Identifiability of Potentially Degenerate Gaussian Mixture Models With Piecewise Affine Mixing

Danru Xu
University of Amsterdam

Sébastien Lachapelle
Samsung AI Lab, Montreal

Sara Magliacane
Saarland University &
University of Amsterdam

Abstract

Causal representation learning (CRL) aims to identify the underlying latent variables from high-dimensional observations, even when variables are dependent with each other. We study this problem for latent variables that follow a potentially degenerate Gaussian mixture distribution and that are only observed through the transformation via a piecewise affine mixing function. We provide a series of progressively stronger identifiability results for this challenging setting in which the probability density functions are ill-defined because of the potential degeneracy. For identifiability up to permutation and scaling, we leverage a sparsity regularization on the learned representation. Based on our theoretical results, we propose a two-stage method to estimate the latent variables by enforcing sparsity and Gaussianity in the learned representations. Experiments on synthetic and image data highlight our method's effectiveness in recovering the ground-truth latent variables.

1 INTRODUCTION

Recovering latent variables from high-dimensional observations, such as images, videos, or text, is essential for building reliable and interpretable models (Bengio et al., 2013). While traditional approaches, e.g., (nonlinear) ICA (Comon, 1994; Hyvärinen, 2013; Hyvarinen and Morioka, 2016, 2017), identify independent or conditionally independent variables, in many real-world settings latent variables often exhibit complex dependencies, e.g., because they are causally related.

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

Causal representation learning (CRL) (Schölkopf et al., 2021) aims to identify latent variables with arbitrary dependencies from observations. Many CRL works consider additional data or information to identify causal variables, e.g., interventions (Brehmer et al., 2022; Lippe et al., 2022; Varici et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Ahuja et al., 2023a), actions or temporal structure (Lippe et al., 2023; Lachapelle et al., 2022, 2024), multi-view structures (von Kügelgen et al., 2021; Ahuja et al., 2023b; Yao et al., 2024; Morioka and Hyvärinen, 2024), grouping of observations by sparsity patterns (Xu et al., 2024), hierarchical structures (Kong et al., 2023, 2024) or conditions on the causal graphs (Zhang et al., 2024; Dong et al., 2024; Song et al., 2024).

In our case, we do not assume any additional data or information besides the observations, but we instead consider a set of parametric assumptions on the latent variables $\mathbf{Z}$ and mixing function $\mathbf{f}$ through which we get the observation $\mathbf{X} = \mathbf{f}(\mathbf{Z})$. In particular, we focus on latent variables that follow a mixture of Gaussian distributions with a piecewise affine mixing function. Similarly to us, Kivva et al. (2022) studies the identifiability of non-degenerate Gaussian Mixture Model (GMM) latent variables with piecewise affine mixing, but to achieve element-wise identifiability they require that all pairs of latent variables must be conditionally independent given an (unobserved) auxiliary variable.

We achieve the same identifiability result without requiring this auxiliary variable, while extending the setting to latent variables that follow a potentially degenerate Gaussian Mixture Model (pdGMM), i.e., a GMM in which the components can be potentially degenerate Gaussians, but we instead require a sparsity assumption. This is inspired by the success of sparse (low-rank, i.e., degenerate) representations. In particular, in many real-world settings, mixtures of low-dimensional subspaces provide a more accurate model than full-rank representations (Buchanan et al., 2025).

We provide a series of progressively stronger identifi-

ability results under reasonable assumptions in this challenging setting, in which the probability density functions are ill-defined and hence previous results cannot be applied. As a final step, we leverage a sparsity regularization on the learned representation, inspired by multi-task learning (Lachapelle et al., 2023a) and partially observable CRL (Xu et al., 2024) to prove identifiability up to permutation, scaling and translation, i.e., that we learn a disentangled representation of the latent variables. We highlight our contributions:

- We provide an identifiability result for pdGMMs, showing that the full distribution can be uniquely identified from an open subset that intersects the support of every component (Thm. 3.2). This is a critical step in our other results, but is also of independent interest due to its proof strategy;
- We introduce a series of identifiability results for recovering latent variables that follow a pdGMMs distribution from observations (Thm. 3.5, 3.7, 3.9), showing that under progressively stronger assumptions, latent variables can be recovered up to (i) affine transformation within components (ATwC), (ii) global affine transformation (AT), and finally, (iii) permutation and scaling (PS), without interventions or supervision;
- Finally, we propose a two-stage algorithm that implements these theoretical results and evaluate it on numerical and image datasets.

2 PROBLEM SETTING AND BACKGROUND

We consider a setting with latent random variables $\mathbf{Z} = (Z_1, ..., Z_n) \in \mathcal{Z} \subseteq \mathbb{R}^n$ that have arbitrary dependences with each other. Given only a set of *observations* $\mathbf{X} \in \mathcal{X} \subseteq \mathbb{R}^d$ generated by an unknown mixing function $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$, our goal is to recover the latent variables $\mathbf{Z}$. In general, we cannot fully recover these latent variables, but we can identify them up to an equivalence class. We first summarize two types of identifiability presented in previous work (Comon, 1994; Khemakhem et al., 2020; Kivva et al., 2022; Lachapelle et al., 2023a).

Definition 2.1. Given an observation $\mathbf{X} = \mathbf{f}(\mathbf{Z})$, a learned representation $\mathbf{g}(\mathbf{X})$ identifies $\mathbf{Z}$ *up to permutation and scaling (PS)* when there exists a permutation matrix P and an invertible diagonal matrix D such that $\mathbf{g}(\mathbf{X}) = PD\mathbf{Z}$ almost surely. A weaker notion of identifiability is *up to a global affine transformation (AT)*, which instead means that there exists a global invertible *affine transformation* $\mathbf{h}$ such that $\mathbf{g}(\mathbf{X}) = \mathbf{h}(\mathbf{Z})$ almost surely.

In this paper, we assume that the observations $\mathbf{X}$ are generated by an injective continuous piecewise affine mixing function $\mathbf{f}$ (Def. B.12). We also assume that $\mathbf{Z}$ follows a *potentially degenerate Gaussian mixture model* (pdGMM) with unknown parameters, i.e., a Gaussian mixture model that explicitly allows for degenerate components (with singular covariance matrices). Gaussian Mixture Models (GMM) are a popular method in clustering. While the standard definition also allows for degenerate components, in practice there is often the assumption that the components are not degenerate. Since in this paper we focus on this setting, we make this choice explicit by slightly modifying their name and defining them as follows:

Definition 2.2. A variable $\mathbf{Z}$ follows a potentially degenerate Gaussian mixture model (pdGMM) if its distribution is $P = \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where for each component $j \in [J]$ the mean is $\boldsymbol{\mu}_j \in \mathbb{R}^n$ and the covariance matrix is $\boldsymbol{\Sigma}_j \in \mathbb{R}^{n \times n}$, such that determinant $|\boldsymbol{\Sigma}_j| \geq 0$. We assume that the weights $\lambda_j > 0$ and $\sum_{j=1}^{J} \lambda_j = 1$. If each component is distinct, i.e., for every $i \neq j \in [J]$ we have $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \neq (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, then we say that the pdGMM is in a *reduced form.*

Other works have already studied latent variables that follow a non-degenerate GMM distribution and considered the problem of identifying these variables from observation. In particular, Kivva et al. (2022) provide the identifiability up to AT of non-degenerate GMMs latent variables with piecewise linear mixing functions. For a stronger notion of identifiability, identifiability up to PS, Kivva et al. (2022) require that all pairs of latent variables must be conditionally independent given an auxiliary variable U, which is an assumption we will not make. Crucially, these results rely on the *analyticity* of the probability density function (pdf) of Gaussian distributions, which does not hold for degenerate Gaussians, since their pdf is no longer well-defined. This means we cannot apply these results directly.

In this challenging setting, we instead leverage results from a different line of work focusing on sparsity (Lachapelle et al., 2023a; Xu et al., 2024) to prove the identifiability up to PS and a constant translation for piecewise affine mixing functions. Lachapelle et al. (2023a) focus on multi-task learning with sparse task-specific predictors, while Xu et al. (2024) focus on partially observable causal representation learning settings, where each observation only captures a small number of "active" latent variables. While these works focus on sparsity, their motivation is related to our choice to allow for potential degenerate representations, since sparse (i.e., low-rank) representations can be seen as representations in which some of the components are degenerate. In many real-world settings we have (nearly)degenerate structures, such as high collinearity or strong dependencies among features, or intrinsically low-rank manifolds. This is typical for

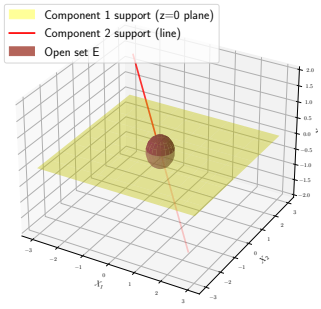

Figure 1: Example of a 3-dimensional pdGMM with 2 components: a rank-2 component with support represented by the yellow plane, and a rank-1 component with support represented by the red line. Knowing the distribution on the open set E (in brown), intersecting both components, is sufficient to identify the pdGMM.

high-dimensional data that require large numbers of sparse features such as language models (Cunningham et al., 2024; Gao et al., 2025; Marks et al., 2025).

3 IDENTIFIABILITY OF PDGMMS

In this section, we provide a series of theoretical results that prove progressively stronger notions of identifiability for latent variables following a potentially degenerate Gaussian Mixture Model (pdGMM) distribution with a piecewise affine mixing function. We provide all proofs in Appendix B. As a first step, we focus on the case in which the pdGMM variables are observed and prove that if two pdGMM variables have the same distribution on an open set, then they have the same distribution over all of $\mathbb{R}^n$, i.e., they are *identified*. The intuition is illustrated in the example in Fig. 1. In addition to being of independent interest, this result is an important step for our other results.

3.1 Identifiability from an open subset

Identifiability results for (pd)GMMs from the whole domain are well-established (Yakowitz and Spragins, 1968), proving that if two (pd)GMMs are the same on their domain, then they have the same components and weights up to a permutation. These results include the potentially degenerate case. On the other hand, in our quest to prove the identifiability of *latent variables* that follow a pdGMM only from observations, we need an identifiability result for observed pdGMMs on a *subdomain*, and more precisely, an open subset. We highlight this result in this subsection, because it is of independent interest beyond the CRL setting.

Since we cannot rely on equality of pdfs because of potential degeneracy, we first need a definition of when two distributions are considered equal on a subdomain.

Definition 3.1. Let $\mu : \mathcal{F} \to [0,1]$ and $\mu' : \mathcal{F} \to [0,1]$ be the probability measures induced by $\mathbf{X}$ and $\mathbf{X}'$, respectively, where $\mathcal{F}$ is the Borel sigma-algebra over $\mathbb{R}^n$. Given a subset $E \in \mathcal{F}$, we define the restriction of μ to E, denoted by $\mu_E : \mathcal{F} \to [0,1]$, as $\mu_E(A) := \mu(A \cap E), \forall A \in \mathcal{F}$. We say that $\mathbf{X}$ and $\mathbf{X}'$ are equal in distribution over a subdomain E, denoted as $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, when $\mu_E = \mu'_E$.

We show that the equality of two pdGMMs on an open subset of $\mathbb{R}^n$ that intersects the support of every Gaussian component in a pdGMM implies the equality of pdGMMs on the whole domain.

Theorem 3.2 (Identifiability of pdGMMs from open set)**.** *Consider two pdGMMs in $\mathbb{R}^n$ in reduced form,*

$$\mathbf{X} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \text{and} \quad \mathbf{X}' \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{1}$$

where $\text{rank}(\boldsymbol{\Sigma}_j) > 0, \forall j \in [J]$ and $\text{rank}(\boldsymbol{\Sigma}'_j) > 0, \forall j \in [J']$. Let $\mathcal{X}_j$ be the supports of $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. If there exists an open set E such that for all $j \in [J]$, $\mathcal{X}_j \cap E \neq \varnothing$ and $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, then $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on the whole domain, $J = J'$, and there exists a permutation $\pi : [J] \to [J']$ such that $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi(j)}, \boldsymbol{\mu}'_{\pi(j)}, \boldsymbol{\Sigma}'_{\pi(j)})$.

We provide the proof in App. B.1 and describe a proof sketch here. The main challenge is the ill-defined pdf of pdGMMs. Kivva et al. (2022) provides similar results for non-degenerate GMM, but their proof strategy relies on the analyticity of the pdf. However, in the degenerate case, the density does not exist on $\mathbb{R}^n$, so their proof cannot be applied. We overcome this issue by projecting a pdGMM in $\mathbb{R}^n$ into a sequence of lower-dimensional spaces $\mathbb{R}^k$, where $k \in [n]$, such that each degenerate Gaussian component becomes non-degenerate in at least one image space. This enables us to apply classical identifiability results to recover the parameters of Gaussian components grouped by their ranks (Yakowitz and Spragins, 1968; Kivva et al., 2022).

To make this strategy work, the projections must satisfy two key properties. First, they must preserve the rank of each component's covariance matrix, thereby maintaining the intrinsic structure and information of the components. Second, a single projection is insufficient to fully recover the parameters of a pdGMM from lower-dimensional spaces; instead, a sufficiently rich set of rank-preserving projections is required to capture all distinguishable parameterizations. We therefore choose to work with a dense subset of the space of projections, and prove that such set of projections always exists.

3.2 Identifiability of latent pdGMMs

In this section, we introduce our main results (Thm. 3.5, 3.7, and 3.9), which form a set of increasingly stronger assumptions on the distribution of $\mathbf{Z}$, each yielding a correspondingly stronger identifiability. As a first step we introduce a weak notion of identifiability for mixture-based latent variables (i.e., the distribution is a mixture of multiple components), *identifiability up to affine transformations within components (ATwC)*. This notion resembles the identifiability up to an affine transformation (AT) from literature, but it provides even less guarantees, since the recovered variables identify the ground truth latent variables only up to an affine function $\mathbf{h}$ within the support of each component, but without any guarantee that $\mathbf{h}$ is affine across all components, thus sacrificing the interpretability of the learned representations. We provide Example E.3 to illustrate the difference between *ATwC* and *AT*. Formally, we define identifiability up to ATwC as:

Definition 3.3. The ground truth mixture-based representation vector $\mathbf{Z}$ is said to be *identified up to affine transformations within components (ATwC)* by a learned representation vector $\mathbf{g}(\mathbf{X})$ when there exists an invertible transformation $\mathbf{h}$, such that $\mathbf{g}(\mathbf{X}) = \mathbf{h}(\mathbf{Z})$ almost surely and $\mathbf{h}$ is *affine* on each $\mathcal{Z}_j$, where $\mathcal{Z}_j$ is the support of j-th component.

To prove identifiability up to ATwC, we introduce a genericity condition on pdGMMs. We consider pdGMMs for which the supports of components with the same rank k may overlap, but there is at least one common point in the intersection where the components can be distinguished, i.e., their projections onto the non-degenerate subspace do not have the same norm. This rules out special cases where two components are geometrically indistinguishable in their overlapping region, which we prove in App. E.2 is a measure zero set in the parameter space under consideration. Intuitively, this means that this assumption only rules out highly symmetric or finetuned cases. We formalize it as follows:

Assumption 3.4 (Genericity of pdGMMs)**.** Consider a pdGMM in reduced form $\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Let $\mathcal{J}_k := \{j \in [J] : rank(\boldsymbol{\Sigma}_j) = k)\}$ be the indices of the components with rank k. For any $k \leq n$ and any subset $\mathcal{J}_k^0 \subseteq \mathcal{J}_k$, suppose that the intersection of the supports of the components $\mathcal{Z}_j$ is non-empty, i.e., $\cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j \neq \varnothing$. We assume there exists at least one point $\mathbf{z}_0 \in \cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j$ s.t. the Mahalanobis distance [1] of $\mathbf{z}_0$ to the distributions of any two distinct Gaussian components $i, j \in \mathcal{J}_k^0$ is not the same.

[1] Mahalanobis distance of a point $\mathbf{z}_0$ from a distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is $\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_j)}$. When $\boldsymbol{\Sigma}_j$ is singular, we use the Moore-Penrose inverse for $\boldsymbol{\Sigma}_j^{-1}$

Based on Ass. 3.4, we show that we can learn a pdGMM representation up to *ATwC* with a perfect reconstruction and enforcing Gaussianity on representations.

Theorem 3.5 (Identifiability up to ATwC of pdGMM latent variables)**.** *Assume that $\mathbf{Z}$ is a pdGMM and the observation $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ follows the data-generating process in Sec. 2 for an injective continuous piecewise affine mixing function $\mathbf{f}$, and $\mathbf{Z}$ satisfies Ass. 3.4. Let $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ be an invertible continuous piecewise affine function and $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ be an invertible piecewise affine function onto its image. If both conditions hold,*

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

$$\mathbf{g}(\mathbf{X}) \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{3}$$

for appropriate values of $\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j$ (in reduced form), then J'=J and $\mathbf{Z}$ is identified by $\mathbf{g}(\mathbf{X})$ up to affine transformation within components, i.e., $\mathbf{g} \circ \mathbf{f}$ is an invertible affine function on $\mathcal{Z}_j$, $j \in [J]$.

We provide a proof in App. B.2.1. While Thm 3.5 obtains the identifiability up to ATwC, this result does not guarantee that there is a single *global affine transformation* $\mathbf{h}$ such that $\mathbf{g}(\mathbf{X}) = \mathbf{h}(\mathbf{Z})$ for all components supports. To achieve this stronger identifiability, the identifiability up to AT, we need an additional condition on $\mathbf{Z}$. The following assumption requires that the supports of all components intersect at at least one point $\mathbf{z}^0$ and that each of these subspaces is spanned by a subset of a shared global basis.

Assumption 3.6 (Common basis and translation vector)**.** Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_J \subseteq \mathbb{R}^n$ be affine spaces such that there exists $\mathbf{z}^0 \in \bigcap_{j=1}^{J} \mathcal{Z}_j$. We assume there exists a basis of $\mathbb{R}^n$, denoted by $\{\mathbf{z}^1, \ldots, \mathbf{z}^n\}$, such that for all $j \in [J]$, there exists $\mathcal{K}_j \subseteq [n]$ s.t. $\{\mathbf{z}^k \mid k \in \mathcal{K}_j\}$ is a basis for the vector space $\mathcal{Z}_j - \mathbf{z}^0$.

This assumption always holds for non-degenerate components. For degenerate components, a violation can happen because the components are not intersecting, or if their individual bases do not form a global shared basis. In App. E.3 we discuss a concrete scenario in which this assumption holds, where only a subset of latent variables are "active" in each observation sample, e.g., they are captured in an image, while others are "inactive", i.e., they are masked by a constant. In this setting, the intersection of the supports is the vector of all the constant values for each variable, which will form a well-defined global shared basis. If this assumption only holds for a subset of components, then we can achieve identifiability up to AT for this subset, while

for the rest can only achieve identifiability up to ATwC using the previous result. We provide an example to show the necessity of this assumption in App. E.3 and use it to achieve identifiability up to AT:

Theorem 3.7 (Identifiability up to AT of pdGMM latent variables)**.** *We assume the same conditions as Thm. 3.5, and that the supports of components $\mathcal{Z}_1, \dots, \mathcal{Z}_J$ satisfy the common basis assumption (Ass. 3.6), then the representation $\mathbf{g}(\mathbf{X})$ identifies $\mathbf{Z}$ up to a single global affine transformation across components, i.e., $\mathbf{g} \circ \mathbf{f}$ is affine on all of $\mathcal{Z}$.*

We provide a proof in App. B.2.2. Identifiability up to an affine transformation (AT) still allows for arbitrary mixing of dimensions through general affine transformations. In many applications, we prefer representations where individual dimensions have separate and interpretable meanings. This motivates us to move to identifiability up to *permutation and scaling* (PS). Permutation and scaling are fundamental indeterminacies that typically cannot be solved without supervision or auxiliary information (Hyvärinen et al., 2024). To get this stronger result, we require a stricter assumption on $\mathbf{Z}$.

Assumption 3.8. Let $\mathcal{Z}_1, \dots, \mathcal{Z}_J \subseteq \mathbb{R}^n$ be the supports of the components of $\mathbf{Z}$. We assume:

- [Common Standard Basis] For all $j \in [J]$, each $\mathcal{Z}_j$ is a vector space, and there exists an index set $\mathcal{K}_j \subseteq [n]$ such that $\{\mathbf{e}^k \mid k \in \mathcal{K}_j\}$ forms a basis for $\mathcal{Z}_j$, where $\{\mathbf{e}^1, \dots, \mathbf{e}^n\}$ denotes the standard basis of $\mathbb{R}^n$ (one-hot vectors).
- [Sufficient Support Basis Index Variability] For every $i \in [n]$, the union of the support index sets $\mathcal{K}_j$ that do not include i covers all other standard basis directions:

$$\bigcup_{j \in [J] \mid i \notin \mathcal{K}_j} \mathcal{K}_j = [n] \setminus \{i\}.$$

This setting is consistent with the sparsity principle proposed by Xu et al. (2024), which shows that by enforcing sparsity of the transformed variables (as expressed in Eq. 4), the corresponding transformation is an element-wise affine function. The Common Standard Basis assumption is a special case of the earlier Common Basis Assumption (Ass. 3.6) where the global basis is fixed to the standard basis and zero translation vector is applied. This implies that the realizations of $\mathbf{Z}$ will have some coordinates equal to 0 with probability greater than zero. For example, assuming $n = 3$ with $\mathcal{K}_1 = \{1\}$ and $\mathcal{K}_2 = \{2, 3\}$, we have that samples from the component $j = 1$ will have the form $(\mathbf{z}_1, 0, 0)$ and samples from the component $j = 2$ will have the form $(0, \mathbf{z}_2, \mathbf{z}_3)$. In other words, the realizations of $\mathbf{Z}$ are sparse, with nonzero coordinates given by $\mathcal{K}_j$.

The Sufficient Support Basis Index Variability was originally proposed by Lachapelle et al. (2023a) in the context of sparse multitask learning. Here, we adapt it into our context and use it to avoid cases in which certain latent variables are always degenerate together across all components, since then we would not be able to disentangle them from the observations. As we show in App. E.5, even if this assumption does not hold strictly in a setting, our framework still supports a weaker form of identifiability: block-wise identifiability, which allows disentanglement across blocks of latent variables, even if variables within each block may remain entangled. This relaxation still enables meaningful interpretability and structure in learned representations. We can now prove our final theoretical result that shows when we can disentangle, or identify pdGMM latent variables up to PS.

