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**[3,7], **Francesco Locatello**[2], and **Sara Magliacane**[1,8]

[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 [5]Mila, Université de Montréal [6]ServiceNow Research [7]University of Cambridge [8]MIT-IBM Watson AI Lab

Abstract

Causal representation learning (CRL) aims at identifying high-level causal variables from low-level data, e.g. images. Most current methods assume that all causal variables are captured in the high-dimensional observations. The few exceptions assume multiple partial observations of the same state, or focus only on the shared causal representations across multiple domains. In this work, we focus on learning causal representations from data under *partial observability*, i.e., when some of the causal variables are masked and therefore not captured in the observations, the observations represent different underlying causal states and the set of masked variables changes across the different samples. We introduce two theoretical results for identifying causal variables in this setting by exploiting a sparsity regularizer. For linear mixing functions, we provide a theorem that allows us to identify the causal variables *up to permutation and element-wise linear transformations* without parametric assumptions on the underlying causal model. For piecewise linear mixing functions, we provide a similar result that allows us to identify Gaussian causal variables *up to permutation and element-wise linear transformations*. We test our theoretical results on simulated data, showing their effectiveness.

1 Introduction

Causal representation learning (CRL) (Schölkopf et al., 2021) aims to identify high-level causal variables from low-level data, e.g. images. Causal reasoning is a promising direction for enhancing machine learning in terms of improving robustness, generalization, and interpretability (Pearl, 2009; Peters et al., 2017; Spirtes et al., 2000). Traditional causality methods assume that the causal variables are given, but in many real-world settings, we only have unstructured, high-dimensional observations of a causal system. A popular approach to identifying 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 components from entangled measurements. Several works generalize this setting to the case in which the latent variables can have causal relations (Ahuja et al., 2023; Brehmer et al., 2022; Buchholz et al., 2023; Lachapelle et al., 2022, 2023; Lippe et al., 2022, 2023; Squires et al., 2023; von Kügelgen et al., 2021, 2023; Wendong et al., 2023; Yao et al., 2022; Zhang et al., 2023), providing different *identifiability* results, given different assumptions on the available data or the data generating process. Most works assume that *all* causal variables are captured in the high-dimensional observations, with the exception of Sturma et al. (2023); Yao et al. (2023). In particular, Yao et al. (2023) focus on the multi-view setting, in which we consider partial observations, or views, of the same latent state, while Sturma et al. (2023) consider recovering a shared causal representation from unpaired observations from multiple domains.

In this work, we focus on learning causal representations from data under *partial observability*, i.e., when some of the causal variables are masked and therefore not captured in the observations, the

Accepted to the NeurIPS 2023 Workshop on Causal Representation Learning.

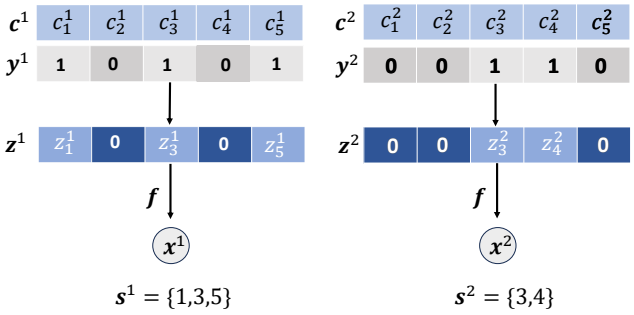


Figure 1: Two example realizations of a *Partially Observable Causal Representation Learning* setting with $n = 5$ latent variables. In our notation, $\mathbf{c}$ are the latent causal variables and $\mathbf{y}$ are the masks that represent which variables are active. The masked causal variables $\mathbf{z}$, which represent the active causal variables (i.e., the ones for which the mask is 1), are the inputs of mixing function $\mathbf{f}$, while the outputs are the observations $\mathbf{x}$. The support indices s represent the indices of the causal variables that not masked in $\mathbf{x}$. We assume all realizations share the same mixing function $\mathbf{f}$, but each observation can have a different masking pattern and represents a potentially different underlying causal state. Our goal is to recover the masked causal variables $\mathbf{z}$ solely from the observations $\mathbf{x}$.

observations represent different underlying causal states and the set of masked variables can change across the different samples. This setting is motivated by real-world applications in which we cannot at all times observe the complete state of the environment. For example, consider a stationary camera taking pictures of a school of fish in a fish tank, where each individual fish can move out of the frame for any image while still affecting the behavior of the visible fish. In this setting, we only observe one image for a given state of the system, and the subsets of causal variables that are observed, i.e., the individual fish, change dynamically across the different images.

We highlight the following contributions:

- We formalize the partial observability problem for CRL, focusing on the case in which there can be different subsets of causal variables that are masked in each sample, and we do not have simultaneous partial observations of the same state of the system.
- We introduce two theoretical results for identifying causal variables under partial observability by exploiting a sparsity regularizer. In particular, Thm. 3.1 proves the identifiability up to permutation and element-wise transformations for the case in which the mixing function is linear and we do not have parametric assumptions on the underlying causal model. Thm. 3.3 proves identifiability up to permutation and element-wise transformations for the piecewise linear case, when the underlying causal variables are Gaussian.
- Finally, we validate the results from Thm. 3.1 and Thm. 3.3 with numerical experiments.

2 Partially Observable Causal Representation Learning

In this section, we formalize the *Partially observable Causal Representation Learning* setting, in which we have a set of high-dimensional observations that are functions of subsets of the true underlying causal variables. In particular, each observation sample $\mathbf{x}^i$ for $i = \{1, \dots, N\}$ is an entangled measurement of a subset of the n latent causal variables $\mathbf{c}^i = \{c_1^i, \dots, c_n^i\}$. Each observation sample can measure different subsets of the latent causal variables, and we assume we do not have access to concurrent measurements of the same realization of causal variables. We also assume that there can be potentially causal relations between the underlying causal variables and that the observations are measurements of i.i.d. samples of the underlying causal variables.

