we do not know the true number of causal variables, but we overestimate them. We provide the heatmaps of the correlation matrix between $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ for one of the seeds with different nn's in Figure 27 and Figure 28.
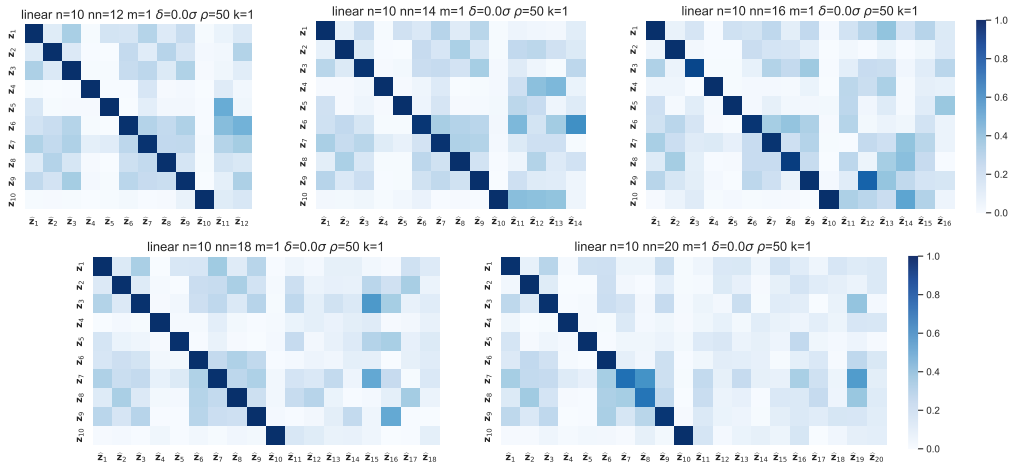
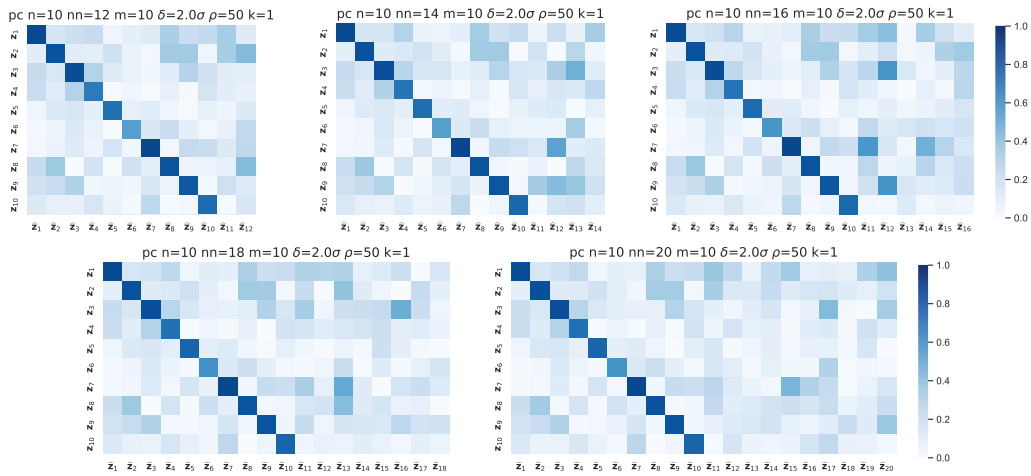Figure 27: Ablation study on number of overestimated representations nn for linear $\mathbf{f}$

Figure 28: Ablation study on the number of overestimated representations nn for piecewise linear $\mathbf{f}$

E.1.5. Causal Discovery on Estimated Latent Causal Variables

In this section, we evaluate how can we use the latent causal variables we learn as input to causal discovery methods in order to learn the causal structure. In causal representation learning often the main difficulty is to identify the causal variables from the observations, while we can then use standard causal discovery methods to estimate an equivalence class of graphs. In this case, intuitively we expect that the closer the estimated causal variables are to the ground truth, the closer the learned graph to the ground truth causal relations.

Since masking might also mask some causal relations in the original graph, we instead consider a group of data without masked variables and then compare the learned graph to the ground truth. For this task, we additionally assume a few standard assumptions in causal discovery (Spirtes et al., 2000), e.g. the causal Markov assumption and the causal faithfulness assumption, which together imply that conditional independences in the data correspond to d-separations in the true underlying causal graph. For simplicity, we also assume causal sufficiency, which in this case means that there are no additional latent confounders of the reconstructed latent variables, and there is no selection bias.