Theorem 3.9 (Identifiability up to PS of pdGMM latent variables)**.** *We assume the same conditions as Thm. 3.5 and that the supports of components $\mathcal{Z}_1, \dots, \mathcal{Z}_J$ satisfy Ass. 3.8. If the following holds:*

$$\mathbb{E} \left\| \mathbf{g}(\mathbf{X}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0, \tag{4}$$

where $\|\cdot\|_0$ means L0 norm, then the representation $\mathbf{g}(\mathbf{X})$ identifies $\mathbf{Z}$ up to PST, i.e., $\mathbf{g} \circ \mathbf{f}$ is a permutation combined with an element-wise linear transformation on $\mathcal{Z}$.

We provide a proof in App. B.2.3. Interestingly, a special case of this result also solves an open question in previous work on partially observed causal representation learning (CRL) (Xu et al., 2024). In that setting the identifiability results required that the component index j is observed for each observation (that is the group index), which is a strong assumption. Modelling this partially observable CRL problem with mixture models allows one to consider settings in which we do not know how to group the observations, i.e., a more realistic partially observable CRL setting.

4 IMPLEMENTATION

We implement our theoretical results in two stages: the identifiability up to AT in Thm. 3.7 serves as a prerequisite for achieving the identifiability up to PS in Thm. 3.9. In the first stage, we use an autoencoder structure to estimate the latent variables directly from observations $\{\mathbf{x}^i\}_{i \in [N]}$ by minimizing the reconstruction error (Eq. 2), where $\mathbf{g}_{\psi_1}$ and $\hat{\mathbf{f}}_{\theta_1}$ denote the encoder and decoder, respectively. To prevent the latent codes from drifting arbitrarily and keep them clustered around the Gaussian prior, we include the L_2 norm in the loss function. According to Thm. 3.7, the representations learned in this stage correspond to an

affine transformation of the ground truth latent variables. In the second stage, we freeze the autoencoder model trained in the first stage and apply a second, inner autoencoder with affine transformations, where $\mathbf{g}_{\psi_2}$ and $\hat{\mathbf{f}}_{\theta_2}$ denote the affine encoder and decoder. Following Xu et al. (2024), we approximate the sparsity constraint in Eq. 4 of Thm. 3.9 by replacing the non-differentiable L_0 norm with the differentiable L_1 norm except at zero. As the ground truth sparsity level is not known, we use a tunable hyperparameter ϵ to control it (we provide sensitivity analysis of ϵ in App. D.5). Combined with the reconstruction loss, this encourages the model to converge to an appropriate sparsity level. To solve the optimization problem, we use Cooper (Gallego-Posada and Ramirez, 2022), a Lagrangian-based optimization toolkit. The two-stage problems are solved using Adam (Kingma and Ba, 2015) and extra-gradient variant (Gidel et al., 2019).

$$\textbf{Stage 1:} \quad \min_{(\theta_1,\psi_1)} \frac{1}{N} \sum_{i\in[N]} \left\| \mathbf{x}^i - \hat{\mathbf{f}}_{\theta_1}(\mathbf{g}_{\psi_1}(\mathbf{x}^i)) \right\|_2^2 + \quad (5)$$
$$log(p(\mathbf{g}_{\psi_1}(\mathbf{x}^i); \mathbf{0}, I))$$
$$\textbf{Stage 2:} \quad \min_{(\theta_2,\psi_2)} \frac{1}{N} \sum_{i\in[N]} \left\| \tilde{\mathbf{x}}^i - \hat{\mathbf{f}}_{\theta_2}(\mathbf{g}_{\psi_2}(\tilde{\mathbf{x}}^i)) \right\|_2^2 \quad (6)$$
$$\text{subject to:} \ \frac{1}{N} \sum_{i\in[N]} \left\| \mathbf{g}_{\psi_2}(\tilde{\mathbf{x}}^i) \right\|_1 \leq \epsilon,$$

where $\tilde{\mathbf{x}}^i := \mathbf{g}_{\psi_1}(\mathbf{x}^i)$. We provide all details in App. C.

5 EXPERIMENTAL RESULTS

Numerical Experiments Setup. We generate simulations with variations in the underlying causal model, the rank of Gaussian components, and mixing function $\mathbf{f}$. We average results for each setup over 5 random seeds. We consider $n = \{5, 10, 20, 40\}$ latent variables $\mathbf{Z}$. For each experiment, we sample a directed acyclic graph $\mathcal{D}$ from an Erdös-Rényi (ER)-k model with $k \in \{0, 1, 2, 3\}$, where each ER-k graph contains $n \cdot k$ directed edges. In particular, ER-0 corresponds to an empty graph, which implies independent latent variables. Given $\mathcal{D}$, we simulate a linear Gaussian Structural Causal Model where edge weights are sampled uniformly from $[-1, -0.2] \cup [0.2, 1]$, and standard Gaussian noise is added at each node. To simulate pdGMMs, we define the number of mixture components $J = 5n$ and introduce the ratio of non-degenerate dimensions $\rho \in \{1\text{VAR}, 50\%, 75\%\}$, where 1VAR indicates only one non-degenerate dimension, and 50% or 75% are the proportion of non-degenerate dimensions relative to n.

Then, we randomly sample J different ρn-hot vectors to define a set of basis index sets $\mathcal{K}_j$, which determines the subspace (i.e., the support) where the component has non-degenerate variance. For dimensions not in $\mathcal{K}_j$ (i.e., those with a 0 in the multi-hot vector), we set their value to the corresponding entry in a predefined translation vector $\mathbf{z}^0$. To further show the different results for Ass. 3.6 and Ass. 3.8, we vary the translation vector $\mathbf{z}^0 = \boldsymbol{\mu} + \delta \cdot \sigma$, where $\boldsymbol{\mu}, \sigma \in \mathbb{R}^n$ are the mean and standard deviation of $\mathbf{Z}$, and $\delta \in \{0, 1, 2, 3\}$. In addition, we apply a $n \times n$ rotation matrix on $\mathbf{Z}$, parameterized by a rotation angle $\theta \in \{0°, 15°, 30°, 45°\}$, to control the deviations from standard basis. ll settings satisfy Ass. 3.6, which means we can always achieve identifiability up to AT, but Ass. 3.8 is only satisfied in the setting with $\delta = 0$ and $\theta = 0$, and thus we expect identifiability up to PS to only be possible there.

To simulate the observed variables $\mathbf{X}$, we apply an invertible piecewise affine mixing function $\mathbf{f}$ to the latent variables $\mathbf{Z}$. We parameterize $\mathbf{f}$ by a multilayer perceptron (MLP) with $m \in \{3, 10, 20\}$ hidden layers, $(m-1)$ Leaky-ReLU activation functions (with slope $\alpha \sim \text{Unif}(0.5, 1.5)$) and a final affine layer. The number of layers m is a proxy for the non-linearity and complexity of the mixing function; the larger the m, the more complex the transformation from $\mathbf{Z}$ to $\mathbf{X}$.

Results for Stage 1 (Thm. 3.7). For the encoder $\mathbf{g}_{\psi_1}$ and the decoder $\hat{\mathbf{f}}_{\theta_1}$ in the first stage, we employ 5-layer MLPs configured with $[50, 100, 100, 50] \times n$ units per layer, where n is the number of latent variables. Each layer uses Leaky-ReLU activations, except for the final one. Batch normalization is applied to control the norm of $\mathbf{g}_{\psi_1}(\mathbf{X})$. To validate the identifiability up to AT presented in Thm. 3.7, the metric we use is the R^2 of the linear regression model with $\mathbf{Z}$ as explanatory variable and $\mathbf{g}_{\psi_1}(\mathbf{X})$ as response. A high R^2 indicates that the ground truth latent variables can be linearly mapped to the learned representation with low error, consistent with our identifiability result. We show the average R^2 over five random seeds in Tab. 1. Our method remains robust with large n, dense causal graph k, varying translation vectors $\mathbf{z}^0(\delta)$ and basis direction θ. Performance decreases slightly only with increased nonlinearity (m) and lower proportions of nondegenerate variables (ρ). This is expected: higher nonlinearity increases the complexity of achieving affine identifiability, while lower ρ allows greater flexibility in how representations are distributed in $\mathbb{R}^n$, making it more challenging to enforce Gaussianity. However, a high R^2 does not imply disentanglement. Following prior work (Khemakhem et al., 2020; Kivva et al., 2022), we also evaluate identifiability up to *PS* via *Mean Correlation Coefficient* (MCC), a metric detailed in App. D.1. Across all row groups, we observe that, as expected, Stage 1 *alone* does not identify the latent variables up to PS accurately, as reflected by the low MCC values reported in the third right column of

n	k	m	ρ	δ	θ	$\mathbf{R}^2 \uparrow$ Stage1	MCC Stage1	MCC ↑ Stage2	MCC ↑ VaDE
5	1	10	50%	0	0	0.93	0.75	0.96	0.45
10	1	10	50%	0	0	0.94	0.56	0.97	0.32
20	1	10	50%	0	0	0.93	0.46	0.92	0.23
40	1	10	50%	0	0	0.94	0.37	0.94	0.25
10	**0**	10	50%	0	0	0.94	0.51	0.97	0.33
10	**1**	10	50%	0	0	0.94	0.56	0.97	0.32
10	**2**	10	50%	0	0	0.93	0.58	0.95	0.31
10	**3**	10	50%	0	0	0.92	0.56	0.91	0.27
10	1	**3**	50%	0	0	0.99	0.50	0.96	0.33
10	1	**10**	50%	0	0	0.94	0.56	0.97	0.32
10	1	**20**	50%	0	0	0.88	0.50	0.93	0.35
10	1	10	**1 var**	0	0	0.75	0.46	0.45	0.28
10	1	10	**50%**	0	0	0.94	0.56	0.97	0.32
10	1	10	**75%**	0	0	0.95	0.57	0.93	0.31
10	1	10	50%	**0**	0	0.94	0.56	0.97	0.32
10	1	10	50%	**1**	0	0.95	0.59	0.88	0.30
10	1	10	50%	**2**	0	0.96	0.59	0.61	0.27
10	1	10	50%	**3**	0	0.96	0.58	0.49	0.28
10	1	10	50%	0	**0**	0.94	0.56	0.97	0.32
10	1	10	50%	0	**15**	0.93	0.58	0.83	0.32
10	1	10	50%	0	**30**	0.93	0.57	0.60	0.31
10	1	10	50%	0	**45**	0.93	0.57	0.62	0.32

Table 1: Results for the numerical experiments. The bold font highlights the parameters that vary in each block. $\mathbf{R}^2$(Stage 1) shows identifiability up to AT after Stage 1 **MCC**(Stage 1) reflects the necessity of sparsity to achieve identifiability up to PS; **MCC**(Stage 2) presents the identifiability up to PS after Stage 2; **MCC**(VaDE) reports the results of Kivva et al. (2022). Full results with Std. Dev. are in Tab. 3.

Tab. 1. This motivates us to move to Stage 2.

Results for Stage 2 (Thm. 3.9). For the affine encoder $\mathbf{g}_{\psi_2}$ and the decoder $\hat{\mathbf{f}}_{\theta_2}$ in the second stage, we employ 7-layer MLPs configured with $[10, 50, 50, 50, 50, 10] \times n$ units per layer, where n is the number of latent variables. Batch normalization is applied to control the norm of $\mathbf{g}_{\psi_2}(\mathbf{X})$. Since we generate a sufficient number of basis indices, our representation should be able to have a one-to-one correspondence with the ground truth latent variables. The last two columns of Tab. 1 reports the average MCC scores for our method and VaDE (Kivva et al., 2022). This is a misspecified setting for VaDE, because VaDE requires non-degeneracy of the data, but we report the results nevertheless as a baseline, and show that as expected we outperform it in all settings. The results indicate consistently good performance across a range of n, k, and m. However, the overall performance is influenced by the quality of representations from Stage 1, with a noticeable decrease under a low nondegeneracy ratio ρ. This is expected, since large translation vectors $\mathbf{z}^0(\delta)$ and non-standard basis (θ) violate Ass. 3.8.

Ablations. We provide an evaluation on independent latent variables in App. D.3, showing comparable results, and an empirical complexity analysis in App. D.4, showing that we can scale our method up to $n = 50$ and partially to $n = 100$ variables. This is in line with other CRL works, which mostly focus on $10 - 20$ latent variables. On the other hand, the inference time increases exponentially with the number of latents. In App. D.5 we report a sensitivity analysis on our hyperparameters, in particular the sparsity level, which we set to $\epsilon = 0.01$ for the main experiments, and the learning rate, which we set to $lr = 1e - 4$. In App. D.6 we report an ablation on the sparsity constraint, showing its importance for the second stage. We also compare with a related method (Xu et al., 2024) that requires additional information (i.e., knowing the label of the component) and show that with the same information, our method provides comparable results. For completeness, we also provide results in App. D.7 with a setting that is fair to the baseline VaDE (Kivva et al., 2022), where we generate the latent with a non-degenerate GMM, showing that our method outperforms VaDE even in these settings. Similarly to other work in CRL, our theoretical results assume that we know the ground truth n, but empirically our method performs well even if we overestimate it, as shown in App. D.8. As an additional measure of the accuracy of our reconstruction, in App. D.9 we evaluate the Structural Hamming Distance (SHD) of the causal graph that we learn with a standard causal discovery method, PC (Spirtes et al., 2000), on our recovered latent variables, showing that the SHD is only slightly worse than the SHD for the ground truth variables.

Settings when the assumptions are violated. In App. E we discuss the intuition, implications and practical applicability of our assumptions, providing also ablations for their violations. In particular, in App. E.1 we show empirically that our method is often still robust in a misspecified setting, i.e., when the variables are not Gaussian (but instead they follow and exponential or Gumbel distribution) or the mixing function is not piecewise affine (but instead is a smooth Leaky-ReLU or sigmoid). On the other hand, we do need all other assumptions. We provide evaluation showing the necessity of Ass. 3.4 (genericity of pdGMMs) in App. E.2, showing that when it does not hold the results are much worse. Similarly, results are much worse when Ass. 3.6 does not hold, so there does not exists a unique shared basis for all Gaussian components. To show a violation of this, we generate the data with a mixture of two different bases: a standard basis and one with rotations, and provide results in App. E.3, showing the necessity of this assumption. Table 1 already shows violations of Ass. 3.8a when δ (the masking value) and θ

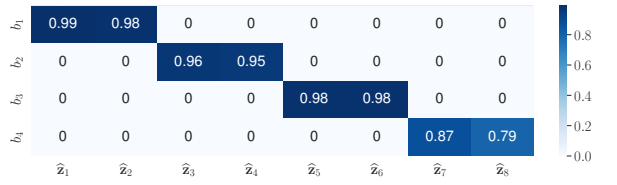

Figure 2: R^2 of the linear regression between ball positions $b_1, \ldots, b_4$ and learned representations $\hat{z}_1, \ldots, \hat{z}_8$. We use two variables for each (x, y) position.

(the rotation of the standard basis) are non-zero, showcasing a decrease in performance when the assumption does not hold. When we have violations of Ass. 3.8b, we can still achieve block-identifiability, as discussed and shown empirically in App. E.5.

Image dataset: Multiple Balls. To show potential applications of our approach in partially observable causal representation learning, we evaluate our method on an image dataset from Ahuja et al. (2022); Xu et al. (2024), which consists of $\mathbf{b}$ moving balls rendered in a 2D space, as shown in App. D.10. The latent causal variables correspond to the (x, y) positions of each ball, which are modeled as Gaussian. We focus on a *fixed position* setting, where the balls usually move freely within the frame but occasionally remain stationary at unknown fixed positions. When a ball is stationary, the corresponding latent dimensions are degenerate. We generate datasets with varying numbers of balls: $\mathbf{b} = 2, 4, 6$. The (x, y) coordinates of each ball $i \in [\mathbf{b}]$ are sampled independently from a truncated 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ bounded within the square $(0.05, 0.95)^2$, where $\boldsymbol{\mu}_i \sim \text{Unif}(0.3, 0.7)^2$, and $\boldsymbol{\Sigma}_i = ((0.01, 0.00)(0.00, 0.01))$. To avoid the high ratio of occlusion, we use rejection sampling to sample $\boldsymbol{\mu}_i$ to ensure the Euclidean distance between any pair of $\boldsymbol{\mu}_i$ values is at least 0.2. Each ball has a probability $p = 0.1$ to be at a fixed position $\boldsymbol{\mu}_i$.

Before we train the model described in Sec. 4, we pre-train a CNN model to reduce the input image dimension from 64×64 to a vector of size $16\mathbf{b} \times 1$. More details about the network architecture are provided in App. D.10. Instead of a fixed-size training dataset, we generate images online until convergence. Since the variables describing the x and y position of the same ball are always degenerate together, they cannot be disentangled in our method (as sufficient variability in Ass. 3.8 does not hold). To evaluate the performance, we perform regression from each representation separately to the corresponding ground-truth position pairs, as illustrated in Fig. 2. The results show that the position of each ball can be recovered through pairs of representations, which is consistent with our theoretical results. Additional results are in App. D.10.

6 RELATED WORK

Our work is inspired by previous work in nonlinear independent component analysis (ICA), causal representation learning (CRL) and sparsity-based methods. ICA (Comon, 1994; Hyvärinen, 2013) is a seminal approach that identifies latent variables under the assumption that the latent variables are independent and that the mixing function is linear. In the non-linear ICA setting, Hyvarinen and Morioka (2016, 2017) reformulate this standard ICA independence assumption as conditional independence given observed auxiliary variables. Khemakhem et al. (2020) show the identifiability of the VAE model with a conditionally independent prior. Willetts and Paige (2021) show that the auxilliary variables can be unobservable and learned; however, this still requires the conditional independence assumption.

Causal representation learning (CRL) extends (nonlinear) ICA to the case in which latent variables have causal relations between them and hence arbitrary dependencies. Our approach fits in this line of work. Many CRL works consider additional data or information to identify causal variables, e.g., interventions (Brehmer et al., 2022; Lippe et al., 2022; Varici et al., 2023; von Kügelgen et al., 2023; Buchholz et al., 2023; Zhang et al., 2023; Ahuja et al., 2023a), actions or temporal structure (Lippe et al., 2023; Lachapelle et al., 2022, 2024), multi-view settings (von Kügelgen et al., 2021; Ahuja et al., 2023b; Yao et al., 2024; Morioka and Hyvärinen, 2024), multiple environments (Liu et al., 2024), grouping of observations by sparsity patterns (Xu et al., 2024), hierarchical structures (Kong et al., 2023, 2024) or conditions on causal graphs (Zhang et al., 2024; Dong et al., 2024; Song et al., 2024).

In our case, we do not assume any additional data or information besides the observations, but instead consider parametric assumptions on the latent variables and mixing function. Similarly to us, Kivva et al. (2022) studies the identifiability of non-degenerate GMM latents with piecewise affine mixing. For PS identifiability, they require that all pairs of latent variables must be conditionally independent given an auxiliary variable. Other works consider restrictions on the mixing function, e.g., Squires et al. (2023); Bing et al. (2024) assume a linear mixing function, Ahuja et al. (2023a) assumes a finite-degree polynomial and either independent supports or perfect interventions, and Lachapelle et al. (2023b) assumes an additive mixing function. Our results do not require conditional independence, independent support or interventions and allow mixing functions composed of an infinite number of affine pieces, allowing the potential possibility to approximate

any nonlinear functions up to an arbitrary precision.

Our focus on low-rank representations is also inspired by prior work that leverages sparsity in (causal) representation learning, e.g., structural sparsity in Jacobians of nonlinear mixing functions (Zheng et al., 2022; Zheng and Zhang, 2023), sparsity constraints on affine mixing (Ng et al., 2023), sparsity on causal graphs (Zhang et al., 2024; Song et al., 2024), sparse decoders (Moran et al., 2022) and sparsity task-specific predictors (Lachapelle et al., 2023a; Fumero et al., 2023). For time-series data, Lachapelle et al. (2022, 2024) propose sparse actions and sparse temporal dependencies, while Li et al. (2025) introduce latent process sparsity, referring to sparse connections between latents both within and across time. Differently from these approaches, which impose sparsity on the model structure or latent interactions, we are enforcing *sparsity of the representation itself*. This aligns with prior work (Tonolini et al., 2020), as well as classical methods *sparse component analysis* (Gribonval and Lesage, 2006), and *sparse dictionary learning* (Mairal et al., 2009). Most works show empirical success on benchmark tasks and do not provide guarantees for identifiability, which is our main focus. Moran et al. (2022) focuses on independent latent variables, while Hu and Huang (2023) studies the identifiability of dictionary learning, but they assume *affine mixing*. Instead, we focus on *piecewise-affine mixing* for dependent latent variables.

7 CONCLUSIONS AND LIMITATIONS

In this work, we focus on identifying pdGMM latent variables with piecewise affine mixing functions. We first show that the entire pdGMM can be identified via an open set. This result is a critical step for a series of increasingly stronger identifiability results, ultimately proving identifiability up to permutation and element-wise transformations, i.e. completely disentangled representations. While our experiments validate our results, there are still several limitations. As discussed in App. E, some of our assumptions might not hold in some settings, potentially leading to partial or no theoretical guarantees. In particular, the Gaussian assumption for mixture components may not hold, so extending our approach to other parametric models is an exciting avenue for future research. Finally, although our method encourages Gaussianity on representations, ensuring this in practice remains challenging.

Acknowledgements

We thank the anonymous reviewers for their helpful and constructive comments. 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.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

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

Aapo Hyvärinen. Independent component analysis: recent advances. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110534, 2013.

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.

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.

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.

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.

Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *Causal Representation Learning Workshop at NeurIPS 2023*, 2023.

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.

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.

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.

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

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.

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, 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.

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.

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.

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. *International Conference on Learning Representations*, 2024.

Hiroshi Morioka and Aapo Hyvärinen. Causal representation learning made identifiable by grouping of observational variables. *International Conference on Machine Learning*, 2024.

Danru Xu, Dingling Yao, Sébastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. *International Conference on Machine Learning*, 2024.

Lingjing Kong, Biwei Huang, Feng Xie, Eric Xing, Yuejie Chi, and Kun Zhang. Identification of nonlinear latent hierarchical models. *Advances in Neural Information Processing Systems*, 36:2010–2032, 2023.