Data generating process. We define our causal variables as a random vector $\mathbf{C} = (\mathbf{C}_1, ..., \mathbf{C}_n)$ that takes values in the causal space $\mathcal{C} = \mathcal{C}_1 \times ... \times \mathcal{C}_n \subseteq \mathbb{R}^n$, which is an open, simply connected latent space. The causal variables follow a distribution $p(\mathbf{C})$, which allows for causal relations between them. We use a binary mask random variable $\mathbf{Y} = \{Y_1, \dots, Y_n\}$ with domain $\mathcal{Y} \subseteq \{0, 1\}^n$

to represent the causal variables that are observed in each of the samples and assume it follows $p(\mathbf{Y})$. The *masked causal variables* $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_n)$ are then the Hadamard product of the causal variables with the random variable representing the binary mask, i.e., $\mathbf{Z} = \mathbf{Y} \odot \mathbf{C}$. This means that in sample $i \in [N]$ and for any causal variable $j \in [n]$ for a realization of the causal variable c_j^i and a corresponding realization of the binary mask y_j^i, if the mask value is 0, the corresponding masked causal variable z_j^i will be 0, and otherwise it will be c_j^i. We define the *support index* random vector $\mathbf{S}$ as the index of non-zero components of $\mathbf{Y}$, i.e., $\mathbf{S} := \{i \in [n] : Y_i \neq 0\}$. The support index vector has a probability mass function $p(\mathbf{s})$ and support $\mathcal{S}$, defined as:

$$\mathbb{P}(\mathbf{s}) := \mathbb{P}\left(\bigwedge_{j \in \mathbf{s}} (\mathbf{Y}_j = 1) \wedge \bigwedge_{j \notin \mathbf{s}} (\mathbf{Y}_j = 0)\right) \quad \text{and} \quad \mathcal{S} := \{\mathbf{s} \subseteq [n] \mid p(\mathbf{s}) > 0\}.$$

We additionally assume that for all $\mathbf{s} \in \mathcal{S}$, the probability measure $\mathbb{P}_{\mathbf{Z}_\mathbf{s}|\mathbf{S}=\mathbf{s}}$ has a density w.r.t. the Lebesgue measure on $\mathbb{R}^{|\mathbf{s}|}$. Finally, we assume that observations $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ are generated by mixing the realizations of the masked causal variables with the same function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^d$, or more precisely $\mathbf{X} = \mathbf{f}(\mathbf{Z})$. We summarize the notation with an example in Fig. 1, where we show the different realizations of five causal variables in two different samples and highlight how the masks and the support indices change across the different samples. Our task is then to recover the masked causal variables from each observation.

To prove our results, we describe the sufficient support index variability, which was originally defined by Lachapelle et al. (2023). A simple case that would satisfy this condition is there is for each causal variable there is at least one observation in which it is the only variable that is masked.

Assumption 2.1. (Sufficient support index variability (Lachapelle et al., 2023)) For all $i \in [n]$, where n is the number of causal variables, 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:

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

Our goal is to identify the masked causal variables from a set of observations under weak assumptions. In general, in CRL we cannot fully identify the (masked) causal variables, but we can only identify the variables up to some transformation. Following Lachapelle et al. (2022), we define two different definitions of identifiability: a weaker form, *identifiability up to linear transformation*, and a stronger form, *identifiability up to permutation and element-wise linear transformation* for arbitrary underlying causal models. In the linear case, we will show that we can leverage sparsity, similar to Lachapelle et al. (2023), to prove identifiability up to permutation and element-wise linear transformation. In the piecewise linear case, we will first prove a weaker form of identifiability, identifiability up to linear transformations, given the binary masks $\mathbf{Y}$, and then provide a theorem that proves the stronger identifiability up to permutation and element-wise linear transformation. We first define the weaker form of identifiability, which recovers the ground truth latent variables up to a linear transformation.

Definition 2.2 (Identifiability up to linear transformation)**.** The ground truth representation vector $\mathbf{Z}$ (a n-dimensional random vector) is said to be identified up to affine transformations by a learned representation vector $\hat{\mathbf{Z}}$ (also a n-dimensional random vector) when there exists an invertible *linear transformation* $\mathbf{h}$ such that, $\hat{\mathbf{Z}} = \mathbf{h}(\mathbf{Z})$ almost surely.

Intuitively, this identifiability definition means there exists a linear function between ground truth variables and learned variables, but this does not imply that each ground truth latent variable is represented in a disentangled way by an estimated latent variable. In order to define a disentangled version of identifiability, we now define identifiability up to element-wise linear transformations:

Definition 2.3 (Identifiability up to permutation and element-wise linear transformations)**.** The ground truth representation vector $\mathbf{Z}$ (a n-dimensional random vector) is said to be identified up to permutation and element-wise linear transformation by a learned representation vector $\hat{\mathbf{Z}}$ (also a n-dimensional random vector) when there exists a permutation matrix $\mathbf{P}$ and an invertible diagonal matrix $\mathbf{D}$ such that $\hat{\mathbf{Z}} = \mathbf{PDZ}$ almost surely.

3 Identifiability via Masked Causal Variables Sparsity Regularization

In this section, we first introduce a theorem (Thm. 3.1) for the identifiability up to permutation and element-wise linear transformations (Def. 2.3) of latent variables when the mixing function $\mathbf{f}$ is linear, inspired by Lachapelle et al. (2022). Then, we present a set of results for the piecewise linear $\mathbf{f}$: we start by presenting an intermediate lemma (Lemma 3.3), which shows that for each realization, we can identify the latent variables up to linear transformation (Def. 2.2) given the binary masks $\mathbf{Y}$. Then we use this lemma to prove identifiability up to permutation and element-wise linear transformations (Def. 2.3) also for this case (Thm. 3.3). All the proofs are in Appendix A.

First, we introduce the theorem for the linear case. This theorem shows that for linear mixing functions, under a perfect reconstruction, a sparsity constraint on the learned representation allows us to identify the ground truth latent variables up to a permutation and element-wise linear transformations.

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 $\mathbf{f}$ is an invertible linear function, and Ass. 2.1 holds. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ 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. If both of the following conditions hold,

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

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

then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$ up to a permutation and element-wise linear transformations (Def. 2.3), i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation composed with element-wise invertible linear transformations.

We provide a proof in App. A.1, based on the proof strategy by Lachapelle et al. (2022). The linearity of $\mathbf{f}$ is a strong assumption and not applicable in many real-world cases. As a first step towards a more general setting, we consider piecewise linear mixing functions $\mathbf{f}$ for causal models with Gaussian variables. In order to prove our results we make the additional assumptions:

Assumption 3.1. (Gaussian causal model) We assume $\mathbf{C}$ follows a *multivariate normal distribution* (MVN), 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.