We consider a set of different algorithms for different underlying causal models, based on the parametric assumptions we use in each setting:

- Linear Gaussian SCM: PC algorithm with partial correlation tests with significance threshold $\alpha = 0.01$ from the *pcalg* (Kalisch et al., 2012) package,
- Linear Exponential SCM: Pairwise LiNGAM (Hyvärinen and Smith, 2013),
- Nonlinear SCM: PC algorithm with Hilbert Schmidt Independence Criterion (HSIC) test with significance threshold $\alpha = 0.01$ from the *kpcalg* (Verbyla et al., 2017) package.

While LiNGAM provides in output a Directed Acyclic Graph (DAG), which can be easily compared with the ground truth DAG, the PC algorithm instead outputs a Completely Partially Oriented DAG (CPDAG). In this case, we consider the ground truth CPDAG (the CPDAG of the Markov Equivalence Class that contains the ground truth DAG).

We evaluate the Structural Hamming distance (SHD) with respect to the ground truth causal graph for LiNGAM and to the ground truth CPDAG for PC. For each setting we consider two distances:

- $SHD_{\mathbf{z}^*}$: the DAG or CPDAG learned with the ground truth causal variables $\mathbf{z}^*$
- $SHD_{\hat{\mathbf{z}}}$ the DAG or CPDAG learned with the estimated causal variables $\hat{\mathbf{z}}$

Since our primary objective is to recover the causal variables, the graph learned with the ground truth $\mathbf{Z}$ is the optimal result we can attain. Therefore we calculate the difference between these two SHDs and we use Δ_{SHD} to denote it. Smaller Δ_{SHD} implies a closer estimated graph to the optimal one.

We provide results for our method for linear mixing functions in Table 13 and for piecewise linear mixing functions in Table 14. We additionally provide results for the oracle version of our method, which uses the known masks for both settings in Table 15. All of these results show a negative correlation between higher MCC (and hence more accurate estimated representations) and lower Δ_{SHD} (and hence more accurate estimated causal graphs w.r.t. to the standard causal discovery setting). This is even clear in the scatterplot for the linear case, shown in Fig. 29. This aligns with the intuition that more well-identified representations tend to lead to graph learning that closely mirrors the one learned from ground truth.

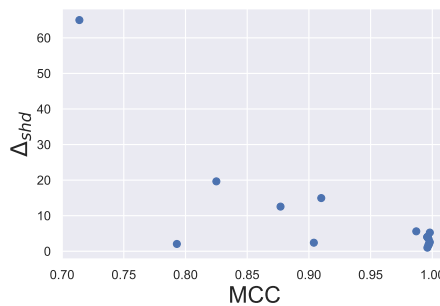Figure 29: MCC-Δ_{shd} of all linear mixing function cases.

Table 13: Results of learning graph for linear mixing functions.

n	k	SCM	ρ	$SHD_{\mathbf{z}^*}$	$SHD_{\hat{\mathbf{z}}}$	Δ_{SHD}
5	1	LIN. GAUSS	50 %	0.65±1.57	2.40±1.90	1.75
10	1	LIN. GAUSS	50 %	1.9±1.89	5.90±3.06	4.00
20	1	LIN. GAUSS	50 %	7.15±3.92	12.75±5.56	5.60
40	1	LIN. GAUSS	50 %	13.70±4.53	79.60±27.08	65.90
10	**1**	LIN. GAUSS	50 %	1.9±1.89	5.90±3.06	4.00
10	**2**	LIN. GAUSS	50 %	16.20±2.38	18.60±4.10	2.40
10	**3**	LIN. GAUSS	50 %	27.20±2.65	29.25±3.29	2.05
10	**1**	LIN. EXP	50 %	2.95±2.26	8.20±4.83	5.25
10	**2**	LIN. EXP	50 %	3.70±2.66	18.65±11.81	14.95
10	**3**	LIN. EXP	50 %	5.20±4.51	24.85±11.39	19.65
10	**1**	NONLINEAR	50 %	5.55±2.50	9±2.94	3.45
10	**2**	NONLINEAR	50 %	14.4±2.04	16.25±3.39	1.85
10	**3**	NONLINEAR	50 %	23.4±2.44	25.4±3.99	1
10	1	LIN. GAUSS	**1VAR**	2.1±1.89	4.65±2.89	2.55
10	1	LIN. GAUSS	**50 %**	1.9±1.89	5.90±3.06	4.00
10	1	LIN. GAUSS	**75 %**	2.15±1.79	14.70±5.25	12.55

Table 14: Results of learning graph for piecewise linear mixing functions with our method.