Lingjing Kong, Guangyi Chen, Biwei Huang, Eric P. Xing, Yuejie Chi, and Kun Zhang. Learning discrete concepts in latent hierarchical models. *Advances in Neural Information Processing Systems*, 2024.

Kun Zhang, Shaoan Xie, Ignavier Ng, and Yujia Zheng. Causal representation learning from multiple distributions: A general setting. *International Conference on Machine Learning*, 2024.

Xinshuai Dong, Ignavier Ng, Biwei Huang, Yuewen Sun, Songyao Jin, Roberto Legaspi, Peter Spirtes, and Kun Zhang. On the parameter identifiability of partially observed linear causal models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Xiangchen Song, Zijian Li, Guangyi Chen, Yujia Zheng, Yewen Fan, Xinshuai Dong, and Kun Zhang. Causal temporal representation learning with nonstationary sparse transition. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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.

Sam Buchanan, Druv Pai, Peng Wang, and Yi Ma. *Learning Deep Representations of Data Distributions*. Online, August 2025. https://ma-lab-berkeley.github.io/deep-representation-learning-book/.

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, 2023a.

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.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *International Conference on Learning Representations*, 2024.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. *International Conference on Learning Representations*, 2025.

Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *International Conference on Learning Representations*, 2025.

Sidney J Yakowitz and John D Spragins. On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 39(1):209–214, 1968.

Aapo Hyvärinen, Ilyes Khemakhem, and Ricardo Monti. Identifiability of latent-variable and structural-equation models: from linear to nonlinear. *Annals of the Institute of Statistical Mathematics*, 76(1):1–33, 2024.

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

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *International Conference on Learning Representations*, 2019.

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

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.

Matthew Willetts and Brooks Paige. I don't need u: Identifiable non-linear ica without side information. *arXiv preprint arXiv:2106.05238*, 2021.

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.

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

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Causal Learning and Reasoning*, pages 843–867. PMLR, 2024.

S. Lachapelle, D. Mahajan, I. Mitliagkas, and S. Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, 2023b.

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.

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

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.

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

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. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zijian Li, Yifan Shen, Kaitao Zheng, Ruichu Cai, Xiangchen Song, Mingming Gong, Guangyi Chen, and Kun Zhang. On the identification of temporal causal representation with instantaneous dependence. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

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.

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.

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.

Richard Caron and Tim Traynor. The zero set of a polynomial. *WSMR Report*, pages 05–02, 2005.

Joachim Bona-Pellissier, François Bachoc, and François Malgouyres. Parameter identifiability of a deep feedforward relu neural network. *Machine Learning*, 112 (11):4431–4493, 2023.

J. R. Munkres. *Topology*. Prentice Hall, Inc., 2 edition, 2000.

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.

BS Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107(3):529–530, 2020.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes. In Section 2, we provide a clean mathematical setting; in Section 3, we clearly list all the assumptions; in Section 4, we provide a full loss function used in our algorithm.]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes. In Appendix D.4, we report both the number of epochs needed for convergence and the average execution time per 1000 epochs in seconds.]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes. We provide all the code for our method at GitHub repository]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes. In Section 3, we clearly list all the assumptions.]
   (b) Complete proofs of all theoretical results. [Yes. We provide all proofs in Appendix B.]
   (c) Clear explanations of any assumptions. [Yes. For each assumption, we provide an intuitive explanation. We also provide a more detailed discussion of all assumptions in Appendix E.]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes. We provide all the code for our method at GitHub repository]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes. We describe the important part of the training and testing process in Section 5. In addition, we provide all the hyperparameters and other experimental details that we used in the training process in the Appendix D.]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes. All the configurations of the experiments are run across multiple seeds, and we report the mean and standard deviation for all of them in both Section 5 and Appendix D.]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes. We provide this information in Appendix D.12.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator if your work uses existing assets. [Yes. We properly cite all the creators whose code we use in the main paper and in App. C.]
   (b) The license information of the assets, if applicable. [Yes, we list all the license information in App. C.]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes.]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Identifiability of potentially degenerate Gaussian Mixture Models with piecewise affine mixing: Supplementary Materials

A Notation

Throughout this paper, we adopt the following conventions for mathematical notation: We use lowercase italic letters such as x to denote scalars, bold lowercase letters such as $\mathbf{x}$ to denote deterministic vectors, bold uppercase letters such as $\mathbf{X}$ to denote random vectors, and uppercase italic letters such as X to denote matrices, and calligraphic uppercase letters such as $\mathcal{X}$ to denote sets (e.g., a set of inputs or a domain). We list all specific notations in the following table:

Symbol	Description
$\mathbf{Z}$	Causal latent variables
$\mathbf{X}$	Observations
$\mathbf{f}$	Mixing function
n	Latent space dimension
d	Observation space dimension
J	Number of components in mixture models
$\mathcal{Z}$	Support of pdGMM latent
$\mathcal{Z}_j$	Support of j-th component of pdGMM latent
$\mathcal{X}$	Support of observation
$\mathcal{K}_j$	Basis index for j-th component
$\mathbf{z}^0$	Translation vector

B Proofs

Our work focuses on Gaussian Mixture Models (GMMs), which are frequently used in machine learning models and possess a number of nice theoretical properties like closure under affine transformation etc. Before presenting our main results, we recall one of the fundamental properties of GMMs, stated in Theorem B.2, which will serve as a recurring tool in our derivations. This theorem builds a one-to-one correspondence-commonly referred to as *identifiability of distributions* between the probability measure and the parameterization of a GMM. Notably, this property holds even when the mixture includes potentially degenerate components (i.e., components with singular covariance matrices), which we refer to as *potentially degenerate GMMs* (pdGMMs).

To proceed, we begin by reporting again the formal definition of pdGMM in a reduced form, i.e., with distinct components. Note that $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ stands for the Gaussian probability measure associated to parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$.

Definition B.1. A variable $\mathbf{Z}$ follows a potentially degenerate Gaussian mixture model (pdGMM) if its distribution is $P = \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where for each component $j \in [J]$ the mean is $\boldsymbol{\mu}_j \in \mathbb{R}^n$ and the covariance matrix is $\boldsymbol{\Sigma}_j \in \mathbb{R}^{n \times n}$, such that determinant $|\boldsymbol{\Sigma}_j| \geq 0$. We assume that the weights $\lambda_j > 0$ and $\sum_{j=1}^{J} \lambda_j = 1$. If each component is distinct, i.e., for every $i \neq j \in [J]$ we have $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \neq (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, then we say that the pdGMM is in a *reduced form*.

We now report a classic result on the identifiability of Gaussian Mixture Models (GMMs) on the whole domain by Yakowitz and Spragins (1968). While this result holds also for pdGMMs, it only proves identifiability when two GMMs are equal in distribution on the whole domain, while in this paper we will focus on identifiability from an open subset of the domain.

Theorem B.2 (Identifiable Gaussian Mixture Models - Proposition 2 in Yakowitz and Spragins (1968))**.** *Consider*

a pair of finite GMMs (in reduced form) in $\mathbb{R}^n$*,*

$$P = \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad P' = \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)\,. \tag{7}$$

we have that $P = P'$ *if and only if the number of mixture components are the same, i.e.,* $J = J'$*, and for some permutation* π *for all* $j \in [J]$ *the mixture components have the same parameters, i.e.,* $\lambda_j = \lambda_{\pi(j)}$ *and* $(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\boldsymbol{\mu}'_{\pi(j)}, \boldsymbol{\Sigma}'_{\pi(j)})$*.*

B.1 Identifiability of pdGMMs from an open subset

In this section, we address the following question: *Can a potentially degenerate Gaussian Mixture Model (pdGMM) be identified using only information from an open subset of its domain?*

Inspired by Kivva et al. (2022), which show that a non-degenerate GMM (so a GMM with $|\boldsymbol{\Sigma}_j| > 0$ for all j) can be uniquely identified from its restriction to a single open subset of the domain (Theorem B.4), we show that a similar identifiability result also holds for pdGMMs (Theorem 3.2).

However, the extension from GMMs to pdGMMs is nontrivial. The main challenge arises from the fact that the proof technique in Kivva et al. (2022) critically relies on the analyticity of the probability density function (w.r.t. the Lebesgue measure) of a non-degenerate Gaussian mixture. However, mixtures of degenerate Gaussians do not have a density function, which means this strategy cannot be applied. This necessitates the development of a new proof strategy to overcome the lack of density function.

As a first step, we introduce a formal definition of equivalence in distribution over a constrained subdomain.

Definition B.3. Let $\mu : \mathcal{F} \to [0,1]$ and $\mu' : \mathcal{F} \to [0,1]$ be the probability measures induced by $\mathbf{X}$ and $\mathbf{X}'$, respectively, where $\mathcal{F}$ is the Borel sigma-algebra over $\mathbb{R}^n$. Given a subset $E \in \mathcal{F}$, we define the restriction of μ to E, denoted by $\mu_E : \mathcal{F} \to [0,1]$, as $\mu_E(A) := \mu(A \cap E), \forall A \in \mathcal{F}$. We say that $\mathbf{X}$ and $\mathbf{X}'$ are equal in distribution over a subdomain E, denoted as $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, when $\mu_E = \mu'_E$.

We then recall the results from Kivva et al. (2022), which show that for non-degenerate GMMs, we can identify them via an open subdomain.

Theorem B.4 (Identifiability of non-degenerate GMMs from open set (Kivva et al., 2022))**.** *Consider a pair of finite non-degenerate GMMs in* $\mathbb{R}^n$ *in reduced form,*

$$\mathbf{X} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad \mathbf{X}' \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{8}$$

where $|\det(\boldsymbol{\Sigma}_j)| > 0, \forall j \in [J]$ *and* $|\det(\boldsymbol{\Sigma}'_j)| > 0, \forall j \in [J']$*. If* $\mathbf{X} \overset{d}{=} \mathbf{X}'$ *on* E*, where* $E \subseteq \mathbb{R}^n$ *is an open set, then,* $\mathbf{X} \overset{d}{=} \mathbf{X}'$*.*

Next, we focus on extending these results to pdGMMs, which we will formalize in the following theorem.

Theorem 3.2 (Identifiability of pdGMMs from open set)**.** *Consider two pdGMMs in* $\mathbb{R}^n$ *in reduced form,*

$$\mathbf{X} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad \mathbf{X}' \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{1}$$

where $rank(\boldsymbol{\Sigma}_j) > 0, \forall j \in [J]$ *and* $rank(\boldsymbol{\Sigma}'_j) > 0, \forall j \in [J']$*. Let* $\mathcal{X}_j$ *be the supports of* $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$*. If there exists an open set* E *such that for all* $j \in [J]$*,* $\mathcal{X}_j \cap E \neq \varnothing$ *and* $\mathbf{X} \overset{d}{=} \mathbf{X}'$ *on* E*, then* $\mathbf{X} \overset{d}{=} \mathbf{X}'$ *on the whole domain,* $J = J'$*, and there exists a permutation* $\pi : [J] \to [J']$ *such that* $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi(j)}, \boldsymbol{\mu}'_{\pi(j)}, \boldsymbol{\Sigma}'_{\pi(j)})$*.*

As previously mentioned, proving the above theorem is challenging because, in the degenerate case, the probability density function no longer exists, making the techniques used in Kivva et al. (2022) inapplicable. To overcome this difficulty, we first provide a weaker result, Theorem B.5, which is requiring all supports intersect at one point at least, as a middle step to get Theorem 3.2.

Theorem B.5 (Identifiability of pdGMMs from open set with common intersection)**.** *Under the same conditions as Theorem 3.2, and further assuming that there exists a point* $\mathbf{z}_0$*, such that* $\mathbf{z}_0 \in \cap_{j\in[J]}\mathcal{X}_j$*, then, the same results stated in Theorem 3.2 hold.*

To prove Theorem B.5, we build upon four key technical lemmas that we introduce and prove below. We now provide a brief overview of our proof strategy. The core idea is to project a pdGMM in $\mathbb{R}^n$ into a sequence of lower-dimensional spaces $\mathbb{R}^k$, where $k \in [n]$. Intuitively, the projection serves to "resolve" the degeneracy by embedding each low-rank component into a subspace where it appears non-degenerate, thus enabling the application of classical identifiability arguments in Theorem B.4.

To make this strategy work, the projections must satisfy two key properties. First, to ensure that no critical information is lost during projection (intuitively, the intrinsic structure of the component is retained under projection), the rank of the covariance matrix for the targeted Gaussian component must be preserved. Second, a single projection is insufficient to fully recover the parameters of a pdGMM; instead, we require a sufficiently rich set of rank-preserving projections to capture all distinguishable parameterizations. The existence of such a family of rank-preserving mappings is guaranteed by Lemma B.6.

Lemma B.6. *Suppose a positive semi-definite and symmetric matrix* $\mathbf{\Sigma} \in \mathbb{R}^{n\times n}$ *has rank* $k \leq n$*. Let*

$$\mathcal{B} := \{A \in \mathbb{R}^{n\times m} : rank(A^T\mathbf{\Sigma}A) = k\}, \tag{9}$$

where $k \leq m \leq n$*. Then, we have that i) the Lebesgue measure of its complement* $\mathcal{B}^c$*, denoted by* $\mathcal{L}(\mathcal{B}^c)$ *is zero, and ii)* $\mathcal{B}$ *is open.*

Proof. The fact that $rank(A\mathbf{\Sigma}A^T) = k$ means there exists at least one $k \times k$ submatrix with full rank k. Consider the sets

$$\mathcal{A}_{I,J} := \{A \in \mathbb{R}^{n\times m} : \det((A^T\mathbf{\Sigma}A)_{I,J}) = 0\} \tag{10}$$

which are the sets of matrices such that $A\mathbf{\Sigma}_j A^T$ has a submatrix with index sets $I, J \subseteq [m]$ with $|I| = |J|$ that is full rank. Then the complement of $\mathcal{B}$, denoted as $\mathcal{B}^c$, can be rewritten as

$$\mathcal{B}^c = \cap_{\{I,J\subseteq[m]:|I|=|J|=k\}}\mathcal{A}_{I,J}. \tag{11}$$

This implies $\mathcal{L}(\mathcal{B}^c) \leq \mathcal{L}(\mathcal{A}_{I,J})$, $\forall(I, J) \in \{I, J \subseteq [n] : |I| = |J| = k\}$. Therefore, if we can find one $\mathcal{A}_{I,J}$ which is measure zero, $\mathcal{B}^c$ has Lebesgue measure zero as well.

Note that $\det((A^T\mathbf{\Sigma}A)_{I,J})$ is a polynomial in $A_{i,j}$, where $i \in I, j \in J$ are individual row and column indices. This is because det is a polynomial and $(A^T\mathbf{\Sigma}A)_{I,J}$ is a polynomial and thus composition of polynomials is a polynomial. Additionally, $\mathcal{A}_{I,J}$ is the set of the roots of this polynomial. By Caron and Traynor (2005), a polynomial from $\mathbb{R}^n$ to $\mathbb{R}$ is either zero everywhere or non-zero almost everywhere. In our context, it means if the polynomial $\det((A^T\mathbf{\Sigma}A)_{I,J})$ is not zero everywhere, then it is nonzero almost everywhere, i.e. its roots $\mathcal{A}_{I,J}$ has zero Lebesgue measure in $\mathbb{R}^{n\times m}$. Thus, all we need to show is that the polynomial $\det((A^T\mathbf{\Sigma}A)_{I,J})$ is not always zero. In other words, we must find a $A_0 \in \mathbb{R}^{n\times m}$, such that for one pair I, J, $\det((A^T\mathbf{\Sigma}A)_{I,J}) \neq 0$. Then in the following, we show how to construct such a A_0.

Since $\text{rank}(\mathbf{\Sigma}) = k$, there exists an invertible $k \times k$ submatrix $\mathbf{\Sigma}_{I,J}$. Let $\tilde{\mathbf{I}} := [\mathbf{e}_1 \ \cdots \ \mathbf{e}_m] \in \mathbb{R}^{n\times m}$ where the vectors $\mathbf{e}_k$ are the elements of the standard basis of $\mathbb{R}^n$ (one-hot vector). Notice that

$$(\tilde{\mathbf{I}}^T\mathbf{\Sigma}\tilde{\mathbf{I}})_{I,J} = (\tilde{\mathbf{I}}^T)_{I,\cdot}\mathbf{\Sigma}\tilde{\mathbf{I}}_{\cdot,J} = (\tilde{\mathbf{I}}_{\cdot,I})^T\mathbf{\Sigma}\tilde{\mathbf{I}}_{\cdot,J} = \mathbf{\Sigma}_{I,J}\,. \tag{12}$$

So we found index sets $I, J \subseteq [m]$ with $|I| = |J| = k$ and a matrix $A_0 := \tilde{\mathbf{I}} \in \mathbb{R}^{n\times m}$ such that

$$\det((A_0^T\mathbf{\Sigma}A_0)_{I,J}) = \det((\tilde{\mathbf{I}}^T\mathbf{\Sigma}\tilde{\mathbf{I}})_{I,J}) = \det(\mathbf{\Sigma}_{I,J}) \neq 0\,,$$

as desired. Hence, the polynomial function $p(A) := \det((A^T\mathbf{\Sigma}A)_{I,J})$ is not everywhere zero and thus, by Caron and Traynor (2005), it is nonzero almost everywhere. In other words, $\mathcal{A}_{I,J}$ has zero Lebesgue measure. We can conclude that $\mathcal{B}^c$ has zero Lebesgue measure.

We now show that $\mathcal{B}$ is open. Take the complement set on Equation (11), we have

$$\mathcal{B} = \cup_{\{I,J\subseteq[n]:|I|=|J|=k\}}\mathcal{A}^c_{I,J}, \tag{13}$$

$$\mathcal{A}^c_{I,J} = \{A \in \mathbb{R}^{n\times m} : \det((A^T\mathbf{\Sigma}A)_{[I,J]}) \neq 0\}. \tag{14}$$

Since $\det((A^T\mathbf{\Sigma}A)_{[I,J]})$ is a continuous function and $\mathcal{A}^c_{I,J}$ is the preimage of open set $\mathbb{R}\backslash\{0\}$, $\mathcal{A}^c_{I,J}$ is open as well. As the union of finite open sets, we can conclude $\mathcal{B}$ is open. □

The following lemma will be useful many times in what follows. For a probability measure μ and a function $\mathbf{f}$, $\mathbf{f}(\mu)$ designates the pushforward of μ under $\mathbf{f}$. The pushforward measure is defined as $\mathbf{f}(\mu)(E) := \mu(\mathbf{f}^{-1}(E))$. We recall that as defined in Def. B.3, μ_E is a restriction of μ to E.

Lemma B.7. *Let $\mu : \mathcal{F}^{(n)} \to \mathbb{R}$ be a measure on $\mathcal{F}^{(n)}$, the Borel sigma-algebra of $\mathbb{R}^n$. Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be an injective function for any set $E \in \mathcal{F}^{(n)}$, we have $\mathbf{f}(\mu_E) = \mathbf{f}(\mu)_{\mathbf{f}(E)}$.*

Proof. Take an arbitrary set $B \in \mathcal{F}^{(m)}$, the Borel sigma-algebra over $\mathbb{R}^m$. We have

$$\mathbf{f}(\mu_E)(B) = \mu_E(\mathbf{f}^{-1}(B)) = \mu(\mathbf{f}^{-1}(B) \cap E) \tag{15}$$

Since $\mathbf{f}$ is injective, we have that $E = \mathbf{f}^{-1}(\mathbf{f}(E))$. This means

$$\begin{aligned}
\mathbf{f}(\mu_E)(B) &= \mu(\mathbf{f}^{-1}(B) \cap \mathbf{f}^{-1}(\mathbf{f}(E))) && (16)\\
&= \mu(\mathbf{f}^{-1}(B \cap \mathbf{f}(E))) && (17)\\
&= \mathbf{f}(\mu)(B \cap \mathbf{f}(E)) && (18)\\
&= \mathbf{f}(\mu)_{\mathbf{f}(E)}(B)\,, && (19)
\end{aligned}$$

where the second equality uses the general fact that taking the preimage preserves intersections. □

The next lemma is a slightly different version from the one showed above where the function is required to be injective only on the support of the measure μ.

Lemma B.8. *Let $\mu : \mathcal{F}^{(n)} \to \mathbb{R}$ be a measure on $\mathcal{F}^{(n)}$, the Borel sigma-algebra of $\mathbb{R}^n$. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the support of μ, and let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be such that its restriction to $\mathcal{X}$ is injective, then for any set $E \in \mathcal{F}^{(n)}$, we have $\mathbf{f}(\mu_E) = \mathbf{f}(\mu)_{\mathbf{f}(E\cap\mathcal{X})}$.*

Proof. Take an arbitrary set $B \in \mathcal{F}^{(m)}$, the Borel sigma-algebra over $\mathbb{R}^m$. We have

$$\begin{aligned}
\mathbf{f}(\mu_E)(B) &= \mu_E(\mathbf{f}^{-1}(B)) && (20)\\
&= \mu(\mathbf{f}^{-1}(B) \cap E) && (21)\\
&= \mu(\mathbf{f}^{-1}(B) \cap E \cap \mathcal{X}) + \mu(\mathbf{f}^{-1}(B) \cap E \cap \mathcal{X}^c) && (22)\\
&= \mu(\mathbf{f}^{-1}(B) \cap (E \cap \mathcal{X}))\,, && (23)
\end{aligned}$$

where the last equality used the fact that $\mu(\mathcal{X}^c) = 0$. Since $\mathbf{f}$ is injective on $\mathcal{X}$ and $(E \cap \mathcal{X}) \subseteq \mathcal{X}$, we have that $(E \cap \mathcal{X}) = \mathbf{f}^{-1}(\mathbf{f}(E \cap \mathcal{X})) \cap \mathcal{X}$. This means

$$\begin{aligned}
\mathbf{f}(\mu_E)(B) &= \mu(\mathbf{f}^{-1}(B) \cap \mathbf{f}^{-1}(\mathbf{f}(E \cap \mathcal{X})) \cap \mathcal{X}) && (24)\\
&= \mu(\mathbf{f}^{-1}(B) \cap \mathbf{f}^{-1}(\mathbf{f}(E \cap \mathcal{X})) \cap \mathcal{X}) + \underbrace{\mu(\mathbf{f}^{-1}(B) \cap \mathbf{f}^{-1}(\mathbf{f}(E \cap \mathcal{X})) \cap \mathcal{X}^c)}_{=0} && (25)\\
&= \mu(\mathbf{f}^{-1}(B) \cap \mathbf{f}^{-1}(\mathbf{f}(E \cap \mathcal{X}))) && (26)\\
&= \mu(\mathbf{f}^{-1}(B \cap \mathbf{f}(E \cap \mathcal{X}))) && (27)\\
&= \mathbf{f}(\mu)(B \cap \mathbf{f}(E \cap \mathcal{X})) && (28)\\
&= \mathbf{f}(\mu)_{\mathbf{f}(E\cap\mathcal{X})}(B)\,, && (29)
\end{aligned}$$

where the second equality uses the fact that $\mu(\mathcal{X}^c) = 0$ and fourth equality uses the general fact that taking the preimage preserves intersections. □

The previous two lemmas show that the injectivity can ensure the exchange order of pushforward and restricted measures. In the following lemma, we show that the rank-preserving projections are injective on the domain we are interested in, and hence we can apply these lemmas.