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

Furthermore, we make an assumption that the piecewise linear function is in a sense "well-behaved", i.e. there exists a ball centered in a point in the support of $\mathbf{f}(\mathbf{Z})$ that only contains one linear piece:

Assumption 3.2. (Existence of a ball with only one linear piece of the piecewise linear $\mathbf{f}$) We assume for the invertible piecewise linear mixing function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^d$, we can always construct an open ball $B(\mathbf{x}_0, \delta) \subseteq \mathbb{R}^d$, where $\mathbf{x}_0 \in \mathbf{f}(\mathcal{Z})$, $\mathcal{Z}$ is the support set of $\mathbf{z}$ and $\delta > 0$, such that it only contains one linear piece of $\mathbf{f}^{-1}$, i.e. on which $\mathbf{f}$ is linear.

We now present a theorem that shows that latent variables that are distributed as potentially degenerate multivariate normals, or *(De)-MVN*, are identifiable up to affine transformation.

Theorem 3.2 (Linear Identifiability for (De)-MVNs with Piecewise Affine $\mathbf{f}$)**.** Assume $\mathbf{f}, \hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ is invertible and piecewise affine. Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the ground truth latent variables, and $\hat{\mathbf{Z}} \sim N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ be the estimated latent variables. We assume $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ follow a (degenerate) multivariate normal distribution, i.e., they are *(De)-MVNs* and Ass. 3.2 holds. If $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\hat{\mathbf{Z}})$ are equally distributed, 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. 2.2).

The proof is in App. A.2 and is inspired by Kivva et al. (2022). The proving strategy has three steps:

i Based on Ass. 3.2 we can construct a ball B that contains a point $\mathbf{x}_0 \in \mathbf{f}(\mathcal{Z})$, where $\mathcal{Z}$ is the support of $\mathbf{z}$, and on which $\mathbf{f}$ is linear;

ii We show a (De)-MVN can be uniquely determined by its value over this ball B;
iii According to the linear closure property of (De)-MVNs, we show that we can construct an affine transformation $\mathbf{h}$ from $\hat{\mathbf{Z}}$ to $\mathbf{Z}$.

We now present the results for piecewise linear case. First, we show that given the information of the binary mask $\mathbf{Y}$, we can identify the latent factors $\mathbf{Z}$ up to a linear transformation (Def. 2.2).

Lemma 3.3 (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, and Ass. 2.1, 3.1, 3.2 hold, and $\mathbf{f}$ is an invertible piecewise linear function. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ be a continuous piecewise linear function and $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible piecewise linear function onto its image. If the following conditions hold,*

$$\mathbb{E}\left\|\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2 = 0\,, \textit{and} \tag{3}$$

$$\mathbf{g}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_{\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{y}}) \quad \text{for some} \quad \boldsymbol{\mu}_{\mathbf{y}} \in \mathbb{R}^n, \boldsymbol{\Sigma}_{\mathbf{y}} \in \mathbb{R}^{n\times n}, \tag{4}$$

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 affine function $\mathbf{h}_{\mathbf{Y}} : \mathbb{R}^n \to \mathbb{R}^n$, such that $\mathbf{h}_{\mathbf{y}}(\mathbf{Z}) \mid (\mathbf{Y} = \mathbf{y}) = \hat{\mathbf{f}}^{-1}(\mathbf{f}(\mathbf{Z})) \mid (\mathbf{Y} = \mathbf{y})$.

The proof is in App. A.3. Based on this lemma, we can now prove that each individual variable is identifiable for piecewise linear $\mathbf{f}$ up to permutation and element-wise transformation.

Theorem 3.3 (Element-wise Identifiability for Piecewise Linear $\mathbf{f}$)**.** Assume the observation $\mathbf{X}$ follows the data-generating process in Sec 2, Ass. 2.1, 3.1, 3.2 hold and $\mathbf{f}$ is an invertible piecewise linear function. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ be a continuous invertible piecewise linear function and let $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible function onto its image. We assume the two conditions of Lemma 3.3 hold. If additionally the following condition holds:

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

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. 2.3).

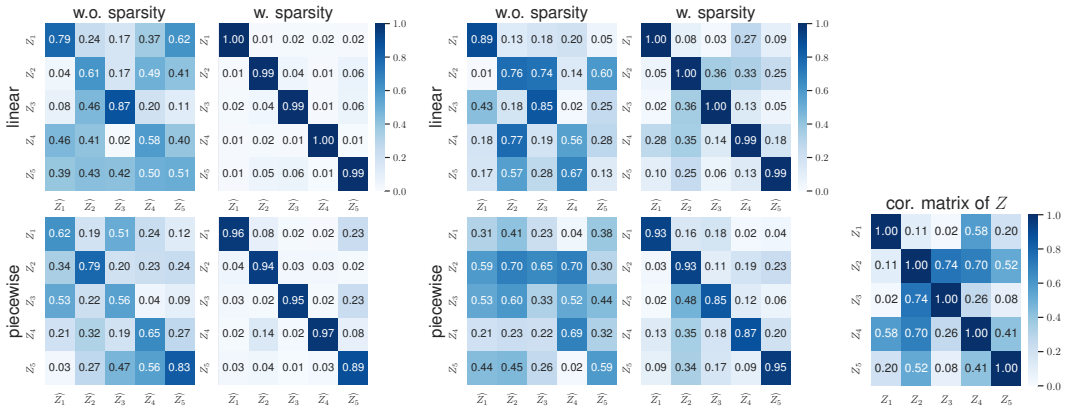
The proof for this theorem is in App. A.4 and it is based on the following reasoning. Intuitively, in Lemma 3.3, we can identify the latent variables up to an affine function $\mathbf{h}_{\mathbf{Y}}$, when the binary mask $\mathbf{Y}$ is given. We start by considering the case in which $\mathbf{Y} = 1$, i.e. there is no masking. In this case, we can know the mask; therefore, we can use Lemma 3.3 to get the reconstruction up to an affine $\mathbf{h}_{\mathbf{Y}=1}$. We also know that we can perfectly reconstruct $\mathbf{Z}$ with $\mathbf{v}(\mathbf{Z}) := \hat{\mathbf{f}}^{-1}(\mathbf{f}(\mathbf{Z}))$ on all $\mathcal{Z}$. This means that $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{Y}=1}$, $\mathbf{h}_{Y=\mathbf{1}}(\mathbf{z}) = \mathbf{v}(\mathbf{z})$. Then, according to Lemma A.7, since the support of $\mathbf{Z}|\mathbf{Y} \neq \mathbf{1}$ is a low dimensional subspace of $\mathbf{Z}|\mathbf{Y} = \mathbf{1}$ (when there is no masking of the causal variables), and we assume that $\mathbf{v}$ is continuous over $\mathbb{R}^n$, then we can derive that $\forall \mathbf{z} \in \mathcal{Z}$, $\mathbf{h}_{Y=\mathbf{1}}(\mathbf{z}) = \mathbf{v}(\mathbf{z})$. Therefore, $\mathbf{v}$ is an invertible affine transformation, and we use a similar strategy to Theorem 3.1 to obtain the element-wise identifiability.