n	m	k	ρ	δ	$SHD_{\mathbf{z}^*}$	$SHD_{\hat{\mathbf{z}}}$	Δ_{SHD}
5	10	1	50 %	2	0.85±1.79	7±1.89	6.15
10	10	1	50 %	2	2±2.08	19.45±2.35	17.45
20	10	1	50 %	2	6.8±3.49	46.75±5.09	39.95
40	10	1	50 %	2	13.3±4.44	113.4±6.85	100.1
10	10	1	50 %	2	2±2.08	19.45±2.35	17.45
10	10	2	50 %	2	15.85±2.52	25.3±2.32	9.45
10	10	3	50 %	2	27.8±2.33	30.75±2.05	2.95
10	3	1	50 %	2	2±2.08	19.05±2.66	17.05
10	10	1	50 %	2	2±2.08	19.45±2.35	17.45
10	20	1	50 %	2	2±2.08	19.5±2.06	17.5
10	10	1	1VAR	2	2±2.08	19.5±1.79	17.5
10	10	1	50 %	2	2±2.08	19.45±2.35	17.45
10	10	1	75 %	2	2±2.08	18.8±2.28	16.8
10	10	1	50 %	0	2±2.08	17.9±2.71	15.9
10	10	1	50 %	3	2±2.08	18.9±2.17	16.9
10	10	1	50 %	5	2±2.08	18.95±1.76	16.95
10	10	1	50 %	10	2±2.08	19.05±1.93	17.05

Table 15: Results of learning graph for piecewise linear mixing functions with oracle method.

n	m	k	ρ	δ	$SHD_{\mathbf{z}^*}$	$SHD_{\hat{\mathbf{z}}}$	Δ_{SHD}
5	10	1	50 %	2	0.85±1.79	7.1±1.97	6.25
10	10	1	50 %	2	2±2.08	21.35±2.68	19.35
20	10	1	50 %	2	6.8±3.49	56.95±4.17	50.15
40	10	1	50 %	2	13.3±4.44	130.85±9.84	117.55
10	10	1	50 %	2	2±2.08	21.35±2.68	19.35
10	10	2	50 %	2	15.85±2.52	26.1±2.77	10.25
10	10	3	50 %	2	27.8±2.33	29.9±2.75	2.1
10	3	1	50 %	2	2±2.08	19.8±2.80	17.8
10	10	1	50 %	2	2±2.08	21.35±2.68	19.35
10	20	1	50 %	2	2±2.08	21.25±2.84	19.25
10	10	1	1 VAR	2	2±2.08	21.15±2.72	19.15
10	10	1	50 %	2	2±2.08	21.35±2.68	19.35
10	10	1	75 %	2	2±2.08	21.3±3.24	19.3
10	10	1	50 %	0	2±2.08	19.55±2.87	17.55
10	10	1	50 %	3	2±2.08	21.2±2.74	19.2
10	10	1	50 %	5	2±2.08	20.35±2.74	18.35
10	10	1	50 %	10	2±2.08	19.65±2.76	17.65

E.2. Image dataset: Multiple Balls

Data generation process We use PyGame (Shinners, 2011) to render images with size $64 \times 64 \times 3$ as shown in Figure 30. The number of balls b varies from 2 to 8.

For the missing ball setting, we generate the x-coordinates of each ball $\mathbf{C}_j$ for $j \in [b]$ from a truncated normal distribution $\mathcal{N}(0.5, 0.1^2)$ with bounds $(0.1, 0.9)$. For the masked position setting, we generate the x and y-coordinates of each ball $\mathbf{C}_i$ for $j \in [b]$ independently from a truncated 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ with bounds $(0.1, 0.9)^2$, where $\boldsymbol{\mu}_i$ is generated from $\text{Unif}(0.4, 0.6)^2$, and $\boldsymbol{\Sigma}_j = ((0.01, 0.005)(0.005, 0.01))$ for all groups. All the aforementioned parameters, including mean and variance/covariance above, are configured to ensure the majority of the samples are located in interval $[0, 1]$. This configuration aims to maintain the truncated distribution as close to a Gaussian distribution as feasible.