Lemma B.9. *For $A \in \mathbb{R}^{n\times m}$, where $m \le n$, if $rank(\mathbf{\Sigma}) = \text{rank}(A^T\mathbf{\Sigma}A) = k \le m$, then, A^T is injective on $col(\mathbf{\Sigma}) + \boldsymbol{\mu}$, $\forall \boldsymbol{\mu} \in \mathbb{R}^n$.*

Proof. Since $rank(A^T\mathbf{\Sigma}A) \le rank(A^T\mathbf{\Sigma})$, we get $k \le rank(A^T\mathbf{\Sigma}) \le m$. Now we consider the null space for $A^T\mathbf{\Sigma}$ and $\mathbf{\Sigma}$.

First, we know $\dim(null(A^T\mathbf{\Sigma})) \in [n-m, n-k]$ and $\dim(null(\mathbf{\Sigma})) = n-k$. Second, $\forall x \in \mathbb{R}^n$, if $\mathbf{\Sigma}x = 0$, then $A^T\mathbf{\Sigma}x = 0$. This means $null(\mathbf{\Sigma}) \subseteq null(A^T\mathbf{\Sigma})$. Combined with the first point, we can get $\dim(null(A^T\mathbf{\Sigma})) = n-k$. Since both of them are affine subspaces with the same dimension and one is the subset of the other, we get $null(\mathbf{\Sigma}) = null(A^T\mathbf{\Sigma})$. This means

$$\forall y \in \mathbb{R}^n, \text{ if } A^T\mathbf{\Sigma}y = 0 \quad \Rightarrow \quad \mathbf{\Sigma}y = 0 \tag{30}$$

Denote $x := \mathbf{\Sigma}y \in col(\mathbf{\Sigma})$, then we have

$$\forall x \in col(\mathbf{\Sigma}), \text{ if } A^T x = 0 \quad \Rightarrow \quad x = 0. \tag{31}$$

Now consider two distinct points $x_1, x_2 \in col(\mathbf{\Sigma}) + \boldsymbol{\mu}$. As $x_1 - x_2 \in col(\mathbf{\Sigma})$ but not equal to 0, by Equation (31), we have

$$A^T(x_1 - x_2) \neq 0 \quad \Rightarrow \quad A^T x_1 \neq A^T x_2, \tag{32}$$

which means A^T is injective on $col(\mathbf{\Sigma}) + \boldsymbol{\mu}$. □

In the following lemma, we show that only the components with well-defined density can be equal to each other, and the same for the components without density. Intuitively, we can group the Gaussian components by rank and analyze each group separately. To discuss the existence of well-defined density, we recall the Radon-Nikodym theorem: informally, if measure μ is absolutely continuous with respect to measure $\mathcal{L}$, denoted as $\mu \ll \mathcal{L}$, then there exists a density function of μ w.r.t. $\mathcal{L}$. In other words, $\mu \ll \mathcal{L}$ means for any set A, if $\mathcal{L}(A) = 0$, then $\mu(A) = 0$.

Lemma B.10. *Let μ, $\tilde{\mu}$, λ, $\tilde{\lambda}$ be nonnegative finite measures on a common measure space such that*

$$\mu + \lambda = \tilde{\mu} + \tilde{\lambda}. \tag{33}$$

Let Λ and $\tilde{\Lambda}$ be the supports of λ and $\tilde{\lambda}$, respectively. If there is a third measure $\mathcal{L}$ over the same space such that (i) $\mu, \tilde{\mu} \ll \mathcal{L}$, and (ii) $\mathcal{L}(\Lambda) = \mathcal{L}(\tilde{\Lambda}) = 0$, then $\mu = \tilde{\mu}$ and $\lambda = \tilde{\lambda}$.

Proof. First, we rewrite the equation equivalently as:

$$\mu - \tilde{\mu} = \tilde{\lambda} - \lambda. \tag{34}$$

Let B be an arbitrary measurable set. We have

$$\begin{aligned}(\mu - \tilde{\mu})(B) &= (\tilde{\lambda} - \lambda)(B) &(35)\\ &= (\tilde{\lambda} - \lambda)(B \cap (\Lambda \cup \tilde{\Lambda})) + (\tilde{\lambda} - \lambda)(B \cap (\Lambda \cup \tilde{\Lambda})^c) &(36)\end{aligned}$$

Since $B \cap (\Lambda \cup \tilde{\Lambda})^c$ has no overlap with the support of neither λ nor $\tilde{\lambda}$, we have $(\tilde{\lambda} - \lambda)(B \cap (\Lambda \cup \tilde{\Lambda})^c) = 0$. This means

$$\begin{aligned}(\mu - \tilde{\mu})(B) &= (\tilde{\lambda} - \lambda)(B \cap (\Lambda \cup \tilde{\Lambda})) &(37)\\ &= (\mu - \tilde{\mu})(B \cap (\Lambda \cup \tilde{\Lambda})) &(38)\\ &= \mu(B \cap (\Lambda \cup \tilde{\Lambda})) - \tilde{\mu}(B \cap (\Lambda \cup \tilde{\Lambda}))\,, &(39)\end{aligned}$$

where the second equality holds by Equation (34). Note that

$$\mathcal{L}(B \cap (\Lambda \cup \tilde{\Lambda})) \leq \mathcal{L}(\Lambda \cup \tilde{\Lambda})) \leq \mathcal{L}(\Lambda) + \mathcal{L}(\tilde{\Lambda}) = 0\,,$$

which means $\mathcal{L}(B \cap (\Lambda \cup \tilde{\Lambda})) = 0$. Since $\mu, \tilde{\mu} \ll \mathcal{L}$, we must have $\mu(B \cap (\Lambda \cup \tilde{\Lambda})) = \tilde{\mu}(B \cap (\Lambda \cup \tilde{\Lambda})) = 0$. This combined with Equation (36) implies $(\mu - \tilde{\mu})(B) = 0$, i.e. $\mu(B) = \tilde{\mu}(B)$. This is true for any B, thus $\mu = \tilde{\mu}$. Plugging this last fact into Equation (34) yields, $\tilde{\lambda} = \lambda$. □

Now we are ready to prove our intermediate result, Theorem B.5, that proves the identifiability of pdGMMs from an open set with a common intersection.

Theorem B.5 (Identifiability of pdGMMs from open set with common intersection)**.** *Under the same conditions as Theorem 3.2, and further assuming that there exists a point* $\mathbf{z}_0$*, such that* $\mathbf{z}_0 \in \cap_{j\in[J]}\mathcal{X}_j$*, then, the same results stated in Theorem 3.2 hold.*

Proof. Let $k_j := rank(\mathbf{\Sigma}_j)$ and $k'_j := rank(\mathbf{\Sigma}'_j)$. W.l.o.g., suppose that the components in each of the two pdGMM are ordered by rank in a decreasing fashion, i.e., $k_1 \geq k_2 \geq ... \geq k_J$, $k'_1 \geq k'_2 \geq ... \geq k'_{J'}$ and that the rank of the first component of the first pdGMM $\mathbf{X}$ is larger or equal to the rank of the first component of the second pdGMM $\mathbf{X}'$, i.e., $k_1 \geq k'_1$.

We define

$$\mathcal{S}_k := \{j \in [J] : k_j \leq k\} \tag{40}$$

as the index set of all components in the first pdGMM $\mathbf{X}$, which have rank lower than or equal to k. Similarly we define the same index set for $\mathbf{X}'$ as

$$\mathcal{S}'_k := \{j \in [J'] : k'_j \leq k\}. \tag{41}$$

We define the set of linear mappings that preserve the rank for the component j of $\mathbf{X}'$ as

$$\mathcal{B}_{k,j} := \{A \in \mathbb{R}^{n\times k} \mid rank(A^T\mathbf{\Sigma}_j A) = k_j\}. \tag{42}$$

We now define the intersection of all the rank-preserving matrices for the components that have rank lower than or equal to k as

$$\mathcal{B}_k := \cap_{\{j\in\mathcal{S}_k\}}\mathcal{B}_{k,j}. \tag{43}$$

Similarly for $\mathbf{X}'$, we define

$$\mathcal{B}'_k := \cap_{\{j\in\mathcal{S}'_k\}}\mathcal{B}'_{k,j} \text{ with } \mathcal{B}'_{k,j} := \{A \in \mathbb{R}^{n\times k} \mid rank(A^T\mathbf{\Sigma}'_j A) = k'_j\}. \tag{44}$$

$\mathcal{B}_k$ and $\mathcal{B}'_k$ involve the linear mappings that preserve the ranks for the components with rank no larger than k. Furthermore, to ensure the components in $\mathbf{X}$ are still different from each other after mapping, we first consider a pair of components in $\mathbf{X}$ and define

$$\mathcal{C}_{k,i,j} := \{A \in \mathbb{R}^{n\times k} \mid A^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \neq \mathbf{0} \text{ or } A^T(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)A \neq \mathbf{0}\}, \tag{45}$$

and then go through all possible pairs and define

$$\mathcal{C}_k := \cap_{\{i\neq j\in\mathcal{S}_k : k_i, k_j = k\}}\mathcal{C}_{k,i,j}. \tag{46}$$

Similarly for $\mathbf{X}'$, we define

$$\mathcal{C}'_k := \cap_{\{i\neq j\in\mathcal{S}'_k : k'_i, k'_j = k\}}\mathcal{C}'_{k,i,j} \tag{47}$$

$$\text{with } \mathcal{C}'_{k,i,j} := \{A \in \mathbb{R}^{n\times k} \mid A^T(\boldsymbol{\mu}'_i - \boldsymbol{\mu}'_j) \neq \mathbf{0} \text{ or } A^T(\mathbf{\Sigma}'_i - \mathbf{\Sigma}'_j)A \neq \mathbf{0}\}. \tag{48}$$

We now analyze the set $\mathcal{A}_k \subseteq \mathbb{R}^{n\times k}$ where $k > 0$. Define

$$\mathcal{A}_k := \mathcal{B}_k \cap \mathcal{B}'_k \cap \mathcal{C}_k \cap \mathcal{C}'_k. \tag{49}$$

We will show that i) $\mathbb{R}^{n\times k}\backslash\mathcal{A}_k$ has Lebesgue measure zero, and ii) $\mathcal{A}$ is an open set in $\mathbb{R}^{n\times k}$. We analyze each of the components

Let $\mathcal{L}$ be the Lebesgue measure. For $\mathcal{B}_k$, by Lemma B.6, we know $\mathcal{L}(\mathcal{B}_{k,j}^c) = 0$ for all $j \in [J]$. Thus, we have

$$0 \leq \mathcal{L}(\mathcal{B}_k^c) = \mathcal{L}(\cup_{j\in\mathcal{S}_k}\mathcal{B}_{k,j}^c) \leq \sum_{j\in\mathcal{S}_k}\mathcal{L}(\mathcal{B}_{k,j}^c) = 0 \tag{50}$$

which of course implies $\mathcal{L}(\mathcal{B}_k^c) = 0$. Next, from Lemma B.6, we also know $\mathcal{B}_{k,j}$ is open for all $j \in [J]$. As $\mathcal{B}_k$ is the finite intersection of $\mathcal{B}_{k,j}$'s, we know $\mathcal{B}_k$ is open as well. We can apply the same arguments and obtain i) $\mathcal{L}(\mathcal{B}_k'^c) = 0$, and ii) $\mathcal{B}_k'$ is an open set.

For $\mathcal{C}_{k,i,j}$, we separate it into two cases:

$$\mathcal{C}_{k,i,j} = \mathcal{C}_{k,i,j}^{\mu} \cup \mathcal{C}_{k,i.j}^{\sigma}, \tag{51}$$

$$\text{where } \mathcal{C}_{k,i,j}^{\mu} := \{A \in \mathbb{R}^{n\times k} \mid A^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \neq \mathbf{0}\}, \tag{52}$$

$$\text{and } \mathcal{C}_{k,i,j}^{\sigma} := \{A \in \mathbb{R}^{n\times k} \mid A^T(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)A \neq \mathbf{0}\}. \tag{53}$$

If $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$, $A^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ can be regarded as m polynomial functions of A from $\mathbb{R}^n$ to $\mathbb{R}$, i.e.

$$A^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \begin{bmatrix} A_{\cdot 1}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ A_{\cdot 2}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \\ \vdots \\ A_{\cdot k}^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \end{bmatrix}. \tag{54}$$

Each $A_{\cdot i}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$, $i \in [k]$ is not zero everywhere as $(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \neq \mathbf{0}$. By Caron and Traynor (2005), a polynomial from $\mathbb{R}^n$ to $\mathbb{R}$ is either zero everywhere or non-zero almost everywhere. Apply to our functions, it means they are all nonzero almost everywhere. Denote the zero set as $\mathcal{O}_\mu$ and we know $\mathcal{L}(\mathcal{O}_\mu) = 0$. Since these polynomial functions all share the same coefficients, we have

$$(\mathcal{C}_{k,i,j}^{\mu})^c = \underbrace{\mathcal{O}_\mu \times \cdots \times \mathcal{O}_\mu}_{m} = \mathcal{O}_\mu^m, \tag{55}$$

which means

$$\mathcal{L}((\mathcal{C}_{k,i,j}^{\mu})^c) = \mathcal{L}(\mathcal{O}_\mu^m) = \mathcal{L}(\mathcal{O}_\mu)^m = 0. \tag{56}$$

If $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$, then consider the diagonal elements of $A^T(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)A$ as n polynomial functions of A, denote as $p_{ii} := A_{\cdot i}^T(\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j)A_{\cdot i}$, $\forall i \in [m]$. Since $\boldsymbol{\Sigma}_i - \boldsymbol{\Sigma}_j \neq \mathbf{0}$, p_{ii} is not zero everywhere. By Caron and Traynor (2005), they are all nonzero almost everywhere. Denote the zero set as $\mathcal{O}_\sigma$ and we know $\mathcal{L}(\mathcal{O}_\sigma) = 0$. Since they are only the diagonal elements, we have

$$(\mathcal{C}_{k,i,j}^{\sigma})^c \subseteq \mathcal{O}_\sigma^m \quad \Rightarrow \quad 0 \leq \mathcal{L}((\mathcal{C}_{k,i,j}^{\sigma})^c) \leq \mathcal{L}(\mathcal{O}_\sigma^m) = \mathcal{L}(\mathcal{O}_\sigma)^m = 0, \tag{57}$$

which means $\mathcal{L}((\mathcal{C}_{k,i,j}^{\sigma})^c) = 0$.

By Equation (51), we have

$$\mathcal{C}_{k,i,j}^c = (\mathcal{C}_{k,i,j}^{\mu})^c \cap (\mathcal{C}_{k,i,j}^{\sigma})^c. \tag{58}$$

As we assume, $\mathbf{X}$ is in reduced form, which means $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ or $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$. This implies at least one of $(\mathcal{C}_{ij}^{\mu})^c$ and $(\mathcal{C}_{ij}^{\sigma})^c$ is Lebesgue measure zero. Then, as the intersection of these two sets, we have $\mathcal{L}(\mathcal{C}_{ij}^c) = 0$. Since the union of measure zero sets is measure zero, we have $\mathcal{L}(\mathcal{C}^c) = 0$.

Since $A^T(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ is continuous function and $(\mathcal{C}_{k,i,j}^{\mu})$ is the preimage of open set $\mathbb{R}^k\backslash\{\mathbf{0}\}$, then $(\mathcal{C}_{k,i,j}^{\mu})$ is open. Similarly, we have $(\mathcal{C}_{k,i,j}^{\sigma})$ is open as well. As the union of two open sets, $\mathcal{C}_{k,i,j}$ is also open. Then as the finite intersection of $\mathcal{C}_{k,i,j}$, we obtain $\mathcal{C})k$ is an open set.

For $\mathcal{C}_k'$, we can apply the same arguments and obtain i) $\mathcal{L}(\mathcal{C}_k'^c) = 0$, and ii) $\mathcal{C}_k'$ is an open set.

To sum up, all the above results mean that $\mathcal{L}(\mathcal{A}_k^c) = \mathcal{L}(\mathcal{B}_k^c \cup \mathcal{B}_k'^c \cup \mathcal{C}_k^c \cup \mathcal{C}_k'^c) = 0$ and thus $\mathcal{A}_k$ is not empty and, furthermore, $\mathcal{A}_k = \mathcal{B}_k \cap \mathcal{B}_k' \cap \mathcal{C}_k \cap \mathcal{C}_k'$ is an open set in $\mathbb{R}^{n\times k}$.

Now, we have $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, by Def. B.3, which means

$$\left(\sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\right)_E = \left(\sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')\right)_E \tag{59}$$

$$\Rightarrow \quad \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_E = \sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_E. \tag{60}$$

We start from the highest rank k_1. Since $\mathcal{A}_{k_1}$ is non-empty, it has an element $A_1 \in \mathcal{A}_{k_1}$. Apply A_1 on both sides of Equation (60) and we get

$$A_1^T \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_E = A_1^T \sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_E \tag{61}$$

$$\sum_{j=1}^{J} \lambda_j A_1^T (N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_E) = \sum_{j=1}^{J'} \lambda_j' A_1^T (N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_E) \tag{62}$$

$$\Rightarrow \quad \sum_{j=1}^{J} \lambda_j A_1^T (N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_{\mathcal{X}_j \cap E}) = \sum_{j=1}^{J'} \lambda_j' A_1^T (N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_{\mathcal{X}_j' \cap E}). \tag{63}$$

Due to $\mathcal{B}_k \cap \mathcal{B}_k'$, we have $rank(A_1^T \boldsymbol{\Sigma}_j A_1) = k_j$ for all $j \in [J]$ and $rank(A_1^T \boldsymbol{\Sigma}_j' A_1) = k_j'$ for all $j \in [J']$. Thus, we can apply Lemma B.9, and obtain A_1 is injective on $col(\boldsymbol{\Sigma}_j) + \boldsymbol{\mu}_j$, which is exactly $\mathcal{X}_j$.

Then, we can apply Lemma B.8 on Equation (63) and get

$$\sum_{j=1}^{J} \lambda_j A_1^T (N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_{\mathcal{X}_j \cap E}) = \sum_{j=1}^{J'} \lambda_j' A_1^T (N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_{\mathcal{X}_j' \cap E}) \tag{64}$$

$$\sum_{j=1}^{J} \lambda_j (A_1^T N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j))_{A_1^T(\mathcal{X}_j \cap E)} = \sum_{j=1}^{J'} \lambda_j' (A_1^T N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j'))_{A_1^T(\mathcal{X}_j' \cap E)} \tag{65}$$

$$\sum_{j=1}^{J} \lambda_j N(A_1^T \boldsymbol{\mu}_j, A_1^T \boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j \cap E)} = \sum_{j=1}^{J'} \lambda_j' N(A_1^T \boldsymbol{\mu}_j', A_1^T \boldsymbol{\Sigma}_j' A_1)_{A_1^T(\mathcal{X}_j' \cap E)}. \tag{66}$$

Next, we split both sides into two terms by the rank value k_j and k_j'

$$\sum_{\{j\in[J]:k_j=k_1\}} \lambda_j N(A_1^T \boldsymbol{\mu}_j, A_1^T \boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j \cap E)} + \sum_{\{j\in[J]:k_j<k_1\}} \lambda_j N(A_1^T \boldsymbol{\mu}_j, A_1^T \boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j \cap E)} \tag{67}$$

$$= \sum_{\{j\in[J']:k_j'=k_1\}} \lambda_j' N(A_1^T \boldsymbol{\mu}_j', A_1^T \boldsymbol{\Sigma}_j' A_1)_{A_1^T(\mathcal{X}_j' \cap E)} + \sum_{\{j\in[J']:k_j'<k_1\}} \lambda_j' N(A_1^T \boldsymbol{\mu}_j', A_1^T \boldsymbol{\Sigma}_j' A_1)_{A_1^T(\mathcal{X}_j' \cap E)}. \tag{68}$$

For the terms in the first sum on both sides, we have $\det(A^T \boldsymbol{\Sigma}_j A) \neq 0$, which implies non-singular covariance matrices after transformation. This means $N(A_1^T \boldsymbol{\mu}_j, A_1^T \boldsymbol{\Sigma}_j A_1) \ll \mathcal{L}$ and $N(A_1^T \boldsymbol{\mu}_j', A_1^T \boldsymbol{\Sigma}_j' A_1) \ll \mathcal{L}$, where $\mathcal{L}$ is the Lebesgue measure[2]. In general, if $\lambda \ll \mu$, we also have $\lambda_E \ll \mu$ and thus we have

$$N(A_1^T \boldsymbol{\mu}_j, A_1^T \boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j \cap E)} \ll \mathcal{L} \quad \text{and} \quad N(A_1^T \boldsymbol{\mu}_j', A_1^T \boldsymbol{\Sigma}_j' A_1)_{A_1^T(\mathcal{X}_j' \cap E)} \ll \mathcal{L} \tag{69}$$

[2]This is another of saying the distribution has a density w.r.t. $\mathcal{L}$, by the Radon-Nikodym theorem

For the terms in the second sum of both sides, we have

$$rank(A_1^T\boldsymbol{\Sigma}_j A_1) = k_j < k_1, \quad rank(A_1^T\boldsymbol{\Sigma}'_j A_1) = k'_j < k_1, \tag{70}$$

which means after transformation, these components are still degenerate. Importantly, it means the support of $N(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1)$ and $N(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1)$ have zero Lebesgue measure. By restricting further, the support of $N(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j\cap E)}$ and $N(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1)_{A_1^T(\mathcal{X}'_j\cap E)}$ still have measure zero.