4 Experimental Results

To validate Theorem 3.1 and Theorem 3.3, we perform numerical experiments in a fully-controlled, finite sample setting, using an autoencoder with sparsity regularization.

Data generation. We generate numerical data, following the assumptions in Section 2. We consider $n = 5$ ground truth causal variables $\mathbf{C} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, modeling *causal relations* among them through the covariance matrix $\boldsymbol{\Sigma}$. We generate each mask Y_i independently from a Bernoulli distribution with parameter $p_i = 0.5$. For $\mathbf{f}$, we use a fully connected layer and 2-layer MLP with LeakyReLU ($\alpha = 0.2$) activation functions. To ensure for invertibility, we use the same process as previous work (von Kügelgen et al., 2021), setting a condition threshold ratio to describe how much the invertibility can be violated in each layer. For the encoder $\mathbf{g}$ and the decoder $\hat{\mathbf{f}}$, we use 7-layer MLPs with LeakyReLU activation functions with [10, 50, 50, 50, 50, 10]$\times n$ units per layer.

Masked causal variable sparsity regularization. Both of our results require that we should not only fit the data perfectly, but also reconstruct a model such that the estimated masked causal variables $\hat{\mathbf{Z}} = \mathbf{g}(\mathbf{X})$ are sparse. To do so, we add a regularization term $\lambda||\hat{\mathbf{Z}}||_1$ to train the encoder $\mathbf{g} : \mathcal{X} \to \hat{\mathcal{Z}}$ and decoder $\hat{\mathbf{f}} : \hat{\mathcal{Z}} \to \mathcal{X}$ additionally to their L2 loss.



(a) Independent $\mathbf{Z}$ (b) Dependent $\mathbf{Z}$ (c) Corr. matrix for (b)

Figure 2: Pearson correlation matrix (after permutation π). For (a) and (b), the top row is for linear $\mathbf{f}$, and the bottom row is for piecewise linear $\mathbf{f}$. The left column corresponds to the reconstructed variables without sparsity regularization. The right column is our approach with sparsity regularization. The ground truth correlation matrix for the case of dependent $\mathbf{Z}$ from (b) is shown in (c).

Metrics. Following previous work (Khemakhem et al., 2020; Lachapelle et al., 2022), 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 obtained by computing the Pearson correlation matrix $\mathrm{Corr}^{n\times n}$ between the learned representations $\hat{\mathbf{Z}}$ and ground truth latent masked causal variables $\mathbf{Z}$. Since our results are up to permutation and element-wise linear transformation, we then compute the MCC on the permutation π that maximizes the average of $|\mathrm{Corr}_{i,\pi(i)}|$ for each index of a ground truth variable $i \in [n]$, i.e. MCC=$\frac{1}{n}\max_{\pi\in perm([n])}\sum_{i=1}^{n}|\mathrm{Corr}_{i,\pi(i)}|$.

Results. In Fig. 2 we show the Pearson correlation matrices between the learned representations and ground truth latent factors for the linear and piecewise linear case, both without sparsity and with our sparsity regularizer. Note that these matrices are reordered using the same permutation π that chooses for each masked causal variable Z_i for $i \in [n]$ the reconstructed variable $\hat{Z}_i$ that maximizes the average absolute value of the correlation, as in the MCC calculation. In order to exclude the effect from the correlation among ground truth latent factors and provide more interpretable results, we first show the results for independent $\mathbf{Z}$ in Fig. 2a. We then show the results for dependent $\mathbf{Z}$ in Fig. 2b, where the dependence is induced by the causal structure between the masked causal variables. As we can see for independent $\mathbf{Z}$ in Fig. 2a, in both the linear (top row) and piecewise linear (bottom row) case, the individual latent variables are identified by our sparsity regularizer, showing a high *disentanglement*, i.e., correlation to only one of the learned representations, while the baseline representation is entangled. When we consider potential causal relation within $\mathbf{Z}$ (Fig. 2b), a latent variable may become predictable by another latent variable, even if the representation we learned was disentangled. Therefore, we can observe that some estimate variable (e.g.$\hat{X}_2$) has a relatively higher correlation with other latent variables. The pattern of correlations in our reconstructed variables is consistent with the ground truth correlation matrix in Fig. 2c.

In Table 1, we report the average of the MCC over 5 random seeds with its standard deviation across four settings, which are combinations of: (i) the type of mixing function $\mathbf{f}$ and (ii) dependence within latent variables $\mathbf{Z}$. An average MCC close to one indicates that there is a one-to-one relation between learned representations and ground truth latent variables, i.e. the latent variables are identified by learned representations up to element-wise linear transformations. For all setups, we can identify each individual variable almost perfectly with sparsity regularization, while the baseline without regularization struggles.

Generative Process		MCC	
Mixing function f	**Dependence**	**w. sparsity**	**w.o. sparsity**
Linear	No	**0.994**$\pm$0.000	0.736$\pm$0.001
Linear	Yes	**0.980**$\pm$0.000	0.692$\pm$0.003
Piecewise Linear	No	**0.935**$\pm$0.001	0.687$\pm$0.003
Piecewise Linear	Yes	**0.913**$\pm$0.002	0.601$\pm$0.003

Table 1: Average MCC for the four settings.

5 Conclusions and Future work

In this work, we focused on learning causal representations from data under *partial observability*, i.e., when some of the causal variables are not observed in the measurements. We introduced two theoretical results for identifying causal variables under partial observability by exploiting a sparsity regularizer, focusing in particular on the linear and piecewise linear mixing functions, and providing identifiability proofs up to permutation and element-wise transformation. In future work, we plan to extend our experimental evaluation and apply our methods to more realistic datasets.

Acknowledgements

This work was initiated at the Second Bellairs Workshop on Causality held at the Bellairs Research Institute, January 6–13, 2022; we thank all workshop participants for providing a stimulating research environment. The research of DX and SM was supported by the Air Force Office of Scientific Research under award number FA8655-22-1-7155. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the United States Air Force. We also thank SURF for the support in using the Dutch National Supercomputer Snellius. DY was supported by an Amazon fellowship and the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Work done outside of Amazon. SL was supported by an IVADO excellence PhD scholarship and by Samsung Electronics Co., Ldt. JvK acknowledges support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B).