For both settings, we generate the masked causal variables as $(\mathbf{Z}_k)_{k=1}^K = \mathbf{y} \cdot \mathbf{C} + (1 - \mathbf{y}) \cdot \mathbf{M} \in \mathbb{R}^{K \times n}$. For the missing ball setting, the latent size is equal to the number of balls $\mathbf{b}$. Therefore, we predefine b masks via strategy ii) in App. D.4. The mask value is set as $\mathbf{0}$. For the masked position setting, the latent size is $2b$ since both x and y-coordinates of each ball are considered. The masked positions for distinct balls are different; avoiding overlap happens if two balls are masked in the same group. The mask value of each latent is predefined by one of these values $\{0.05, 0.1, 0.9, 0.95\}$.
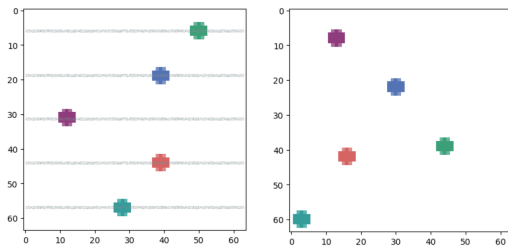

Figure 30: Example images of two settings in multiple ball dataset, based on the rendering code provided by (Ahuja et al., 2022). (left) *Missing ball* setting: balls can move along gray paths only and are not visible when they move out of view. (right) *Masked position* setting: balls can move freely inside the frame and their coordinates are masked when they have a specific value $\mathbf{M}$; e.g., the left-bottom position represents a masked version of the x and y coordinate for the olive ball.

E.3. Image dataset: Causal3DIdent

A qualitative visualization of the result for one object class is presented in Fig. 32 where estimated and ground truth latent are well aligned on the diagonal.
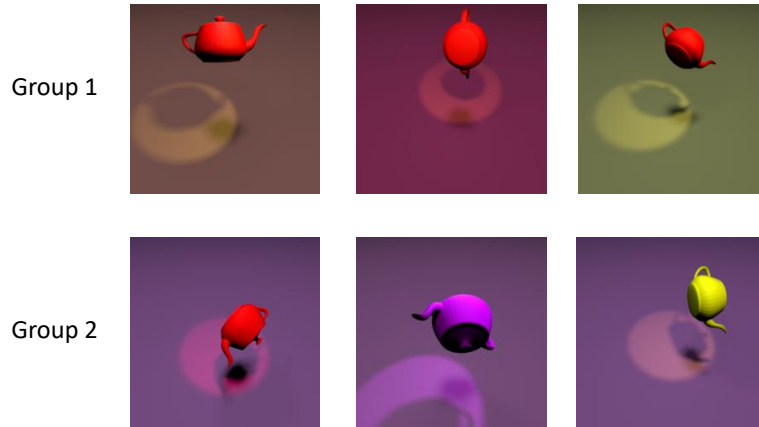
Figure 31: Example images from two groups in PartialCausal3DIdent dataset, resampled from Causal3DIdent datasert provided by (von Kügelgen et al., 2021). In Group 1, the variable *object hue* is masked to a constant. In Group 2, the variable *background hue* is masked to a constant.

Since as we show in Table 3, we did not observe significant performance differences between different object classes, we conduct an ablation study of the masking value δ only on **OBJECT CLASS ID 0**. We show the results in Table 16. Consistent with the observation in numerical experiments, when we increase the value of δ, which measures the distance between mask value and original mean, our method achieves better identification.

Table 16: MCC on **OBJECT CLASS ID=0** regarding various masking values δ with standard deviation $\sigma = 0.1$. Results averaged over three random seeds.

δ	0	2σ	3σ	5σ
MCC (MEAN ± STD)	0.586 ± 0.056	0.632 ± 0.044	0.700 ± 0.039	0.828 ± 0.011

F. Image credits

The images in Fig. 1 are adapted from a VectorPortal image that is covered by CC BY 4.0. We segmented the cars, and removed some of them or moved their position in the image.

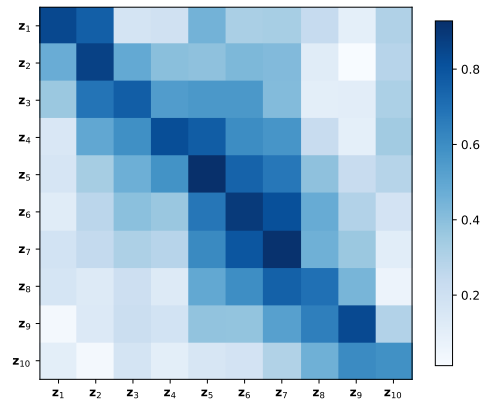Figure 32: Pearson correlation matrix $\mathrm{Corr}_\pi^{n \times n}$ with the permutation π between estimated and ground truth latent variables for PartialCausal3DIdent with masking distance $\delta = 10\sigma$ for all latent variables.