Therefore, we can apply Lemma B.10 and obtain the equality for the terms with density

$$\sum_{\{j\in[J]:k_j=k_1\}} \lambda_j N(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1)_{A_1^T(\mathcal{X}_j\cap E)} = \sum_{\{j\in[J']:k'_j=k_1\}} \lambda'_j N(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1)_{A_1^T(\mathcal{X}'_j\cap E)}. \tag{71}$$

Let $U := \cap_{\{j\in[J]:k_j=k_1\}} A_1^T(\mathcal{X}_j \cap E)$. Since $A_1^T(\mathcal{X}_j \cap E)$ is an open set in $\mathbb{R}^{k_1}$ for all $\{j \in [J] : k_j = k_1\}$, and have common element $A_1\mathbf{z}_0$ by condition. Thus U is an open and non-empty set. We further restrict the measure of Equation (71) to U:

$$\sum_{\{j\in[J]:k_j=k_1\}} \lambda_j N_U(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1) = \sum_{\{j\in[J]:k'_j=k_1\}} \lambda'_j N_U(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1) \tag{72}$$

$$\left(\sum_{\{j\in[J]:k_j=k_1\}} \lambda_j N(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1)\right)_U = \left(\sum_{\{j\in[J]:k'_j=k_1\}} \lambda'_j N(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1)\right)_U. \tag{73}$$

Since each component in Equation (73) has density functions, and they are all different due to $\mathcal{C}\cap\mathcal{C}'$. As we discussed before, U is an open and non-empty set, thus by Theorem B.4, we can get

$$\sum_{\{j\in[J]:k_j=k_1\}} \lambda_j N(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1) = \sum_{\{j\in[J]:k'_j=k_1\}} \lambda'_j N(A_1^T\boldsymbol{\mu}'_j, A_1^T\boldsymbol{\Sigma}'_j A_1), \tag{74}$$

which means $|\{j \in [J] : k_j = k_1\}| = |\{j \in [J'] : k'_j = k_1\}|$, and there exits a permutation $\pi : \{j \in [J] : k_j = k_1\} \to \{j \in [J'] : k'_j = k_1\}$, such that

$$(A_1^T\boldsymbol{\mu}_j, A_1^T\boldsymbol{\Sigma}_j A_1) = (A_1^T\boldsymbol{\mu}'_{\pi(j)}, A_1^T\boldsymbol{\Sigma}'_{\pi(j)} A_1), \quad \lambda_j = \lambda'_{\pi(j)}, \quad \forall A_1 \in \mathcal{A}_{k_1}. \tag{75}$$

Consider two functions:

$$f_1(A_1) := (A_1^T(\boldsymbol{\mu}_j - \boldsymbol{\mu}'_{\pi(j)}))_1 = 0, \quad f_2(A_1) := \left(A_1^T(\boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}'_{\pi(j)})A_1\right)_{1,1} = 0, \quad \forall A_1 \in \mathcal{A}_{k_1}. \tag{76}$$

Since these hold for all $A_1 \in \mathcal{A}_{k_1}$, and $\mathcal{A}_{k_1}$ is open and non-empty. Thus, we can take the first and second derivatives on f_1 and f_2, respectively, w.r.t. the elements in the first column of A_1. We get

$$f'_1(A_1) := \boldsymbol{\mu}_j - \boldsymbol{\mu}'_{\pi(j)} = \mathbf{0}^{n\times 1}, \quad f''_2(A_1) = \boldsymbol{\Sigma}_j - \boldsymbol{\Sigma}'_{\pi(j)} = \mathbf{0}^{n\times n}, \quad \forall A_1 \in \mathcal{A}_{k_1}. \tag{77}$$

Thus we conclude

$$\exists \pi : \{j \in [J] : k_j = k_1\} \to \{j \in [J'] : k'_j = k_1\}, s.t. \boldsymbol{\mu}_j = \boldsymbol{\mu}'_{\pi(j)}, \quad \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}'_{\pi(j)}, \quad \lambda_j = \lambda'_{\pi(j)}. \tag{78}$$

Now go back to Equation (63), we can cancel out the terms with rank k_1 both sides and get

$$\sum_{\{j\in[J]:k_j<k_1\}} \lambda_j N_E(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{\{j\in[J']:k'_j<k_1\}} \lambda'_j N_E(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j) \tag{79}$$

W.l.o.g, suppose $k_2 \geq k'_2$, the same argument can apply to k_2. And continue progressing from the largest to the smallest rank to cover all elements in both $\mathbf{X}$ and $\mathbf{X}'$, and finally, we can get

$$J = J', \text{ and } \exists \pi : [J] \to [J'], s.t. \boldsymbol{\mu}_j = \boldsymbol{\mu}'_{\pi(j)}, \quad \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}'_{\pi(j)}, \quad \lambda_j = \lambda'_{\pi(j)}, \tag{80}$$

and by Theorem B.2, we can conclude $\mathbf{X} \overset{d}{=} \mathbf{X}'$. □

We have concluded the proof of Theorem B.5, which proves the identifiability of two pdGMMs from an open set assuming a common non-empty intersection set for all components. Based on these results, we can prove our final goal of this section: Theorem 3.2, which removes the assumption of the common non-empty intersection set for all components. The basic idea is to group the components by whether they intersect on the same point and apply Theorem B.5 on each group.

Theorem 3.2 (Identifiability of pdGMMs from open set)**.** *Consider two pdGMMs in $\mathbb{R}^n$ in reduced form,*

$$\mathbf{X} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad \mathbf{X}' \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{1}$$

where $rank(\boldsymbol{\Sigma}_j) > 0, \forall j \in [J]$ and $rank(\boldsymbol{\Sigma}'_j) > 0, \forall j \in [J']$. Let $\mathcal{X}_j$ be the supports of $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. If there exists an open set E such that for all $j \in [J]$, $\mathcal{X}_j \cap E \neq \varnothing$ and $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, then $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on the whole domain, $J = J'$, and there exists a permutation $\pi : [J] \to [J']$ such that $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi(j)}, \boldsymbol{\mu}'_{\pi(j)}, \boldsymbol{\Sigma}'_{\pi(j)})$.

Proof. We first focus on $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$. Pick up one point $\mathbf{z}_1 \in \mathcal{X}_1 \cap E$, and define index set

$$u_1 := \{j \in [J] : \mathbf{z}_1 \in \mathcal{X}_j\}, \quad u'_1 := \{j \in [J'] : \mathbf{z}_1 \in \mathcal{X}'_j\}. \tag{81}$$

Then, we can construct a open ball $B(\mathbf{z}_1, \delta) \subset E$, such that $\forall j \in [J] \backslash u_1$, $\mathcal{X}_j \cap B(\mathbf{z}_1, \delta) = \varnothing$. By condition we have $\mathbf{X} \overset{d}{=} \mathbf{X}'$ on E, we restrict it on $B(\mathbf{z}_1, \delta)$ and eliminate the zero measures on both sides, then we get

$$\sum_{j \in u_1} \lambda_j N_B(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{j \in u'_1} \lambda'_j N_B(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j) \tag{82}$$

$$\left(\sum_{j \in u_1} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right)_B = \left(\sum_{j \in u'_1} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j) \right)_B . \tag{83}$$

By Theorem B.5, we have

$$\sum_{j \in u_1} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{j \in u'_1} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{84}$$

also $|u_1| = |u'_1|$, and there exists a permutation $\pi_1 : u_1 \to u'_1$, such that $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi_1(j)}, \boldsymbol{\mu}'_{\pi_1(j)}, \boldsymbol{\Sigma}'_{\pi_1(j)})$.

Then, we move to the rest $t \in [J] \backslash \{1\}$. Pick up one point $\mathbf{z}_t \in \mathcal{X}_t \cap E$, and define index set

$$u_t := \{j \in [J] \backslash (\cup_{j<t} u_j) : \mathbf{z}_t \in \mathcal{X}_j\}, \quad u'_t := \{j \in [J'] \backslash (\cup_{j<t} u'_j) : \mathbf{z}_t \in \mathcal{X}'_j\}. \tag{85}$$

If u_t is not empty, this means we haven't idenfity component t yet. Then as before, we construct a open ball $B(\mathbf{z}_t, \delta) \subset E$, such that $\forall j \in [J] \backslash u_t$, $\mathcal{X}_j \cap B(\mathbf{z}_t, \delta) = \varnothing$. We restrict it on $B(\mathbf{z}_t, \delta)$ and apply Theorem B.5 to get

$$\sum_{j \in u_t} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = \sum_{j \in u'_t} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{86}$$

also $|u_t| = |u'_t|$, and there exists a permutation $\pi_t : u_t \to u'_t$, such that $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi_t(j)}, \boldsymbol{\mu}'_{\pi_t(j)}, \boldsymbol{\Sigma}'_{\pi_t(j)})$.

By going through all $t \in [J]$, we can combine all results to conclude that $\mathbf{X} \overset{d}{=} \mathbf{X}'$. Also $J = J'$, and there exists a permutation $\pi : J \to J'$, such that $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) = (\lambda'_{\pi(j)}, \boldsymbol{\mu}'_{\pi(j)}, \boldsymbol{\Sigma}'_{\pi(j)})$. □

B.2 Proofs for the identifiability of pdGMM latent variables

In this section, we consider the setting where latent variables follow pdGMMs and aim to identify the them from observations. In Appendix B.2.1 we provide the proof of Theorem 3.5, which shows the identifiability up to affine transformation within components. In Appendix B.2.2 we provide the proof of Theorem 3.7, which shows the identifiability up to affine transformation. Finally, in Appendix B.2.3 we provide the proof of Theorem 3.9, which shows the identifiability up to permutation and element-wise affine transformation.

B.2.1 Affine Identifiability within component of pdGMM latent variables mixed by piecewise affine function

In this section, we consider a set of latent variables that follow pdGMM and are transformed into observations via a continuous, invertible, piecewise affine mixing function. Our objective is to recover the latent variables from the observed data, without any auxiliary information, except for knowing that the latent variables are pdGMM.

First, following Bona-Pellissier et al. (2023), we define continuous piecewise affine functions by *closed polyhedra* as follows:

Definition B.11 (Closed polyhedron (Bona-Pellissier et al., 2023))**.** We say $P \subset \mathbb{R}^n$ is a *closed polyhedron* if and only if there exists $q \in \mathbb{N}$, $\mathbf{a}_1, ..., \mathbf{a}_q \in \mathbb{R}^n$ and $b_1, ..., b_q \in \mathbb{R}$ such that

$$\mathbf{x} \in P \quad \Leftrightarrow \quad \begin{cases} \mathbf{a}_1^T \mathbf{x} + b_1 \leq 0 \\ \vdots \\ \mathbf{a}_q^T \mathbf{x} + b_q \leq 0. \end{cases} \tag{87}$$

Definition B.12 (Continuous piecewise affine function (Bona-Pellissier et al., 2023))**.** We say that a function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a *continuous piecewise affine* function if there exists a finite set of closed polyhedra whose union is $\mathbb{R}^n$ and such that $\mathbf{f}$ is affine over each polyhedron.

A simple application of the pasting lemma (see e.g. Munkres (2000)) shows that a continuous piecewise affine function is indeed continuous. For a continuous piecewise affine function $\mathbf{f}$, there are infinitely many finite covers of $\mathbb{R}^n$ by closed polyhedra such that $\mathbf{f}$ is affine over each polyhedron. We will be interested into a specific type of covers called *admissible*.

Definition B.13. [(Bona-Pellissier et al., 2023)] Let $\mathbf{f}$ be a continuous piecewise affine function. Let Π be a finite set of closed polyhedra of $\mathbb{R}^n$. We say that Π is *admissible* with respect to $\mathbf{f}$ if and only if

1. $\bigcup_{P \in \Pi} P = \mathbb{R}^n$;
2. for all $P \in \Pi$, $\mathbf{f}$ is affine on P; and
3. for all $P \in \Pi$, the interior of P, denoted by P°, is not empty.

Bona-Pellissier et al. (2023, Proposition 22) show that such an admissible set of closed polyhedra exists for all continuous piecewise affine functions.

We first present a result by Kivva et al. (2022) that studies the identifiability of non-degenerate GMM latent variables with piecewise affine mixing function, which serves as a primary inspiration for our approach.

Theorem B.14 (Affine Identifiability of GMM latent variables (Kivva et al., 2022))**.** *Assume the observation* $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ *follows the data-generating process in Sec. 2. Suppose* $\mathbf{Z}$ *follows a Gaussian mixture model*

$$\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \tag{88}$$

with $\boldsymbol{\mu}_j \in \mathbb{R}^n$, $\boldsymbol{\Sigma}_j \in \mathbb{R}^{n \times n}$ *and* $|\boldsymbol{\Sigma}_j| > 0$, $\lambda_j > 0$, *and* $\sum_{j=1}^{J} \lambda_j = 1$. *For every* $i \neq j \in [J]$ *we have* $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \neq (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. *Then* $\mathbf{Z}$ *is identifiable from* $\mathbf{X}$ *up to an affine transformation.*

However, the analysis of Theorem B.14 is restricted to non-degenerate Gaussian mixture models, assuming full-rank covariance matrices for all components.

We begin by presenting an intermediate theorem (Thm. B.16), which shows that any continuous invertible piecewise affine function capable of transforming one pdGMM into another pdGMM must be affine on the support of each component in pdGMM. However, this property does not hold universally. We show now a counterexample in Example B.15, where $\mathbf{f}$ is not affine over the support of any Gaussian components but still preserves the same distribution.

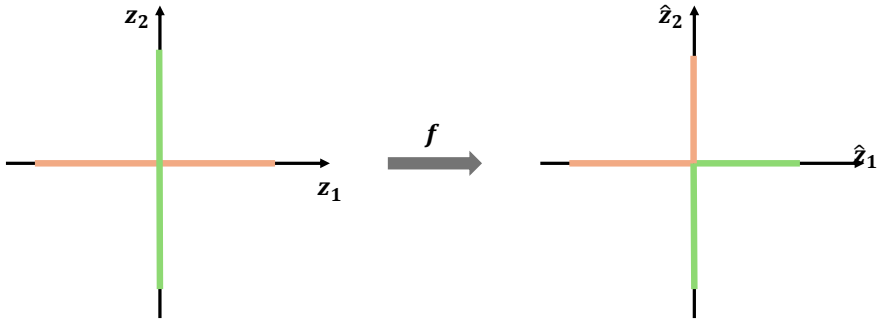

Figure 3: Visualization of Example B.15: $\mathbf{f}$ switches half axes of $\mathbf{z}_1$ and $\mathbf{z}_2$. The orange area refers to the support of component 1, and the green area refers to the support of component 2 before transformation.

Example B.15. Consider a pdGMM with dimension $n = 2$ and Gaussian components $J = 2$,

$$\mathbf{Z} \sim 0.5N((0,0), \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}) + 0.5N((0,0), \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}) \tag{89}$$

Denote the two dimensions as $\mathbf{z}_1$ and $\mathbf{z}_2$. One can construct a function $\mathbf{f} : \mathcal{Z} \to \mathbb{R}^2$ that is clearly not affine on $\mathcal{Z}$:

$$\mathbf{f}(\mathbf{z}) = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \mathbf{z}, \text{ if } \mathbf{z}_1 < 0 \text{ and } \mathbf{z}_2 = 0, \\ \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \mathbf{z}, \text{ if } \mathbf{z}_1 = 0 \text{ and } \mathbf{z}_2 > 0, \\ \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \mathbf{z}, \text{ if } \mathbf{z}_1 = 0 \text{ and } \mathbf{z}_2 < 0 \text{ or } \mathbf{z}_1 > 0 \text{ and } \mathbf{z}_2 = 0. \end{cases} \tag{90}$$

In other words, $\mathbf{f}$ switches the relative position of axes, while retaining the same distribution, i.e. $\mathbf{f}(\mathbf{Z}) \overset{d}{=} \mathbf{Z}$.

To avoid these special cases, we propose Ass. 3.4, which describes a general class of models.

Assumption 3.4 (Genericity of pdGMMs)**.** Consider a pdGMM in reduced form $\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Let $\mathcal{J}_k := \{j \in [J] : rank(\boldsymbol{\Sigma}_j) = k)\}$ be the indices of the components with rank k. For any $k \leq n$ and any subset $\mathcal{J}_k^0 \subseteq \mathcal{J}_k$, suppose that the intersection of the supports of the components $\mathcal{Z}_j$ is non-empty, i.e., $\cap_{j \in \mathcal{J}_k^0} \mathcal{Z}_j \neq \varnothing$. We assume there exists at least one point $\mathbf{z}_0 \in \cap_{j \in \mathcal{J}_k^0} \mathcal{Z}_j$ s.t. the Mahalanobis distance [3] of $\mathbf{z}_0$ to the distributions of any two distinct Gaussian components $i, j \in \mathcal{J}_k^0$ is not the same.

The idea is: the supports of components with the same rank k may overlap, but there is at least one common point in the intersection where the components can be distinguished, i.e., their projections onto the non-degenerate subspace do not have the same norm.

We provide proof and example to illustrate the genericity of Ass. 3.4 in Appendix E.

The following theorem shows that any continuous invertible piecewise affine function capable of transforming one pdGMM into another pdGMM must be affine on the support of each component in pdGMM.

Theorem B.16 (Identifiability of pdGMMs from open set with common intersection)**.** *Consider a pair of finite*

[3] Mahalanobis distance of a point $\mathbf{z}_0$ from a distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is $\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_j)}$. When $\boldsymbol{\Sigma}_j$ is singular, we use the Moore-Penrose inverse for $\boldsymbol{\Sigma}_j^{-1}$

pdGMMs in $\mathbb{R}^n$ in reduced form,

$$\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad \mathbf{Z}' \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{91}$$

where $|\det(\boldsymbol{\Sigma}_j)| \geq 0, \forall j \in [J]$ and $|\det(\boldsymbol{\Sigma}'_j)| \geq 0, \forall j \in [J']$. Suppose $\mathbf{Z}$ satisfies Assumption 3.4. If there exists a continuous, invertible, piecewise affine function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying $\mathbf{Z}' \overset{d}{=} \mathbf{f}(\mathbf{Z})$, then for all $j_0 \in [J]$, $\mathbf{f}$ is affine on $\mathcal{Z}_{j_0}$, where $\mathcal{Z}_{j_0}$ is the support of $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})$.

We introduce some useful theoretical results before we start the proof of Theorem B.16. The first lemma shows that the full-ranked matrices sharing the same Gram matrix can be converted to each other via rotations.

Lemma B.17. *Consider two matrices $A, B \in \mathbb{R}^{n\times k}$ with $n \geq k$ and A is full-column rank with $rank(A) = k$. If $AA^T = BB^T$, then there exists an orthogonal matrix $Q \in \mathbb{R}^{k\times k}$ such that $A = BQ$.*

Proof. Since $AA^T = BB^T$ and A is full-column-rank, we can derive $A = BB^TA(A^TA)^{-1}$. Then, we aim to prove $Q := B^TA(A^TA)^{-1}$ is an orthogonal matrix by showing $Q^TQ = I$, where I is a identity matrix:

$$Q^TQ = [B^TA(A^TA)^{-1}]^TB^TA(A^TA)^{-1} = (A^TA)^{-1}A^TBB^TA(A^TA)^{-1} = I. \tag{92}$$

□

In the following lemma, we show that if we can find an open cover, such that two probability measure are the same over any open set in this open cover, then they are the same measure.

Lemma B.18. *Let $\mu : \mathcal{F} \to [0, 1]$ and $\lambda : \mathcal{F} \to [0, 1]$ be two probability measures on the Borel sigma-algebra $\mathcal{F}$ over $\mathbb{R}^n$. Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the support of μ. Let $\mathcal{O}$ be an open cover of $\mathcal{X}$. If for all $O \in \mathcal{O}$, we have*

$$\mu_{O\cap\mathcal{X}} = \lambda_{O\cap\mathcal{X}}\,,$$

then $\mu = \lambda$.

Proof. Let $B \in \mathcal{F}$. Let $\mathcal{O}_B$ be a cover of B. Since B is a subset of a second-countable space, one can find a countable open subcover $\{O_1, O_2, \dots\}$ of B. We construct a partition B as follow: For all $k \in \mathbb{N}$, define

$$B_1 := B \cap O_1 \quad \text{and} \quad B_k := B \cap O_k \cap \left(\bigcap_{i=1}^{k-1} B_i^c\right).$$

We now show that $\{B_1, B_2, \dots\}$ is indeed a partition of B. For any $x \in B$, we show that there exists a unique B_k such that $x \in B_k$. Let $k_x := \min\{k \in \mathbb{N} \mid x \in O_k\}$. Clearly, $x \notin B_i$ for $i < k_x$ since $x \notin O_i$ for $i < k_x$. Moreover, $x \in B_{k_x}$ since $x \in O_{k_x}$ and $x \in B_i^c$ for all $i < k_x$. It is then clear that $x \notin B_i$ for $i > k_x$ since $x \notin B_{k_x}^c$.

We write

$$\mu(B) = \mu(B \cap \mathcal{X}) = \mu\Big(\Big(\bigcup_{k\geq 1} B_k\Big) \cap \mathcal{X}\Big) = \mu\Big(\bigcup_{k\geq 1}(B_k \cap \mathcal{X})\Big) = \sum_{k=1}^{\infty} \mu(B_k \cap \mathcal{X}) \tag{93}$$

$$= \sum_{k=1}^{\infty} \mu\Big(B \cap O_k \cap \Big(\bigcap_{i=1}^{k-1} B_i^c\Big) \cap \mathcal{X}\Big) = \sum_{k=1}^{\infty} \mu_{O_k\cap\mathcal{X}}\Big(B \cap \Big(\bigcap_{i=1}^{k-1} B_i^c\Big)\Big) \tag{94}$$

$$= \sum_{k=1}^{\infty} \lambda_{O_k\cap\mathcal{X}}\Big(B \cap \Big(\bigcap_{i=1}^{k-1} B_i^c\Big)\Big) = \cdots = \lambda(B \cap \mathcal{X})\,, \tag{95}$$

$$\tag{96}$$

where the $\cdots$ means we apply the exact same steps as before in reverse order. By taking $B = \mathcal{X}$, we get $1 = \mu(\mathcal{X}) = \lambda(\mathcal{X} \cap \mathcal{X}) = \lambda(\mathcal{X})$, which means $\lambda(\mathcal{X}^c) = 0$. This means

$$\mu(B) = \lambda(B \cap \mathcal{X}) = \lambda(B \cap \mathcal{X}) + \lambda(B \cap \mathcal{X}^c) = \lambda(B)\,,$$

which is what we intended to show. □

Next, we report two results from Xu et al. (2024), which show the identifiability results of potentially degenerate multivariate normal distributions (pdMVNs), as we will use them as important final steps in the proof of Theorem B.16 later.