References

Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, pages 372–407. PMLR, 2023. [Cited on page 1.]

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. [Cited on page 1.]

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. [Cited on page 1.]

S.H. Friedberg, A.J. Insel, and L.E. Spence. *Linear Algebra*. Pearson Education, 2014. ISBN 9780321998897. URL https://books.google.ca/books?id=KyB0DAAAQBAJ. [Cited on page 11.]

Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016. [Cited on page 1.]

Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017. [Cited on page 1.]

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. [Cited on page 1.]

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. [Cited on pages 1 and 6.]

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. [Cited on pages 4 and 12.]

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. [Cited on pages 1, 3, 4, and 6.]

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. [Cited on pages 1 and 3.]

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. [Cited on page 1.]

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. [Cited on page 1.]

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X. [Cited on page 1.]

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319. [Cited on page 1.]

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. [Cited on page 1.]

Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge MA, 2nd edition, 2000. [Cited on page 1.]

Chandler Squires, Anna Seigal, Salil Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In *40th International Conference on Machine Learning*, 2023. [Cited on page 1.]

Nils Sturma, Chandler Squires, Mathias Drton, and Caroline Uhler. Unpaired multi-domain causal representation learning. *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Cited on page 1.]

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. [Cited on pages 1 and 5.]

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. [Cited on page 1.]

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. [Cited on page 1.]

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. [Cited on page 1.]

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. [Cited on page 1.]

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. [Cited on page 1.]

A Proofs

A.1 Proof of Theorem 3.1

We will first introduce the definition of *dependent inputs*, which intuitively will be the set of variables on which a reconstruction of a given variable depends. Our end goal will be to prove that a method provides a reconstruction in which there exists a permutation such that only the variable $i \in [n]$ itself is in the set of dependent inputs N_i.

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

$$N_i := \{j \in [n] \mid \exists \mathbf{z}^0_{-j} \in \mathcal{Z}_{-j} \text{ s.t. } \mathbf{v}_i(z_j, \mathbf{z}^0_{-j}) \text{ is not a constant function of } z_j\}\,, \tag{6}$$

where $\mathbf{z}^0_{-j}$ is an $n-1$ dimensional vector with domain:

$$\mathcal{Z}_{-j} := \{\mathbf{z}_{-j} : (z_j, \mathbf{z}_{-j}) \in \mathcal{Z}\}\,, \tag{7}$$

We now prove a lemma that will be useful in the proof of the theorem, that shows that for any diffeomorphism and any variable, there exists always a permutation that ensured 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 will depend on the variable itself.

Lemma A.2 (Existence of permutation π s.t. $i \in N_{\pi(i)}$)**.** *Let $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^n$ be a diffeomorphism with variables $\mathbf{z} = (z_1, ..., z_n)$ and domain $\mathcal{Z}$. 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. A.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.

$$\det(D\mathbf{v}(\mathbf{z}^0)) := \sum_{\pi \in \mathfrak{S}_n} \operatorname{sign}(\pi) \prod_{i=1}^{n} D\mathbf{v}(\mathbf{z}^0)_{\pi(i),i} \neq 0\,, \tag{8}$$

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:

$$\exists \pi \in \mathfrak{S}_n, \forall i \in [n], D\mathbf{v}(\mathbf{z}^0)_{\pi(i),i} \neq 0\,. \tag{9}$$

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. A.1, $i \in N_{\pi(i)}$. □

We now prove identifiability up to permutation and element-wise linear transformations for 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 $\mathbf{f}$ is an invertible linear function, and Ass. 2.1 holds. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ 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. If both of the following conditions hold,

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

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

then $\mathbf{Z}$ is identified by $\hat{\mathbf{f}}^{-1}(\mathbf{X})$ up to a permutation and element-wise linear transformations (Def. 2.3), i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ is a permutation composed with element-wise invertible linear transformations.

Proof. Since $\mathbf{X} = \mathbf{f}(\mathbf{Z})$, we can rewrite Equation (1) (perfect reconstruction) as

$$\mathbb{E}||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 = 0\,. \tag{10}$$

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.,

$$\mathbf{f}(\mathbf{z}) = \hat{\mathbf{f}} \circ \mathbf{g} \circ \mathbf{f}(\mathbf{z}) \,, \forall \mathbf{z} \in \mathcal{Z} \,. \tag{11}$$

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 inverse of $\hat{\mathbf{f}}$ on both sides to obtain

$$\hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}) = \underbrace{\mathbf{g} \circ \mathbf{f}}_{\mathbf{v}:=}(\mathbf{z}) \,, \forall \mathbf{z} \in \mathcal{Z} \,. \tag{12}$$

Since both $\mathbf{f}$ and $\mathbf{g}$ are invertible linear functions, $\mathbf{v}$ is also an invertible linear function.

We now show that $\mathbf{v}$ is a permutation composed with an element-wise linear transformation. To do this, we leverage the sparsity constraint (2):

$$\mathbb{E} \left\| \mathbf{g}(\mathbf{X}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 \tag{13}$$

$$\mathbb{E} \left\| \mathbf{g}(\mathbf{f}(\mathbf{Z})) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 \tag{14}$$

$$\mathbb{E} \left\| \mathbf{v}(\mathbf{Z}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 \tag{15}$$

We reuse the definition of the support indices $\mathbf{S} := \{i \in [n] : Z_i \neq 0\}$ and analyze each side of inequality (15), starting with its right-hand side.

$$\mathbb{E}||\mathbf{Z}||_0 = \mathbb{E} \sum_{i=1}^{n} \mathbb{1}(\mathbf{Z}_i \neq 0) \tag{16}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \mathbb{E} \left[\sum_{i=1}^{n} \mathbb{1}(\mathbf{Z}_i \neq 0) \mid \mathbf{S} = \mathbf{s} \right] \tag{17}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(\mathbf{Z}_i \neq 0) \mid \mathbf{S} = \mathbf{s}] \tag{18}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{1}(i \in \mathbf{s}) \tag{19}$$

Now analyzing the left hand side of (15), starting with similar steps as previously we get

$$\mathbb{E}||\mathbf{v}(\mathbf{Z})||_0 = \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(\mathbf{v}_i(\mathbf{Z}) \neq 0) \mid \mathbf{S} = \mathbf{s}] \tag{20}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) \neq 0] \tag{21}$$

$$= \sum_{\mathbf{s} \in \mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \left(1 - \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0]\right) \,. \tag{22}$$