Theorem B.19 (Affine Identifiability for pdMVNs with Piecewise Affine $\mathbf{f}$ (Xu et al. (2024)))**.** *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 $\mathbf{f}(\mathbf{Z}) \overset{d}{=} \hat{\mathbf{f}}(\hat{\mathbf{Z}})$, then there exists an invertible affine transformation $\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{h}(\mathbf{Z}) \overset{d}{=} \hat{\mathbf{Z}}$.*

Lemma B.20 (Linearity of $\mathbf{f}$ for pdMVNs (Xu et al. (2024)))**.** *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 $\mathbf{f}(\mathbf{Z}) \overset{d}{=} \mathbf{Z}$. Then $\mathbf{f}$ is affine over $\mathcal{Z}$, the support of $\mathbf{Z}$.*

Now we are ready to prove Theorem B.16. Below, we outline the structure of the proof. We begin by focusing on a fixed Gaussian component j_0 from $\mathbf{Z}$ and partitioning its support into polyhedral regions such that the function $\mathbf{f}$ is affine within each subdomain. Within each polyhedral region, by applying Theorem B.4, we show that there exists a component j_0' in $\mathbf{Z}'$ such that component j_0 is mapped to j_0'. Next, under Ass. 3.4, we show that $\mathbf{f}$ consistently assigns j_0 to the same j_0' across all polyhedra. This successfully builds a one-to-one correspondence between the Gaussian components on both sides via $\mathbf{f}$. Finally, using the results from Xu et al. (2024), we conclude that $\mathbf{f}$ must be affine on the support of each component.

Theorem B.16 (Identifiability of pdGMMs from open set with common intersection)**.** *Consider a pair of finite pdGMMs in $\mathbb{R}^n$ in reduced form,*

$$\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \quad \textit{and} \quad \mathbf{Z}' \sim \sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j'), \tag{91}$$

where $|\det(\boldsymbol{\Sigma}_j)| \geq 0, \forall j \in [J]$ and $|\det(\boldsymbol{\Sigma}_j')| \geq 0, \forall j \in [J']$. Suppose $\mathbf{Z}$ satisfies Assumption 3.4. If there exists a continuous, invertible, piecewise affine function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^n$ satisfying $\mathbf{Z}' \overset{d}{=} \mathbf{f}(\mathbf{Z})$, then for all $j_0 \in [J]$, $\mathbf{f}$ is affine on $\mathcal{Z}_{j_0}$, where $\mathcal{Z}_{j_0}$ is the support of $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})$.

Proof. Fix $j_0 \in [J]$. We must show that $\mathbf{f}$ is affine on $\mathcal{Z}_{j_0}$, the support of the j_0th component. To achieve this, we will first show that, there exists an $j_0' \in [J']$, such that

$$\mathbf{f}(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})) = N(\boldsymbol{\mu}_{j_0'}', \boldsymbol{\Sigma}_{j_0'}').$$

To get this, we will show that for all $\mathbf{z}_0 \in \mathcal{Z}_{j_0}$, there exists an open ball B centered at $\mathbf{z}_0$ such that

$$(\mathbf{f}(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})))_{\mathbf{f}(B)} = (N(\boldsymbol{\mu}_{j_0'}, \boldsymbol{\Sigma}_{j_0'}))_{\mathbf{f}(B)}.$$

Let us fix some $\mathbf{z}_0 \in \mathcal{Z}_{j_0}$ and let Π be an admissible set of closed polyhedra for $\mathbf{f}$ (Definition B.13). The existence of admissible set is ensured by Proposition 22 from Bona-Pellissier et al. (2023). We consider two cases:

- Case 1. There exists $P \in \Pi$ such that $\mathbf{z}_0 \in P^\circ$, where P° is the interior part of P as defined in Def. B.13.
- Case 2. For all $P \in \Pi$, $\mathbf{z}_0 \notin P^\circ$, where P° is the interior part of P.

Case 1. Case 1 is simple. Since $\mathbf{z}_0$ is in the interior of P, we can find an open set with interior point $\mathbf{z}_0$, B, such that $B \subset P$. By definition of an admissible cover, $\mathbf{f}$ is affine on P, and thus on B, and thus on $B \cap \mathcal{Z}_j$. Since $\mathbf{f}$ is injective, we know there exists an invertible matrix H and weight vector $\mathbf{b}$ such that $\mathbf{f}(\mathbf{z}) = H\mathbf{z} + \mathbf{b}$, $\forall \mathbf{z} \in B$. Define $\phi(\mathbf{z}) := H\mathbf{z} + \mathbf{b}$ for all $\mathbf{z} \in \mathbb{R}^n$. We can thus constrain the probability measure to $\mathbf{f}(B)$ on both sides of $\mathbf{f}(\mathbf{Z}) \overset{d}{=} \mathbf{Z}'$ and write

$$\sum_{j \in J} \lambda_j \mathbf{f}\left(N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_B\right) = \sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_{\mathbf{f}(B)}\,, \tag{97}$$

$$\sum_{j \in J} \lambda_j \phi\left(N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)_B\right) = \sum_{j=1}^{J'} \lambda_j' N(\boldsymbol{\mu}_j', \boldsymbol{\Sigma}_j')_{\mathbf{f}(B)}\,. \tag{98}$$

By Lemma B.7 and $\mathbf{f}$ is injective, we have

$$\sum_{j\in J}\lambda_j\left(N(H\boldsymbol{\mu}_j+\mathbf{b},H\boldsymbol{\Sigma}_jH^T)\right)_{\mathbf{f}(B)}=\sum_{j=1}^{J'}\lambda'_jN(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(B)}\,. \tag{99}$$

Since B is open set and $\mathbf{f}$ is an open map on B (all surjective linear maps are), $\mathbf{f}(B)$ is an open set as well. Thus we can apply Theorem 3.2 and get there must exists an $j'_0\in[J']$, such that $\mathbf{f}(N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0}))=N(\boldsymbol{\mu}'_{j'_0},\boldsymbol{\Sigma}'_{j'_0})$ on $\mathbf{f}(B)$.

Case 2. Case 2 is much more technical and will thus be the focus of the rest of the proof. Since Π covers $\mathbb{R}^n$, there must be at least one $P_1\in\Pi$ such that $\mathbf{z}_0\in P_1$. Since $\mathbf{z}_0$ is not in the interior of P_1, it must be in its boundary, ∂P_1. We now argue that $\mathbf{z}_0$ must also be in a distinct polyhedron. We argue by contradiction. Suppose that is false. This means $\mathbf{z}_0$ is in the complement of all polyhedra $P\in\Pi\setminus\{P_1\}$. Since these polyhedra are closed, their complements are open, thus we can find an open set U containing $\mathbf{z}_0$ that do not intersect with any $P\in\Pi\setminus\{P_1\}$. But since $\mathbf{z}_0$ is on the boundary of P_1, this means a non-empty portion of U lies outside P_1. But this would mean this portion is outside of all $P\in\Pi$, which is a contradiction since Π covers $\mathbb{R}^n$. This means there must be at least one other polyhedron $P_2\neq P_1$ containing $\mathbf{z}_0$. Again, since $\mathbf{z}_0$ is not in the interior of P_2, it must be on its boundary, ∂P_2.

To summarize, we found two distinct polyhedra, P_1 and P_2, such that $\mathbf{z}_0\in P_1\cap P_2$ and such that $\mathbf{z}_0\in\partial P_1$ and $\mathbf{z}_0\in\partial P_2$. In general, $\mathbf{z}_0$ could be in more than two polyhedra. Let $P_1,\ldots,P_L\in\Pi$ be the polyhedra that contain $\mathbf{z}_0$. We know from the previous steps that $L\geq 2$.

Define the set of all components that intersects the point $\mathbf{z}_0$.

$$J_{\mathbf{z}_0}:=\{j\in[J]:\mathbf{z}_0\in\mathcal{Z}_j\}, \tag{100}$$

Of course, $j_0\in J_{\mathbf{z}_0}$. Since the $\mathcal{Z}_j$ are affine subspaces, they are closed sets. This means we can find an open set B' containing $\mathbf{z}_0$ small enough such that B' intersects only the components $j\in J_{\mathbf{z}_0}$. Furthermore, since all $P\in\Pi$ are closed, we can choose an open set B'' containing $\mathbf{z}_0$ that is small enough so has to not intersect with the polyhedra other than $P_1,\ldots,P_L$. Take $B_{\mathbf{z}_0}:=B'\cap B''$. Since $\mathbf{z}_0$ is on the boundaries of each $P_1,\ldots,P_L$, any open set must intersect with each P_l for $l\in[L]$. We can thus define the non-empty sets $S_l:=B_{\mathbf{z}_0}\cap P_l$ for all $l\in[L]$. Define the index set

$$J_l:=\{j\in[J]:S_l\cap\mathcal{Z}_j\neq\varnothing\}\,, \tag{101}$$

i.e. this is the set of components $\mathcal{Z}_j$ that intersects the polyhedron P_l and the open set $B_{\mathbf{z}_0}$. Notice that $J_l\subseteq J_{\mathbf{z}_0}$.

Then our goal for the following part is to show that there exists an $j'_0\in[J']$, such that $\mathbf{f}(N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0}))=N(\boldsymbol{\mu}'_{j'_0},\boldsymbol{\Sigma}'_{j'_0})$ on S_l.

But we do not need to go through all $l\in[L]$. Instead, since we are focusing on component j_0 now, in Case 2, we first explore the one on which $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})$ has positive probability, and later come back to discuss the rest of the polyhedra. For this reason, we introduce one more index notation:

$$L_{j_0}:=\{l\in[L]:N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})(S_l)>0)\}\,, \tag{102}$$

where $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})(S_l)$ means the probability that measure $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})$ take on S_l.

Pick some polyhedron $l\in[L_{j_0}]$. By condition we have $\mathbf{Z}'\overset{d}{=}\mathbf{f}(\mathbf{Z})$, we restrict the measure on both sides to $\mathbf{f}(S_l)$ and then we get

$$\left(\mathbf{f}(\sum_{j=1}^{J}\lambda_jN(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j))\right)_{\mathbf{f}(S_l)}=\left(\sum_{j=1}^{J'}\lambda'_jN(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)\right)_{\mathbf{f}(S_l)} \tag{103}$$

Since $\mathbf{f}$ is injective, we can use Lemma B.7 to switch constrain and $\mathbf{f}$ on the left hand side and we get

$$\mathbf{f}\left(\left(\sum_{j=1}^{J}\lambda_j N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\right)_{S_l}\right)=\left(\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)\right)_{\mathbf{f}(S_l)}, \tag{104}$$

$$\mathbf{f}\left(\sum_{j=1}^{J}\lambda_j N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)_{S_l}\right)=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)}, \tag{105}$$

$$\sum_{j=1}^{J}\lambda_j \mathbf{f}\left(N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)_{S_l}\right)=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)}\,, \tag{106}$$

where the second line uses the fact that the restriction of a mixture is the mixture of the restriction and the third line uses the fact that the pushforward of a mixture is the mixture of the pushforwards.

Notice that if $\mathcal{Z}_j \cap S_l = \varnothing$, we have that $N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)_{S_l} = 0$, i.e. it is the zero measure. The pushforward of the zero measure is also the zero measure, thus

$$\sum_{j\in J_l}\lambda_j \mathbf{f}\left(N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)_{S_l}\right)=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)} \tag{107}$$

Since $\mathbf{f}$ is piecewise affine and injective, we know there exists an invertible matrix H^l and weight vector $\mathbf{b}^l$ such that $\mathbf{f}(\mathbf{z}) = H^l\mathbf{z} + \mathbf{b}^l$, $\forall \mathbf{z} \in S_l$. Define $\phi^l(\mathbf{z}) := H^l\mathbf{z} + \mathbf{b}^l$ for all $\mathbf{z} \in \mathbb{R}^n$. We can thus write

$$\sum_{j\in J_l}\lambda_j \phi^l\left(N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)_{S_l}\right)=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)}\,, \tag{108}$$

since $\mathbf{f} = \phi^l$ on S_l. Since ϕ^l is injective, we can apply Lemma B.7 to get

$$\sum_{j\in J_l}\lambda_j \phi^l\left(N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)\right)_{\phi^l(S_l)}=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)} \tag{109}$$

$$\sum_{j\in J_l}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)_{\mathbf{f}(S_l)}=\sum_{j=1}^{J'}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)}\,. \tag{110}$$

Define $J'_l := \{j \in [J'] : \mathcal{Z}'_j \cap \mathbf{f}(S_l) \neq \varnothing\}$. This allows us to write

$$\sum_{j\in J_l}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)_{\mathbf{f}(S_l)}=\sum_{j\in J'_l}\lambda'_j N(\boldsymbol{\mu}'_j,\boldsymbol{\Sigma}'_j)_{\mathbf{f}(S_l)}. \tag{111}$$

Now we discuss the possible relations between $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})$ and S_l. If $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})$ takes probability zero on S_l, then we can skip it. If not, then we face two cases:

- Case 2.1 $\mathcal{Z}_{j_0}$ intersects with the interior of S_l, i.e. $\mathcal{Z}_{j_0} \cap S_l^\circ \neq \varnothing$, where $S_l^\circ := B_{\mathbf{z}_0} \cap P_l^\circ$.
- Case 2.2 $\mathcal{Z}_{j_0}$ only intersects with the boundary of S_l with positive probability, i.e. $\mathcal{Z}_{j_0} \cap S_l^\circ = \varnothing$ and $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})(\partial S_l) > 0$, where $\partial S_l := B_{\mathbf{z}_0} \cap \partial P_l$.

Case 2.1. For Case 2.1, define the index set that also intersects with the interior part of S_l:

$$J_l^\circ := \{j \in J_l : \mathcal{Z}_j \cap S_l^\circ \neq \varnothing\}, \quad J_l'^\circ := \{j \in J'_l : \mathcal{Z}'_j \cap \mathbf{f}(S_l^\circ) \neq \varnothing\}.$$

We further constrain Equation (111) on $\mathbf{f}(S_l^\circ)$ and cancel out the purely zero measure terms on both sides, i.e. the terms are not in J_l° and $J_l'^\circ$

$$\sum_{j\in J_l^\circ}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)_{\mathbf{f}(S_l^\circ)}=\sum_{j\in J_l'^\circ}\lambda_j' N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j')_{\mathbf{f}(S_l^\circ)}. \tag{112}$$

$\mathbf{f}$ is an open map on S_l° since it is affine on S_l°. Therefore, $\mathbf{f}(S_l^\circ)$ is an open set as well. By the definition of J_l°, we have $\forall j\in J_l^\circ$, $\mathcal{Z}_j\cap S_l^\circ\neq\varnothing$, then we can derive $\mathbf{f}(\mathcal{Z}_j\cap S_l^\circ)\neq\varnothing, \forall j\in J_l^\circ$. Since $\mathbf{f}$ is invertible function, we can further derive $\forall j\in J_l, \mathbf{f}(\mathcal{Z}_j)\cap\mathbf{f}(S_l)\neq\varnothing$. Therefore, we can use Theorem 3.2 and get

$$\sum_{j\in J_l^\circ}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)=\sum_{j\in J_l'^\circ}\lambda_j' N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j'), \tag{113}$$

which means $|J_l^\circ|=|J_l'^\circ|$ and there exists a permutation $\pi^l: J_l^\circ\to J_l'^\circ$ such that

$$H^l\boldsymbol{\Sigma}_j(H^l)^T=\boldsymbol{\Sigma}'_{\pi^l(j)},\quad H^l\boldsymbol{\mu}_j+\mathbf{b}^l=\boldsymbol{\mu}'_{\pi^l(j)}. \tag{114}$$

Case 2.2. For Case 2.2, because of the definition of L_{j_0} (Equation (102)) where we pick up the S_l, we know $N(\boldsymbol{\mu}_{j_0},\boldsymbol{\Sigma}_{j_0})$ has positive probability on ∂S_l. Since we know $\mathcal{Z}_{j_0}$ is an affine space, it can only intersect with one facet of ∂S_l. We denote this facet $\partial S_l^{j_0}$. Define the index sets that the supports intersect with this facet with positive probability:

$$J_l^{j_0}:=\{j\in J_l\backslash J_l^\circ: N(\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)(\partial S_l^{j_0})>0\},\quad J_l'^{j_0}:=\{j\in J_l'\backslash J_l'^\circ: N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j')(\mathbf{f}(\partial S_l^{j_0}))>0\}.$$

Constrain Equation (111) further on $\mathbf{f}(\partial S_l^{j_0})$ and eliminate the probability zero terms then we get

$$\sum_{j\in J_l^{j_0}}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)_{\mathbf{f}(\partial S_l^{j_0})}=\sum_{j\in J_l'^{j_0}}\lambda_j' N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j')_{\mathbf{f}(\partial S_l^{j_0})}. \tag{115}$$

Since all the Gaussian components in $J_l^{j_0}$ take a positive probability on $\partial S_l^{j_0}$, which implies their rank is lower than or equal to the rank of $\partial S_l^{j_0}$ and their support is parallel with $\partial S_l^{j_0}$. Then it must have no intersection with the interior of all adjacent polyhedra. Thus, we can construct an open set $E\subset\mathbb{R}^n$, such that $\partial S_l^{j_0}\subset E$. And expand the constrained measure from $\mathbf{f}(\partial S_l^{j_0})$ to $\mathbf{f}(E)$ and then we have

$$\sum_{j\in J_l^{j_0}}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)_{\mathbf{f}(E)}=\sum_{j\in J_l'^{j_0}}\lambda_j' N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j')_{\mathbf{f}(E)}. \tag{116}$$

$\mathbf{f}$ is an open map on E. Therefore, $\mathbf{f}(E)$ is also an open set. Therefore, we can use Theorem 3.2 and get

$$\sum_{j\in J_l^{j_0}}\lambda_j N(H^l\boldsymbol{\mu}_j+\mathbf{b}^l, H^l\boldsymbol{\Sigma}_j(H^l)^T)=\sum_{j\in J_l'^{j_0}}\lambda_j' N(\boldsymbol{\mu}_j',\boldsymbol{\Sigma}_j'), \tag{117}$$

which means $|J_l^{j_0}|=|J_l'^{j_0}|$ and there exists a permutation $\pi^l: J_l^{j_0}\to J_l'^{j_0}$ such that

$$H^l\boldsymbol{\Sigma}_j(H^l)^T=\boldsymbol{\Sigma}'_{\pi^l(j)},\quad H^l\boldsymbol{\mu}_j+\mathbf{b}^l=\boldsymbol{\mu}'_{\pi^l(j)}. \tag{118}$$

To sum up, Equation (114) holds for all S_l in Case 2.1, and Equation (118) holds for all S_l in Case 2.2, which covers all $\mathcal{Z}_{j_0}\cap B_{\mathbf{z}_0}$. Thus, we can summarize it as for all $l\in[L_{j_0}]$

$$H^l\boldsymbol{\Sigma}_j(H^l)^T=\boldsymbol{\Sigma}'_{\pi^l(j)},\quad H^l\boldsymbol{\mu}_j+\mathbf{b}^l=\boldsymbol{\mu}'_{\pi^l(j)}. \tag{119}$$

In the following part, we focus on the points $\mathbf{z}_0$ satisfying Ass. 3.4, denote this set as $\mathcal{S}_\mathcal{Z}$. We aim to show that, consider $\mathbf{z}_0\in\mathcal{S}_\mathcal{Z}$, for all $l\in[L_{j_0}]$, permutations π^l map j_0 to the same index in $\mathcal{Z}'$.

We prove the claim for all $\mathbf{z}_0\in\mathcal{S}_\mathcal{Z}$ by contradiction: suppose there exist $l_1,l_2\in[L_{j_0}]$, such that $\pi^{l1}(j_0)\neq\pi^{l2}(j_0)$. Denote $\pi^{l1}(j_0):=j_0'$, this means, there exists a polyhedron $l*\in[L_{j_0}]$, such that $(\pi^{l*})^{-1}(j_0')\neq(\pi^{l_1})^{-1}(j_0')$. Denote $j*:=(\pi^{l*})^{-1}(j_0')$, we have $j*\neq j_0$.

We can derive this

$$\begin{cases} H^{l_1}\boldsymbol{\Sigma}_{j_0}(H^{l_1})^T = \boldsymbol{\Sigma}'_{j'_0}, & H^{l_1}\boldsymbol{\mu}_{j_0} + \mathbf{b}^{l_1} = \boldsymbol{\mu}'_{j'_0}, \\ H^{l*}\boldsymbol{\Sigma}_{j*}(H^{l*})^T = \boldsymbol{\Sigma}'_{j'_0}, & H^{l*}\boldsymbol{\mu}_{j*} + \mathbf{b}^{l*} = \boldsymbol{\mu}'_{j'_0}. \end{cases} \tag{120}$$

$$\Rightarrow \begin{cases} H^{l_1}\boldsymbol{\Sigma}_{j_0}(H^{l_1})^T = H^{l*}\boldsymbol{\Sigma}_{j*}(H^{l*})^T, \\ H^{l_1}\boldsymbol{\mu}_{j_0} + \mathbf{b}^{l_1} = H^{l*}\boldsymbol{\mu}_{j*} + \mathbf{b}^{l*}. \end{cases} \tag{121}$$

Since $\mathbf{z}_0$ is in both polyhedra l_1 and $l*$, we get

$$H^{l_1}\mathbf{z}_0 + \mathbf{b}^{l_1} = \mathbf{f}(\mathbf{z}_0) = H^{l*}\mathbf{z}_0 + \mathbf{b}^{l*} \tag{122}$$

By taking the difference between Equation (122) and Equation (121), we have

$$H^{l_1}(\mathbf{z}_0 - \boldsymbol{\mu}_{j_0}) = H^{l*}(\mathbf{z}_0 - \boldsymbol{\mu}_{j*}). \tag{123}$$

Suppose $rank(\boldsymbol{\Sigma}'_{j'_0}) = k$. By Equation (121), since both H^{l_1} and H^{l_2} are invertible matrix, we have

$$rank(\boldsymbol{\Sigma}'_{j'_0}) = rank(\boldsymbol{\Sigma}_{j_0}) = rank(\boldsymbol{\Sigma}_{j*}) = k. \tag{124}$$

By the definition of degenerate Gaussian random variable, we can find a $n \times k$ matrix A_1 where $A_1A_1^T = \boldsymbol{\Sigma}_{j_0}$ and $rank(A_1) = k$, and have

$$A_1\boldsymbol{\varepsilon} + \boldsymbol{\mu}_{j_0} \sim N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0}),$$

where $\boldsymbol{\varepsilon}$ follows a standard k-dimensional normal distribution. Hence, there exists a unique value $\boldsymbol{\varepsilon}_1 \in \mathbb{R}^k$, such that $\mathbf{z}_0 = A_1\boldsymbol{\varepsilon}_1 + \boldsymbol{\mu}_{j_0}$. The same analysis for $N(\boldsymbol{\mu}_{j*}, \boldsymbol{\Sigma}_{j*})$, there exists a unique $\boldsymbol{\varepsilon}_2 \in \mathbb{R}^k$ such that $\mathbf{z}_0 = A_2\boldsymbol{\varepsilon}_2 + \boldsymbol{\mu}_{j*}$.