For $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ to be a permutation composed with an element-wise invertible linear transformation, it is enough to show there exists a permutation $\pi : [n] \to [n]$ such that, for every i, $N_i = \{\pi(i)\}$. To achieve this, we are going to first show that

$$\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \mathbb{1}(N_i \cap \mathbf{s} = \varnothing) \,. \tag{23}$$

Since $\mathbf{v}_i$ is linear, we have that $\mathbf{v}_i(\mathbf{Z}) = \mathbf{w}^i \cdot \mathbf{Z}$ for some $\mathbf{w}^i \in \mathbb{R}^n$. Furthermore, $N_i = \{j \in [N] \mid \mathbf{w}^i_j \neq 0\}$. Thus,

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

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

$$\mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{v}_i(\mathbf{Z}) = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}^i_{N_i} \cdot \mathbf{Z}_{N_i} = 0] = \mathbb{P}_{\mathbf{Z}|\mathbf{S}=\mathbf{s}}[\mathbf{w}^i_{N_i} \cdot \mathbf{0} = 0] = 1\,.$$

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

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

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

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

Putting (15), (19), (22) and (23) together, we obtain

$$\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [1 - \mathbb{1}(N_i \cap \mathbf{s} = \varnothing)] \leq \sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} \mathbb{1}(i \in \mathbf{s}) \tag{24}$$

By Lemma A.2, 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:

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

$$\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [1 - \mathbb{1}(N_{\pi(i)} \cap \mathbf{s} = \varnothing) - \mathbb{1}(i \in \mathbf{s})] \leq 0 \tag{25}$$

$$\sum_{\mathbf{s}\in\mathcal{S}} p(\mathbf{s}) \sum_{i=1}^{n} [\mathbb{1}(N_{\pi(i)} \cap \mathbf{s} \neq \varnothing) - \mathbb{1}(i \in \mathbf{s})] \leq 0\,. \tag{26}$$

Note how, for all 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 $i \in N_{\pi(i)}$. If $i \notin \mathbf{s}$, the function can have either value 0 or 1, but in any case not negative. Hence the inequality in (26) is actually an equality and hence for all $\mathbf{s} \in \mathcal{S}$ and all $i \in [n]$,

$$\mathbb{1}(N_{\pi(i)} \cap \mathbf{s} \neq \varnothing) = \mathbb{1}(i \in \mathbf{s}) \tag{27}$$

$$\mathbb{1}(N_{\pi(i)} \cap \mathbf{s} = \varnothing) = \mathbb{1}(i \notin \mathbf{s})\,. \tag{28}$$

Importantly, this means

$$\forall i \in [n], \forall \mathbf{s} \in \mathcal{S},\ i \notin \mathbf{s} \implies N_{\pi(i)} \cap \mathbf{s} = \varnothing \implies N_{\pi(i)} \subseteq \mathbf{s}^c\,, \tag{29}$$

which can be rewritten as

$$\forall i \in [n], N_{\pi(i)} \subseteq \bigcap_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}} \mathbf{s}^c\,. \tag{30}$$

We now rewrite Assumption 2.1 below and take the complement on both sides:

$$\forall i \in [n], \bigcup_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}} \mathbf{s} = [n] \setminus \{i\} \tag{31}$$

$$\bigcap_{\mathbf{s}\in\mathcal{S}|i\notin\mathbf{s}} \mathbf{s}^c = \{i\} \tag{32}$$

Combining (30) with (32) implies that $N_{\pi(i)} = \{i\}$ for all i, which concludes the proof. □

A.2 Proof of Theorem 3.2

We first prove a lemma that will be useful for our final result.

Lemma A.3. *(Degenerate) Multivariate Normals 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* $\mathbf{AZ}$*, where* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *is also a potentially degenerate multivariate normal variable.*

Proof. Let $\hat{\mathbf{Z}} = \mathbf{AZ}$, then $P_{\hat{\mathbf{Z}}} = \mathbf{A}N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = N(\mathbf{A}\mu, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)$ 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. □

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 from Theorem C.3 by Kivva et al. (2022).

Theorem A.4 (Identifiability of non-degenerate MVNs (Kivva et al., 2022))**.** *Consider a pair of non-degenerate MVNs in* $\mathbb{R}^n$*. If*

$$P = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \textit{and} \quad P' = N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{33}$$

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, that 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 A.5 (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*

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \textit{and} \quad \mathbf{X}' \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{34}$$

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 $\boldsymbol{\Sigma}$:

$$\boldsymbol{\Sigma} = QDQ^T, \tag{35}$$

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 $k < n$, 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

$$\mathbf{Y} \sim N(Q^T\boldsymbol{\mu}, Q^TQDQ^TQ) = N(Q^T\mu, D) \tag{36}$$

$$\mathbf{Y}' \sim N(Q^T\boldsymbol{\mu}', Q^T\Sigma'Q) \tag{37}$$

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 A.4 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 $\mathbf{X} \equiv \mathbf{X}'$. □

Lemma A.6 (Identifiability of (De)-MVNs)**.** *Consider a pair of random vectors* $\mathbf{X}$, $\mathbf{X}'$ *in* $\mathbb{R}^n$ *distributed as*

$$\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \textit{and} \quad \mathbf{X}' \sim N(\boldsymbol{\mu}', \boldsymbol{\Sigma}'), \tag{38}$$

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

$$\boldsymbol{\Sigma} = QDQ^T, \tag{39}$$

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 σ_1^2, σ_2^2, ..., σ_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

$$\mathbf{Y} \sim N(Q^T\boldsymbol{\mu}, Q^TQDQ^TQ) = N(Q^T\boldsymbol{\mu}, D) \tag{40}$$

$$\mathbf{Y}' \sim N(Q^T\boldsymbol{\mu}', Q^T\Sigma'Q) \tag{41}$$

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 A.5 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 (40), 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:

$$\mathbf{X} = Q\mathbf{Y} = Q({\mathbf{Y}_{1:k}}^T, {\mathbf{Y}_{k+1:n}}^T) \tag{42}$$

$$\mathbf{X}' = Q\mathbf{Y}' = Q({\mathbf{Y}'_{1:k}}^T, {\mathbf{Y}'_{k+1:n}}^T), \tag{43}$$

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 $\mathbf{X} \equiv \mathbf{X}'$. □

Theorem 3.2 (Linear Identifiability for (De)-MVNs with Piecewise Affine $\mathbf{f}$)**.** Assume $\mathbf{f}, \hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ is invertible and piecewise affine. Let $\mathbf{Z} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the ground truth latent variables, and $\hat{\mathbf{Z}} \sim N(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ be the estimated latent variables. We assume $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ follow a (degenerate) multivariate normal distribution, i.e., they are *(De)-MVNs* and Ass. 3.2 holds. If $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\hat{\mathbf{Z}})$ are equally distributed, 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. 2.2).

Proof. We construct a ball $B(\mathbf{x}_0, \delta) \subseteq \mathbb{R}^d$,where $\mathbf{x}_0 \in \mathbf{f}(\mathcal{Z})$, $\mathcal{Z}$ is the support set of $\mathbf{z}$ and $\delta > 0$. By assumption $\mathbf{f}$ and $\hat{\mathbf{f}}$ are invertible and equally distributed, so they are also invertible on $B(\mathbf{x}_0, \delta) \cap \mathbf{f}(\mathbb{R}^n)$. Additionally, since we assume they are invertible piecewise affine, the inverse functions are also piecewise affine. Moreover, by assumption 3.2, we also have that the ball $B(\mathbf{x}_0, \delta)$ we construct can only contain one linear piece of $\mathbf{f}^{-1}$ and $\hat{\mathbf{f}}^{-1}$, i.e. both $\mathbf{f}^{-1}$ and $\hat{\mathbf{f}}^{-1}$ on $B(\mathbf{x}_0, \delta)$ are defined by affine functions.

Let $L \subseteq \mathbb{R}^d$ be an affine subspace, such that $B(\mathbf{x}_0, \delta) \cap \mathbf{f}(\mathbb{R}^n) = B(\mathbf{x}_0, \delta) \cap L$. Let $\mathbf{h}_\mathbf{f}, \mathbf{h}_{\hat{\mathbf{f}}} : \mathbb{R}^n \to L$ be a pair of invertible affine functions such that $\mathbf{h}_\mathbf{f}^{-1}$ coincides with $\mathbf{f}^{-1}$ on $B(\mathbf{x}_0, \delta) \cap L$ and $\mathbf{h}_{\hat{\mathbf{f}}}^{-1}$ coincides with $\hat{\mathbf{f}}^{-1}$ on $B(\mathbf{x}_0, \delta) \cap L$. This means that distributions $\mathbf{h}_\mathbf{f}(\mathbf{Z})$ and $\mathbf{h}_{\hat{\mathbf{f}}}(\hat{\mathbf{Z}})$ coincide on $B(\mathbf{x}_0, \delta) \cap L$, since $\mathbf{f}(\mathbf{Z}) \equiv \hat{\mathbf{f}}(\hat{\mathbf{Z}})$. Moreover, since $\mathbf{h}_\mathbf{f}$ and $\mathbf{h}_{\hat{\mathbf{f}}}$ are affine transformations and $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ are (De)-MVNs by assumption, by Lemma A.3, $\mathbf{h}_\mathbf{f}(\mathbf{Z})$ and $\mathbf{h}_{\hat{\mathbf{f}}}(\hat{\mathbf{Z}})$ are also (De)-MVNs with mean $\mathbf{h}_\mathbf{f}(\boldsymbol{\mu})$ and $\mathbf{h}_\mathbf{f}(\hat{\boldsymbol{\mu}})$. Additionally, since $\mathbf{x}_0 \in \mathbf{f}(\mathcal{Z})$ and $\mathbf{f}$ is affine over $B(\mathbf{x}_0, \delta)$, $\mathbf{x}_0 \in L$. Then, $\mathbf{x}_0 \in B(\mathbf{x}_0, \delta) \cap L$. By Lemma A.6, $\mathbf{h}_\mathbf{f}(\mathbf{Z}) \equiv \mathbf{h}_{\hat{\mathbf{f}}}(\hat{\mathbf{Z}})$. We define an invertible affine transformation $\mathbf{h} := \mathbf{h}_{\hat{\mathbf{f}}}^{-1} \circ \mathbf{h}_\mathbf{f}$. Then we can prove the theorem's claim by showing that $\mathbf{h}(\mathbf{Z}) = \mathbf{h}_{\hat{\mathbf{f}}}^{-1}(\mathbf{h}_\mathbf{f}(\mathbf{Z})) \equiv \hat{\mathbf{Z}}$. □

A.3 Proof of Lemma 3.3

Lemma 3.3 (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, and Ass. 2.1, 3.1, 3.2 hold, and $\mathbf{f}$ is an invertible piecewise linear function. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ be a continuous piecewise linear function and $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible piecewise linear function onto its image. If the following conditions hold,*

$$\mathbb{E}\left\|\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))\right\|_2^2 = 0\,, \text{and} \tag{3}$$

$$\mathbf{g}(\mathbf{X}) \mid (\mathbf{Y} = \mathbf{y}) \sim N(\boldsymbol{\mu}_\mathbf{y}, \boldsymbol{\Sigma}_\mathbf{y}) \quad \text{for some} \quad \boldsymbol{\mu}_\mathbf{y} \in \mathbb{R}^n, \boldsymbol{\Sigma}_\mathbf{y} \in \mathbb{R}^{n \times n}, \tag{4}$$

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 affine function $\mathbf{h}_\mathbf{Y} : \mathbb{R}^n \to \mathbb{R}^n$, such that $\mathbf{h}_\mathbf{y}(\mathbf{Z}) \mid (\mathbf{Y} = \mathbf{y}) = \hat{\mathbf{f}}^{-1}(\mathbf{f}(\mathbf{Z})) \mid (\mathbf{Y} = \mathbf{y})$.