Substitute into Equation (123), we get

$$H^{l_1}A_1\boldsymbol{\varepsilon}_1 = H^{l_2}A_2\boldsymbol{\varepsilon}_2. \tag{125}$$

Additionally, we have $A_1A_1^T = \boldsymbol{\Sigma}_{j_0}$ and $A_2A_2^T = \boldsymbol{\Sigma}_{j*}$. Substitute into Equation (121), we get

$$H^{l_1}A_1(H^{l_1}A_1)^T = H^{l_2}A_2(H^{l_2})A_2)^T.$$

First we have $rank(H^{l_1}A_1) = k$ as H^{l_1} is invertible. By Lemma B.17, this implies there exists an orthogonal matrix $Q \in \mathbb{R}^{k\times k}$, such that $H^{l_1}A_1 = H^{l_2}A_2Q$. Substitute into Equation (125), we have

$$H^{l_1}A_1(\boldsymbol{\varepsilon}_1 - Q^T\boldsymbol{\varepsilon}_2) = 0.$$

Then we can derive $\boldsymbol{\varepsilon}_1 - Q^T\boldsymbol{\varepsilon}_2 = 0$, which implies $||\boldsymbol{\varepsilon}_1|| = ||\boldsymbol{\varepsilon}_2||$.Note that $||\boldsymbol{\varepsilon}_1|| = \sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_{j_0})^T\boldsymbol{\Sigma}_{j_0}^{-1}(\mathbf{z}_0 - \boldsymbol{\mu}_{j_0})}$ and $||\boldsymbol{\varepsilon}_2|| = \sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_{j_*})^T\boldsymbol{\Sigma}_{j_*}^{-1}(\mathbf{z}_0 - \boldsymbol{\mu}_{j_*})}$. This means for all $\mathbf{z}_0 \in \mathcal{S}_{\mathcal{Z}}$, the Mahalanobis distance of $\mathbf{z}_0$ to $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})$ and $N(\boldsymbol{\mu}_{j_*}, \boldsymbol{\Sigma}_{j_*})$ are the same.

We can find it will be contradictory to Assumption 3.4 if $\pi^{l_1}(j_0) \neq \pi^{l_2}(j_0)$. Therefore, we can claim there exists $j'_0 \in [J']$, such that $\pi^l(j_0) = j'_0$ for all $l \in [L_{j_0}]$. Due to Equation (114) and the identifiability of pdMVN, we have

$$N(H^l\boldsymbol{\mu}_{j_0} + \mathbf{b}^l, H^l\boldsymbol{\Sigma}_{j_0}(H^l)^T) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0}), \quad \forall l \in [L_{j_0}]. \tag{126}$$

Now we finished all the discussion of $l \in [L_{j_0}]$, i.e. all the S_l where $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})$ can take positive probability. However, as our goal is to show $\mathbf{f}(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})$ over $\mathcal{Z}_{j_0} \cap B$, there is a probability-zero subdomain missing in the above analysis. Next, we discuss the rest part $l \in [L]\backslash[L_{j_0}]$. These cases can only happen if $\mathcal{Z}_{j_0}$ intersects with the boundary of S_l only, i.e. $\mathcal{Z}_{j_0} \cap S_l^\circ = \varnothing$, $\mathcal{Z}_{j_0} \cap \partial S_l \neq \varnothing$ and $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})(\partial S_l) = 0$. For this case, since $\mathcal{Z}_{j_0} \cap \partial S_l \neq \varnothing$ and $N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})(\partial S_l) = 0$, we know for every point $\mathbf{z}* \in \mathcal{Z}_{j_0} \cap \partial S_l$, there must exist an adjecent S_{l*} such that $\mathbf{z}* \in S_{l*}$ and $l* \in [L_{j_0}]$. Therefore, we can conclude that $\mathcal{Z}_{j_0} \cap B \subseteq \cup_{l\in[L_{j_0}]} S_l$.

We now define a partition of the set $B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}$ by S_l, $l \in [L_{j_0}]$.

For any non-empty subset $I \subseteq [L_{j_0}]$, we define

$$D_I := (\cap_{l' \in I} S_{l'}) \setminus (\cup_{l' \notin I} S_{l'}) .$$

Note that all the affine functions $H^l \mathbf{z} + \mathbf{b}^l$ with $l \in I$ agree on D_I since $D_I \subseteq \cap_{l \in I} P_l$. For simplicity, we can directly use $l_I := \min I$, i.e. the smallest index in I. When $l = l_I$ we constrain the measures on both sides of Equation (126) on subdomain $\mathbf{f}(D_I)$ and get

$$N(H^{l_I} \boldsymbol{\mu}_{j_0} + \mathbf{b}^{l_I}, H^{l_I} \boldsymbol{\Sigma}_{j_0} (H^{l_I})^T)_{\mathbf{f}(D_I)} = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(D_I)} \tag{127}$$

Sum up all the $I \subseteq [L_{j_0}]$, we have

$$\sum_{I \subseteq [L_{j_0}]} N(H^{l_I} \boldsymbol{\mu}_{j_0} + \mathbf{b}^{l_I}, H^{l_I} \boldsymbol{\Sigma}_{j_0} (H^{l_I})^T)_{\mathbf{f}(D_I)} = \sum_{I \subseteq [L_{j_0}]} N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(D_I)}, \tag{128}$$

$$\sum_{I \subseteq [L_{j_0}]} \mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})_{D_I}\right) = \sum_{I \subseteq [L_{j_0}]} N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(D_I)}. \tag{129}$$

Note that D_I's are disjoint from each other. Thus, we can switch sum into union and get

$$\mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})_{\cup_{I \subseteq [L_{j_0}]} D_I}\right) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(\cup_{I \subseteq [L_{j_0}]} D_I)}. \tag{130}$$

Since D_I is partition set of $B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}$, which means $\cup_{I \subseteq [L_{j_0}]} D_I = B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}$, substitute into Equation (130) and then we have

$$\mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})_{B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}}\right) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0})}, \tag{131}$$

$$N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})_{B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}} = \mathbf{f}^{-1}\left(N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})_{\mathbf{f}(B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0})}\right). \tag{132}$$

Since $\mathbf{f}$ is bijective on $\mathbb{R}^n$, we have $\mathbf{f}^{-1}$ is injective. Thus by apply Lemma B.7, we can conclude Case 2:

$$N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})_{B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}} = \mathbf{f}^{-1}\left(N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})\right)_{B_{\mathbf{z}_0} \cap \mathcal{Z}_{j_0}}. \tag{133}$$

Note that Equation (133) is true for all $\mathbf{z}_0 \in \mathcal{S}_{\mathcal{Z}}$, i.e. we can construct an open set $B_{\mathbf{z}_0}$ containing $\mathbf{z}_0$ for all $\mathbf{z}_0 \in \mathcal{S}_{\mathcal{Z}}$. Notice that $\{B_{\mathbf{z}_0} \mid \mathbf{z}_0 \in\in \mathcal{S}_{\mathcal{Z}}\}$ is thus an open cover of the support of the $\mathcal{Z}_{j_0}$. We can thus apply Lemma B.18 to get

$$N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0}) = \mathbf{f}^{-1}\left(N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})\right) \tag{134}$$

$$\mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})\right) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0}). \tag{135}$$

Then, since $\mathbf{f}$ is injective and piecewise affine, by Theorem B.19 from Xu et al. (2024), there exists an invertible affine transformation $\mathbf{h}_{j_0} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathbf{h}_{j_0}(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})) = N(\boldsymbol{\mu}'_{j'_0}, \boldsymbol{\Sigma}'_{j'_0})$. Substitute into Equation (135) then we have

$$\mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})\right) = \mathbf{h}_{j_0}(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})) \quad \Rightarrow \quad \mathbf{h}_{j_0}^{-1} \circ \mathbf{f}\left(N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0})\right) = N(\boldsymbol{\mu}_{j_0}, \boldsymbol{\Sigma}_{j_0}). \tag{136}$$

As $\mathbf{h}_{j_0}^{-1} \circ \mathbf{f}$ is a continuous piecewise affine function, by Lemma B.20 from Xu et al. (2024), we can conclude that $\mathbf{h}_{j_0}^{-1} \circ \mathbf{f}$ is affine on $\mathcal{Z}_{j_0}$. Furthermore, we can conclude that $\mathbf{f}$ is affine on $\mathcal{Z}_{j_0}$. This result holds for all $j_0 \in [J]$. □

Theorem 3.5 (Identifiability up to ATwC of pdGMM latent variables)**.** *Assume that* $\mathbf{Z}$ *is a pdGMM and the observation* $\mathbf{X} = \mathbf{f}(\mathbf{Z})$ *follows the data-generating process in Sec. 2 for an injective continuous piecewise affine mixing function* $\mathbf{f}$*, and* $\mathbf{Z}$ *satisfies Ass. 3.4. Let* $\mathbf{g} : \mathcal{X} \to \mathbb{R}^n$ *be an invertible continuous piecewise affine function and* $\hat{\mathbf{f}} : \mathbb{R}^n \to \mathbb{R}^d$ *be an invertible piecewise affine function onto its image. If both conditions hold,*

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

$$\mathbf{g}(\mathbf{X}) \sim \sum_{j=1}^{J'} \lambda'_j N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j), \tag{3}$$

for appropriate values of $\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j$ *(in reduced form), then J'=J and* $\mathbf{Z}$ *is identified by* $\mathbf{g}(\mathbf{X})$ *up to affine transformation within components, i.e.,* $\mathbf{g} \circ \mathbf{f}$ *is an invertible affine function on* $\mathcal{Z}_j$*,* $j \in [J]$.

Proof. Since $\mathbf{X} = \mathbf{f}(\mathbf{Z})$, we can rewrite Equation (2) (perfect reconstruction) as

$$\mathbb{E}||\mathbf{f}(\mathbf{Z}) - \hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))||_2^2 = 0\,. \tag{137}$$

This means that $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\mathbf{g}(\mathbf{f}(\mathbf{Z})))$ are equal $\mathbb{P}_{\mathbf{Z}}$-almost everywhere, which implies $\mathbf{f}(\mathbf{Z})$ and $\hat{\mathbf{f}}(\mathbf{g}(\mathbf{X}))$ are equally distributed with the same image space. Therefore, we can define $\mathbf{h} := \hat{\mathbf{f}}^{-1} \circ \mathbf{f}$ on $\mathcal{Z}$, which is a continuous and invertible piecewise affine function, and we have $\mathbf{Z} \overset{d}{=} \mathbf{h}^{-1}(\mathbf{g}(\mathbf{X}))$, where both $\mathbf{Z}$ and $\mathbf{g}(\mathbf{X})$ are pdGMMs.

Since $\mathbf{Z}$ satisfies Ass. 3.4, and $\mathbf{h}^{-1}$ is a continuous and invertible piecewise affine function, we can apply Theorem B.16 and get $\mathbf{h}^{-1}$ is affine on $\mathcal{Z}'_j$, where $\mathcal{Z}'_j$ is the support of $N(\boldsymbol{\mu}'_j, \boldsymbol{\Sigma}'_j)$. One step more, we can conclude that $\mathbf{h}$, i.e., $\hat{\mathbf{f}}^{-1} \circ \mathbf{f}$, is affine on $\mathcal{Z}_j$.

□

B.2.2 From affine on components to affine everywhere

While Theorem 3.5 proves that $\mathbf{g} \circ \mathbf{f}$ is affine within each individual component, it does not guarantee that there is a global affine function across all components. In this section, we seek conditions under which $\mathbf{g} \circ \mathbf{f}$ is globally affine, i.e., affine on the whole support for the pdGMM. In the following proposition (Prop. B.21), we show that if the supports of all components share a common basis (i.e., each support can be spanned by a subset of this common basis), then it is possible to construct a consistent affine function across the entire domain.

Assumption 3.6 (Common basis and translation vector)**.** Let $\mathcal{Z}_1, \dots, \mathcal{Z}_J \subseteq \mathbb{R}^n$ be affine spaces such that there exists $\mathbf{z}^0 \in \bigcap_{j=1}^J \mathcal{Z}_j$. We assume there exists a basis of $\mathbb{R}^n$, denoted by $\{\mathbf{z}^1, \dots, \mathbf{z}^n\}$, such that for all $j \in [J]$, there exists $\mathcal{K}_j \subseteq [n]$ s.t. $\{\mathbf{z}^k \mid k \in \mathcal{K}_j\}$ is a basis for the vector space $\mathcal{Z}_j - \mathbf{z}^0$.

Proposition B.21. *Let $\mathcal{Z}_1, \dots, \mathcal{Z}_J \subseteq \mathbb{R}^n$ be affine spaces that satisfy Ass. 3.6, and let $f : \mathbb{R}^n \to \mathbb{R}$ be a function such that, for all $j \in [J]$, the restriction of f to $\mathcal{Z}_j$ is affine. Then the restriction of f to $\mathcal{Z} := \cup_{i\in[m]}\mathcal{Z}_i$ is affine.*

Proof. Our goal is to construct an affine function $\mathbf{a}\mathbf{z} + b$ with $\mathbf{a} \in \mathbb{R}^{1\times n}$ and $b \in \mathbb{R}$ such that this function agrees with f on $\mathcal{Z}$.

First notice that, since f is affine on each $\mathcal{Z}_j$, we have: for each $j \in [J]$, there exists $\mathbf{a}^j \in \mathbb{R}^{1\times n}$ and $b^j \in \mathbb{R}$ such that, for all $\mathbf{z} \in \mathcal{Z}_j$, $f(\mathbf{z}) = \mathbf{a}^j\mathbf{z} + b^j$.

Since $\mathbf{z}^0 \in \cap_{j\in[J]}\mathcal{Z}_j$, we have that, for all $j \in [J]$, $f(\mathbf{z}^0) = \mathbf{a}^j\mathbf{z}^0 + b^j$ for all $j \in [J]$.

Let us define $\mathbf{B} := [\mathbf{z}^1 \ \cdots \ \mathbf{z}^n] \in \mathbb{R}^{n\times n}$, which is invertible since its columns form a basis by assumption.

Fix $j \in [J]$ and $k \in \mathcal{K}_j$. We know that

$$\begin{aligned}
f(\mathbf{z}^k + \mathbf{z}^0) &= \mathbf{a}^j(\mathbf{z}^k + \mathbf{z}^0) + b^j && (138)\\
&= \mathbf{a}^j\mathbf{z}^k + \mathbf{a}^j\mathbf{z}^0 + b^j && (139)\\
&= \mathbf{a}^j\mathbf{z}^k + f(\mathbf{z}^0) && (140)\\
&= \underbrace{\mathbf{a}^j\mathbf{B}}_{\tilde{\mathbf{a}}^j :=}\mathbf{e}^k + f(\mathbf{z}^0) && (141)\\
&= \tilde{\mathbf{a}}^j\mathbf{e}^k + f(\mathbf{z}^0) && (142)\\
&= \tilde{\mathbf{a}}^j_k + f(\mathbf{z}^0) && (143)\\
\implies \tilde{\mathbf{a}}^j_k &= f(\mathbf{z}^k + \mathbf{z}^0) - f(\mathbf{z}^0) \quad \forall k \in \mathcal{K}_j\,. && (144)
\end{aligned}$$

Let us define

$$\begin{aligned}
\mathbf{a} &:= \left[\left(f(\mathbf{z}^1 + \mathbf{z}^0) - f(\mathbf{z}^0)\right) \quad \cdots \quad \left(f(\mathbf{z}^n + \mathbf{z}^0) - f(\mathbf{z}^0)\right)\right]\mathbf{B}^{-1} \in \mathbb{R}^{1\times n} \text{and} && (145)\\
b &:= f(\mathbf{z}^0) - \mathbf{a}\mathbf{z}^0\,. && (146)
\end{aligned}$$

We must now show that, for all $\mathbf{z} \in \cup_{j\in[J]}\mathcal{Z}_j$, $f(\mathbf{z}) = \mathbf{a}\mathbf{z} + b$. Fix some arbitrary $\mathbf{z} \in \cup_{j\in[J]}\mathcal{Z}_j$. We know there has to be one $j_0 \in [J]$ s.t. $\mathbf{z} \in \mathcal{Z}_{j_0}$. Hence $f(\mathbf{z}) = \mathbf{a}^{j_0}\mathbf{z} + b^{j_0}$. We showed earlier that $b^j = f(\mathbf{z}^0) - \mathbf{a}^j\mathbf{z}^0$ for all $j \in [J]$,

hence we have $f(\mathbf{z}) = \mathbf{a}^{j_0}\mathbf{z} + f(\mathbf{z}^0) - \mathbf{a}^{j_0}\mathbf{z}^0$. Since $\mathbf{B}$ is invertible, there exists $\tilde{\mathbf{z}} \in \mathbb{R}^n$ such that $\mathbf{z} - \mathbf{z}^0 = B\tilde{\mathbf{z}}$. Since $\{\mathbf{z}^k \mid k \in \mathcal{K}_j\}$ forms a basis for $\mathcal{Z}_j - \mathbf{z}^0$, we have that whenever $k \notin \mathcal{K}_j$, $\tilde{\mathbf{z}}_k = 0$. We can thus write

$$
\begin{aligned}
f(\mathbf{z}) &= \mathbf{a}^{j_0}\mathbf{z} + b^{j_0} && (147)\\
&= \mathbf{a}^{j_0}(B\tilde{\mathbf{z}} + \mathbf{z}^0) + b^{j_0} && (148)\\
&= \mathbf{a}^{j_0}B\tilde{\mathbf{z}} + \mathbf{a}^{j_0}\mathbf{z}^0 + b^{j_0} && (149)\\
&= \tilde{\mathbf{a}}^{j_0}\tilde{\mathbf{z}} + f(\mathbf{z}^0) && (150)\\
&= \sum_{k\in\mathcal{K}_j} \tilde{\mathbf{a}}_k^{j_0}\tilde{\mathbf{z}}_k + \sum_{k\notin\mathcal{K}_j} \tilde{\mathbf{a}}_k^{j_0}\underbrace{\tilde{\mathbf{z}}_k}_{0} + f(\mathbf{z}^0) && (151)\\
&= \sum_{k\in\mathcal{K}_j} \tilde{\mathbf{a}}_k^{j_0}\tilde{\mathbf{z}}_k + f(\mathbf{z}^0) && (152)\\
&= \sum_{k\in\mathcal{K}_j} \left(f(\mathbf{z}^k + \mathbf{z}^0) - f(\mathbf{z}^0)\right)\tilde{\mathbf{z}}_k + f(\mathbf{z}^0) && (153)\\
&= \sum_{k\in\mathcal{K}_j} \left(f(\mathbf{z}^k + \mathbf{z}^0) - f(\mathbf{z}^0)\right)\tilde{\mathbf{z}}_k + \sum_{k\notin\mathcal{K}_j} \left(f(\mathbf{z}^k + \mathbf{z}^0) - f(\mathbf{z}^0)\right)\underbrace{\tilde{\mathbf{z}}_k}_{0} + f(\mathbf{z}^0) && (154)\\
&= \begin{bmatrix}(f(\mathbf{z}^1 + \mathbf{z}^0) - f(\mathbf{z}^0)) & \cdots & (f(\mathbf{z}^n + \mathbf{z}^0) - f(\mathbf{z}^0))\end{bmatrix}\tilde{\mathbf{z}} + f(\mathbf{z}^0) && (155)\\
&= \begin{bmatrix}(f(\mathbf{z}^1 + \mathbf{z}^0) - f(\mathbf{z}^0)) & \cdots & (f(\mathbf{z}^n + \mathbf{z}^0) - f(\mathbf{z}^0))\end{bmatrix}\mathbf{B}^{-1}(\mathbf{z} - \mathbf{z}^0) + f(\mathbf{z}^0) && (156)\\
&= \mathbf{a}(\mathbf{z} - \mathbf{z}^0) + f(\mathbf{z}^0) && (157)\\
&= \mathbf{a}\mathbf{z} + \underbrace{f(\mathbf{z}^0) - \mathbf{a}\mathbf{z}^0}_{b:=} && (158)\\
&= \mathbf{a}\mathbf{z} + b\,. && (159)
\end{aligned}
$$

Since $\mathbf{z} \in \cup_{j\in[J]}\mathcal{Z}_j$ was arbitrary, the above equality holds for all $\mathbf{z}$ in $\mathcal{Z} = \cup_{j\in[J]}\mathcal{Z}_j$. This concludes the proof. □

It is easy to construct examples of collections of subspaces $\mathcal{Z}_1, \ldots, \mathcal{Z}_J$ satisfying the condition of Prop. B.21. One can simply choose some basis $\mathbf{z}^1, \cdots \mathbf{z}^n$ of $\mathbb{R}^n$ and construct each $\mathcal{Z}_j$ as the span of some subset of this basis. In particular, one can use the standard basis $\mathbf{e}^1, \ldots, \mathbf{e}^n$, which will yield axis-aligned subspaces of $\mathbb{R}^n$. We provide a counter-example in Appendix E.