Proof. From the perfect reconstruction constraint (3), we can derive

$$\mathbb{E}||\mathbf{X} - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))||_2^2 = 0 \tag{44}$$

$$\mathbb{E}||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z}))))||_2^2 = 0 \tag{45}$$

$$\mathbb{E}\left\{\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 \mid \mathbf{Y}\right]\right\} = 0 \tag{46}$$

$$\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 \mid \mathbf{Y}\right] = 0 \quad \mathbb{P}_\mathbf{Y}\text{−a.e.} \tag{47}$$

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 (47), we have for any value $\mathbf{y} \in \mathcal{Y}$ we have

$$\mathbb{E}\left[||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{v}(\mathbf{Z}))||_2^2 \mid \mathbf{Y} = \mathbf{y}\right] = 0, \tag{48}$$

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. Since $\mathbf{Z}$ and $\mathbf{v}(\mathbf{Z})$ are potentially degenerate MVNs, by Theorem 3.2, there exists an invertible affine transformation $\mathbf{h}_\mathbf{y} : \mathbb{R}^n \to \mathbb{R}^n$ such that

$$\mathbf{h}_\mathbf{y}(\mathbf{Z}) \equiv \mathbf{v}(\mathbf{Z}) \tag{49}$$

This proves that in the case of known mask $\mathbf{Y} = \mathbf{y}$ we can identify the masked causal variables mixed through a piecewise linear function up to a linear transformation. □

A.4 Proof of Theorem 3.3

Before we proof Theorem 3.3, we first prove an intermediate Lemma.

Lemma A.7 (Consistent affine transformation across all realization)**.** *Suppose we have a n-dimensional random vector* $\mathbf{Z}$ *with an open, connected support space* $\mathcal{Z} \subseteq \mathbb{R}^n$*. Let* $\mathcal{Z}^0$ *be a lower-dimensional subspace of* $\mathcal{Z}$*. Assume* $\mathbf{v} : \mathbb{R}^n \to \mathbb{R}^n$ *is a homeomorphism function. If we have another homeomorphism function* $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ *such that*

$$\mathbf{f}(\mathbf{z}) = \mathbf{v}(\mathbf{z}) \qquad \forall \mathbf{z} \in \mathcal{Z} \setminus \mathcal{Z}^0$$

then, since $\mathbf{f}$ *and* $\mathbf{v}$ *are both continuous, we can intuitively conclude that:*

$$\mathbf{f}(\mathbf{z}) = \mathbf{v}(\mathbf{z}) \qquad \forall \mathbf{z} \in \mathcal{Z}.$$

Proof. Let $\mathbf{d}(\mathbf{z}) := \mathbf{f}(\mathbf{z}) - \mathbf{v}(\mathbf{z})$, $\forall \mathbf{z} \in \mathcal{Z}$. Since both $\mathbf{f}$ and $\mathbf{v}$ are continuous functions, $\mathbf{d}$ is continuous function as well.

Let set $A := \mathbf{d}^{-1}(\mathbb{R}^n \setminus \mathbf{0})$, where $\mathbf{0}$ is the vector of all zeros. Since $\mathbb{R}^n \setminus \mathbf{0}$ is an open set and $\mathbf{d}$ is continuous, we can derive that A is an open set as well. We furthermore know that $A \subseteq \mathcal{Z}^0$, because it is the preimage of the non-zero values of $\mathbf{d}$, i.e. the collection of $\mathbf{z} \in \mathcal{Z}$ for which $\mathbf{f}(\mathbf{z}) \neq \mathbf{v}(\mathbf{z})$.

Since $\mathcal{Z}^0$ is a lower-dimensional subspace of $\mathcal{Z}$, the measure of $\mathcal{Z}^0$ is zero. Therefore, the Lebesgue measure of set A which is contained in $\mathcal{Z}^0$ is zero as well.

Since A is an open set with measure zero, and in this measure all non-empty open sets have a non-zero measure, we can derive $A = \varnothing$. Thus, $\mathbf{f}(\mathbf{z}) = \mathbf{v}(\mathbf{z}) \quad \forall \mathbf{z} \in \mathcal{Z}$. □

With this results, we can now prove the following Theorem 3.3.

Theorem 3.3 (Element-wise Identifiability for Piecewise Linear $\mathbf{f}$)**.** Assume the observation $\mathbf{X}$ follows the data-generating process in Sec 2, Ass. 2.1, 3.1, 3.2 hold and $\mathbf{f}$ is an invertible piecewise linear function. Let $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^n$ be a continuous invertible piecewise linear function and let $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible function onto its image. We assume the two conditions of Lemma 3.3 hold. If additionally the following condition holds:

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

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. 2.3).

Proof. From Lemma 3.3, we know that for all $\mathbf{y} \in \mathcal{Y}$, given mask $\mathbf{Y} = \mathbf{y}$, $\mathbf{v}(\mathbf{Z}) \equiv \mathbf{h}_{\mathbf{y}}(\mathbf{Z})$, where $\mathbf{h}_{\mathbf{y}} : \mathbb{R}^n \to \mathbb{R}^n$ is an invertible affine transformation.

We start by considering the case in which $\mathbf{Y} = 1$, i.e. there is no masking. In this case, we can know the mask; therefore, we can use Lemma 3.3 to get the reconstruction up to an affine $\mathbf{h}_{\mathbf{Y}=1}$. We also know that we can perfectly reconstruct $\mathbf{Z}$ with $\mathbf{v}(\mathbf{Z}) := \hat{\mathbf{f}}^{-1}(\mathbf{f}(\mathbf{Z}))$ on all $\mathcal{Z}$. This means that $\forall \mathbf{z} \in \mathcal{Z}_{\mathbf{Y}=1}$, $\mathbf{h}_{Y=\mathbf{1}}(\mathbf{z}) = \mathbf{v}(\mathbf{z})$. Then, according to Lemma A.7, since the support of $\mathbf{Z}|\mathbf{Y} \neq \mathbf{1}$ is a low dimensional subspace of $\mathbf{Z}|\mathbf{Y} = \mathbf{1}$ (when there is no masking of the causal variables), and we assume that $\mathbf{v}$ is continuous over $\mathbb{R}^n$, then we can derive that $\forall \mathbf{z} \in \mathcal{Z}$, $\mathbf{h}_{Y=\mathbf{1}}(\mathbf{z}) = \mathbf{v}(\mathbf{z})$. Therefore, $\mathbf{v}$ is an invertible affine transformation.

We can now apply the inverse of $\hat{\mathbf{f}}$ on both sides of equation 3 to obtain

$$\hat{\mathbf{f}}^{-1} \circ \mathbf{f}(\mathbf{z}) = \underbrace{\mathbf{g} \circ \mathbf{f}}_{\mathbf{v}:=}(\mathbf{z}) , \forall \mathbf{z} \in \mathcal{Z} . \tag{50}$$

where $\mathbf{v}$ is an invertible affine function.

To show that $\mathbf{v}$ is a permutation composed with an element-wise linear transformation, we leverage the sparsity constraint (5) and reuse the same strategy from (15) to (32) to conclude the proof.

□