Then by combining Ass. 3.6 with Theorem B.16, we can directly conclude the identifiability up to affine transformation in the following theorem.

Theorem 3.7 (Identifiability up to AT of pdGMM latent variables)**.** *We assume the same conditions as Thm. 3.5, and that the supports of components $\mathcal{Z}_1, \ldots, \mathcal{Z}_J$ satisfy the common basis assumption (Ass. 3.6), then the representation $\mathbf{g}(\mathbf{X})$ identifies $\mathbf{Z}$ up to a single global affine transformation across components, i.e., $\mathbf{g} \circ \mathbf{f}$ is affine on all of $\mathcal{Z}$.*

Proof. Due to Theorem 3.5, we have $\mathbf{g} \circ \mathbf{f}$ is affine restricted on $\mathcal{Z}_1, ..., \mathcal{Z}_J$ individually. In addition, since we assume $\mathcal{Z}_1, ..., \mathcal{Z}_J$ satisfying Ass. 3.6, by applying Prop. B.21, we can conclude $\mathbf{g} \circ \mathbf{f}$ is affine on $\mathcal{Z} = \cup_{j=1}^{J}\mathcal{Z}_j$. □

B.2.3 From affine identifiability to element-wise identifiability via sparsity principle

In this section, we take a step further to investigate the possibility of achieving a stronger level of identifiability—namely, up to permutation and scaling indeterminacies only. We first introduce the following assumption, which is a special case of Ass. 3.6 that assumes zero translation vector and standard common basis of $\mathbb{R}^n$.

Assumption 3.8. Let $\mathcal{Z}_1, \ldots, \mathcal{Z}_J \subseteq \mathbb{R}^n$ be the supports of the components of $\mathbf{Z}$. We assume:

- [Common Standard Basis] For all $j \in [J]$, each $\mathcal{Z}_j$ is a vector space, and there exists an index set $\mathcal{K}_j \subseteq [n]$ such that $\{\mathbf{e}^k \mid k \in \mathcal{K}_j\}$ forms a basis for $\mathcal{Z}_j$, where $\{\mathbf{e}^1, \ldots, \mathbf{e}^n\}$ denotes the standard basis of $\mathbb{R}^n$ (one-hot vectors).

- [Sufficient Support Basis Index Variability] For every $i \in [n]$, the union of the support index sets $\mathcal{K}_j$ that do not include i covers all other standard basis directions:

$$\bigcup_{j \in [J] \mid i \notin \mathcal{K}_j} \mathcal{K}_j = [n] \setminus \{i\}.$$

We then adapt a key argument due to Lachapelle et al. (2023a) in the context of sparse multitask learning and later adapted by Xu et al. (2024) in the context of partial observability. In our case, the masked latent variables exactly follow a piecewise-degenerate Gaussian mixture model (pdGMM) with supports that share a common standard basis. The adapted result is formalized as follows:

Lemma B.22 (Element-wise Identifiability from sparsity (Lachapelle et al., 2023a; Xu et al., 2024))**.** *Assume that the latent variables* $\mathbf{Z}$ *with support* $\mathcal{Z}$ *follows a pdGMM (Def. B.1). The supports of its components satisfy Ass. 3.8. Let the function* $\mathbf{v} : \mathcal{Z} \to \mathbb{R}^n$ *be invertible and affine on* $\mathcal{Z}$*, and*

$$\mathbb{E} \left\| \mathbf{v}(\mathbf{Z}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 . \tag{160}$$

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

Since from Theorem 3.7 the mixing process is successfully reduced from the piecewise affine to affine, we can directly apply Lemma B.22 on this results under Ass. 3.8 to conclude the following theorem.

Theorem 3.9 (Identifiability up to PS of pdGMM latent variables)**.** *We assume the same conditions as Thm. 3.5 and that the supports of components* $\mathcal{Z}_1, \ldots, \mathcal{Z}_J$ *satisfy Ass. 3.8. If the following holds:*

$$\mathbb{E} \left\| \mathbf{g}(\mathbf{X}) \right\|_0 \leq \mathbb{E} \left\| \mathbf{Z} \right\|_0 , \tag{4}$$

where $\|\cdot\|_0$ *means L0 norm, then the representation* $\mathbf{g}(\mathbf{X})$ *identifies* $\mathbf{Z}$ *up to PST, i.e.,* $\mathbf{g} \circ \mathbf{f}$ *is a permutation combined with an element-wise linear transformation on* $\mathcal{Z}$*.*

Proof. Since satisfying Ass. 3.8 implies that Ass. 3.6 holds, by Theorem 3.7, we have $\mathbf{g} \circ \mathbf{f}$ is affine and invertible on $\mathcal{Z}$. By Lemma. B.22, we can conclude $\mathbf{g} \circ \mathbf{f}$ is a permutation composed with an element-wise invertible affine transformation on $\mathcal{Z}$. □

C Implementation details and Licences

This section provides further details about the experiment implementation in Section 4. The implementation is built on the open-source code by Lachapelle et al. (2022) released under Apache 2.0 License; the code by Ahuja et al. (2022) released under MIT License; the code by Zheng et al. (2018) released under Apache 2.0 License. We list all the hyperparameters we used for training in Table 2, and provide all the code for our method, and the experiments at GitHub repository.

Table 2: Parameters for experiments results in Sec. 5 and App. D.

	Numerical Stage 1	Numerical Stage 2	Balls CNN	Balls Stage 1	Balls Stage 2
Sparsity level ϵ		0.01			0.01
Optimizer	Adam	ExtraAdam	Adam	Adam	ExtraAdam
Primal optimizer learning rate	1e-4	1e-4	1e-3	1e-3	1e-4
Dual optimizer learning rate		5e-5			5e-5
Batch size	6144	6144	256	256	256
Group number J	$5n$	$5n$	$2^{\mathbf{b}}$	$2^{\mathbf{b}}$	$2^{\mathbf{b}}$
# Seeds	5	5	5	5	5
# Iterations	5000	80000	20000	10000	5000

D Full experimental results

D.1 Metrics

Metric for identifiability up to affine transformation R^2 metric is the *coefficient of determination* of the linear regression model with ground truth latents $\mathbf{Z}$ as explanatory and our learned representations $\mathbf{g}(\mathbf{X})$ as response. R^2 normally ranges from 0 to 1, and the closer it is to 1, the better its affine identifiability.

Metric for identifiability up to permutation and scaling. MCC 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 latent 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 $\text{Corr}_{\pi}^{n\times n}$. MCC can take value in $[0,1]$, and the closer it is to 1, the better element-wise affine identifiability.

D.2 Numerical experiments

We first show the full version (including standard deviation) of Table 1.

Table 3: Results for the numerical experiments (Average ± Standard Deviation). The bold font indicates which parameters are varying in each block of rows. $\mathbf{R}^2$(Stage 1) shows the identifiability up to AT after Stage 1 (Thm. 3.7); **MCC**(Stage 1) reflects the neccesarity of sparsity to achieve PTS identifiability; **MCC**(Stage 2) presents the identifiability up to PTS after Stage 2 (Thm. 3.9).

n	k	m	ρ	δ	θ	$\mathbf{R}^2$ ↑(Stage1)	**MCC** (Stage1)	**MCC** (Stage1)	**MCC** ↑(VaDE)
5	1	10	50%	0	0	0.93±0.02	0.75±0.03	0.96±0.01 ‘	0.45±0.05
10	1	10	50%	0	0	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
20	1	10	50%	0	0	0.93±0.02	0.46±0.01	0.92±0.05	0.23±0.15
40	1	10	50%	0	0	0.94±0.01	0.37±0.01	0.94±0.03	0.25±0.11
10	**0**	10	50%	0	0	0.94±0.02	0.51±0.11	0.97±0.01	0.33±0.05
10	**1**	10	50%	0	0	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
10	**2**	10	50%	0	0	0.93±0.01	0.58±0.01	0.95±0.01	0.31±0.03
10	**3**	10	50%	0	0	0.92±0.03	0.56±0.02	0.91±0.04	0.27±0.03
10	1	**3**	50%	0	0	0.99±0.01	0.50±0.07	0.96±0.03	0.33±0.04
10	1	**10**	50%	0	0	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
10	1	**20**	50%	0	0	0.88±0.02	0.50±0.07	0.93±0.01	0.35±0.07
10	1	10	**1 var**	0	0	0.75±0.03	0.46±0.02	0.45±0.04	0.28±0.03
10	1	10	**50%**	0	0	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
10	1	10	**75%**	0	0	0.95±0.02	0.57±0.03	0.93±0.04	0.31±0.04
10	1	10	50%	**0**	0	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
10	1	10	50%	**1**	0	0.95±0.01	0.59±0.01	0.88±0.04	0.30±0.02
10	1	10	50%	**2**	0	0.96±0.01	0.59±0.03	0.61±0.04	0.27±0.03
10	1	10	50%	**3**	0	0.96±0.01	0.58±0.01	0.49±0.02	0.28±0.03
10	1	10	50%	0	**0**	0.94±0.02	0.56±0.03	0.97±0.01	0.32±0.03
10	1	10	50%	0	**15**	0.93±0.02	0.58±0.03	0.83±0.02	0.32±0.03
10	1	10	50%	0	**30**	0.93±0.02	0.57±0.03	0.60±0.02	0.31±0.04
10	1	10	50%	0	**45**	0.93±0.02	0.57±0.03	0.62±0.01	0.32±0.03

We show the Pearson Correlation matrix $\text{Corr}_{\pi}^{n\times n}$ with the permutation π between ground truth latent variables $\mathbf{Z}$ and the learned representations $\hat{\mathbf{Z}} = \mathbf{g}(\mathbf{X})$. 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.

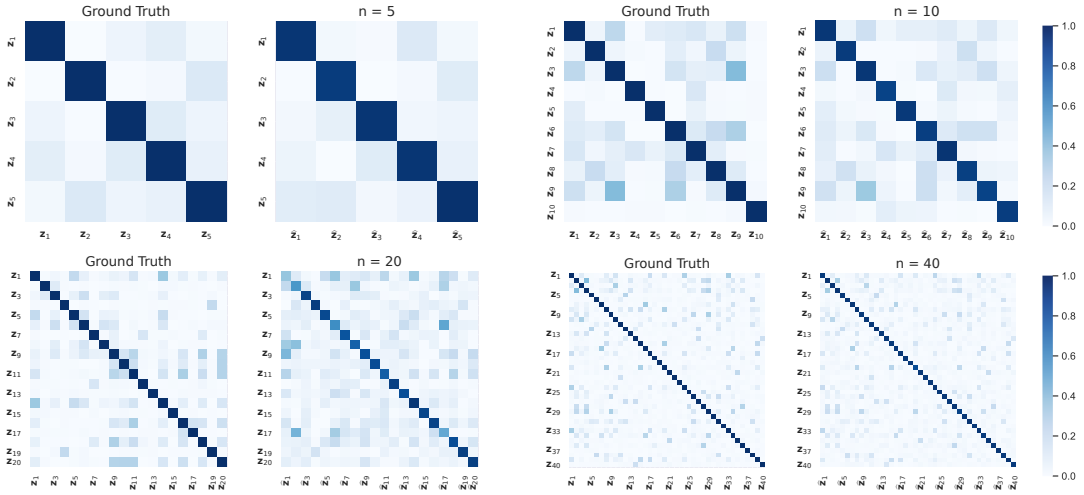

Figure 4: Ablation study on increasing the latent dimension n from 5 to 40 and fixing $k = 1$, $m = 10$, $\rho = 50\%$, $\delta = 0.0\sigma$, $\theta = 0$. The case with a higher n is more complicated to learn the latent variables.

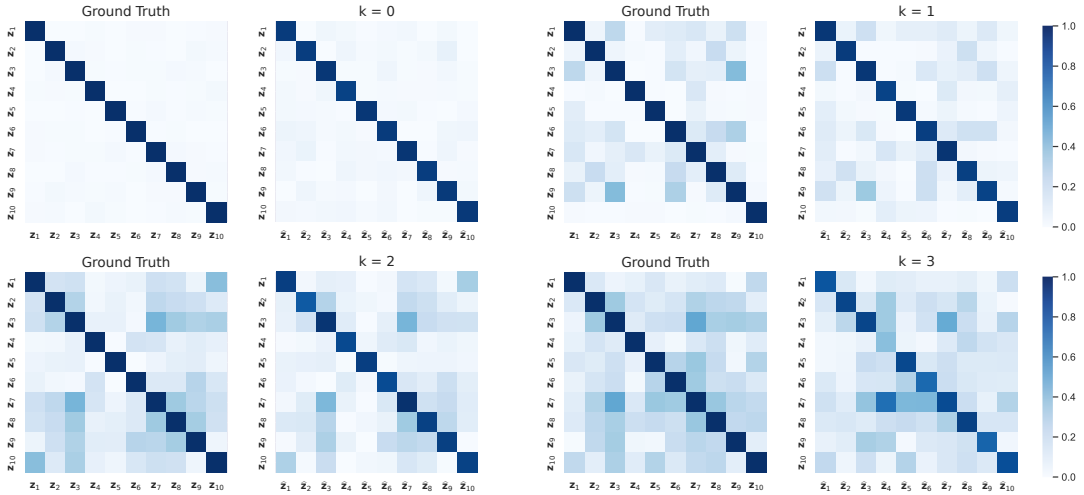

Figure 5: Ablation study on increasing the density of causal graphs k from 0 to 3 and fixing $n = 10$, $m = 10$, $\rho = 50\%$, $\delta = 0.0\sigma$, $\theta = 0$. In the case with a denser graph, it is more complicated to learn the latent variables.

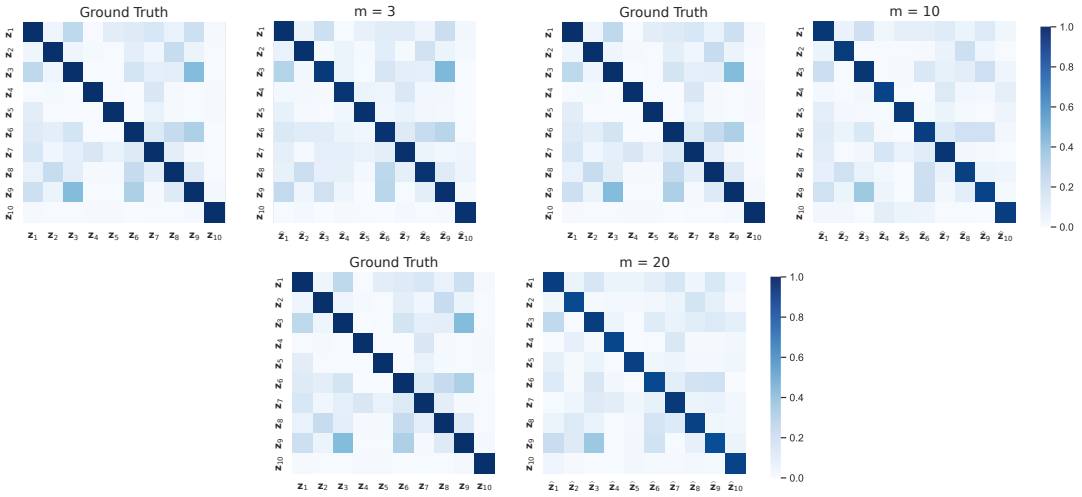

Figure 6: Ablation study on increasing the number of Leaky-ReLU layers $(m-1)$ from 3 to 20 and fixing $n = 10$, $k = 1$, $\rho = 50\%$, $\delta = 0.0\sigma$, $\theta = 0$. The case with a higher n is more complicated to learn the latent variables. The case with larger m is more complicated to learn the latent variables due to a greater extent of nonlinearity.

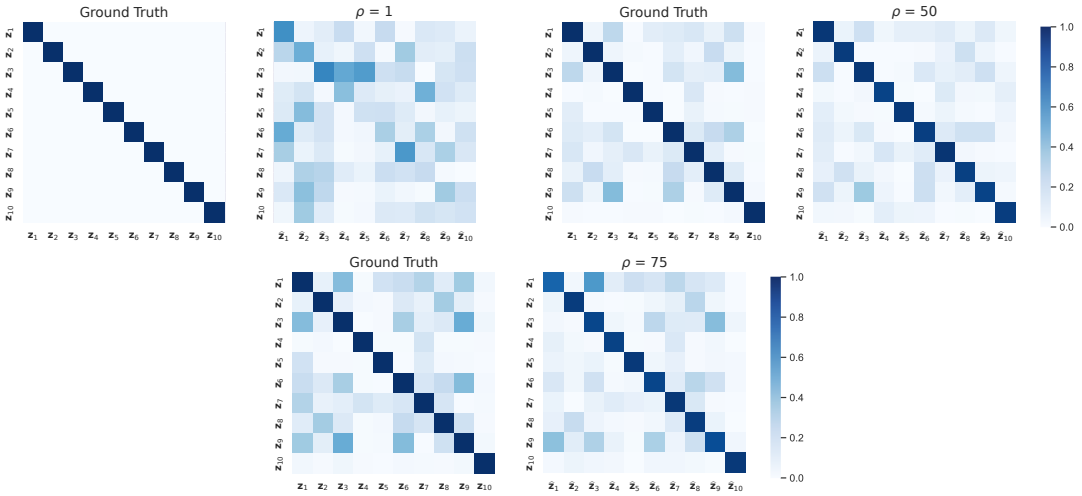

Figure 7: Piecewise linear mixing function with linear causal relation and Gaussian noise: ablation study on increasing the ratio of active (unmasked) variables ρ from 1 variable only to 75% and fixing $n = 10$, $k = 1$, $m = 10$, $\delta = 0.0\sigma$, $\theta = 0$. Learning the latent variables is more complicated in the case with a larger portion of active variables.

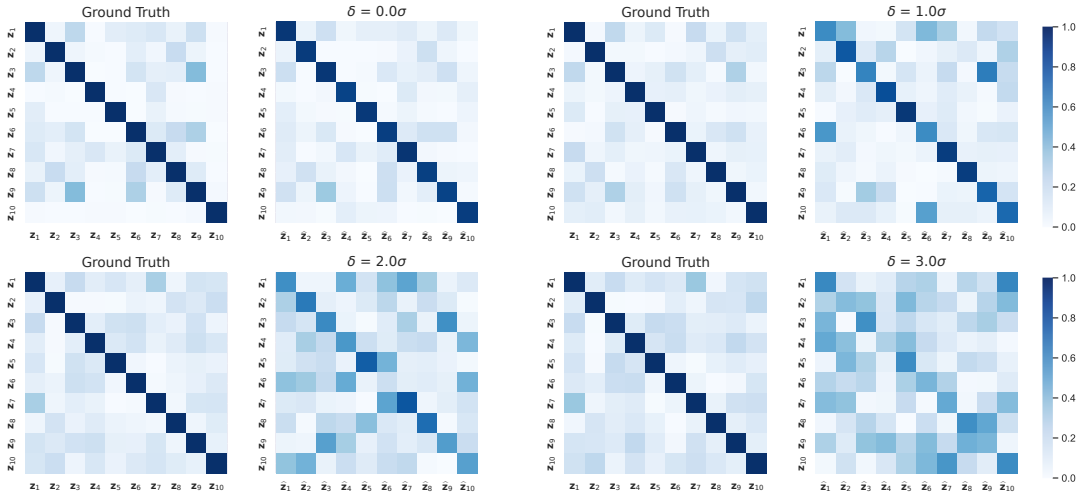

Figure 8: Ablation study on increasing the distance between masked value and mean of latents from 0.0σ to 3.0σ, where σ is the standard deviation of latents, and fixing $n = 10$, $k = 1$, $m = 10$, $\rho = 50\%$, $\theta = 0$. Disentangling the latent variables is more complicated in the case of a larger deviation from zero translation.

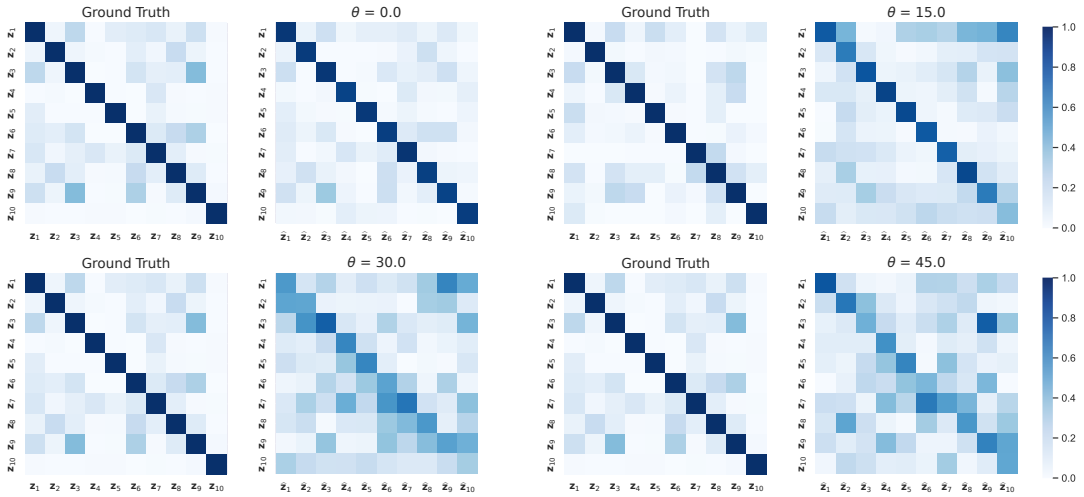

Figure 9: Ablation study on increasing the rotation of common basis from 0.0° to 45.0°, and fixing $n = 10$, $k = 1$, $m = 10$, $\rho = 50\%$, $\delta = 0.0\sigma$. Disentangling the latent variables is more complicated in the case of a larger deviation from the standard basis.

D.3 Test on independent variables

To exclude the natural dependency between latent variables that potentially increases the value of metric MCC that might cause a false positive measurement, we provide additional experiment results with independent latent variables using five random seeds in Table 9. As we can see, it keeps the consistent performance as those in Table 1, where we repeat the results in the right column of Table 9 for easier comparison. This provides evidence that the MCC obtained by our method does not come from the inherent dependencies among latents.

Table 4: Results for the numerical experiments (Stage 2) tested on independent data. The bold font indicates which parameters are varying in each block of rows.

n	k	m	ρ	δ	θ	**MCC** Independent	**MCC** in Tab. 1
5	1	10	50 %	0	0	0.96±0.01	0.96±0.01
10	1	10	50 %	0	0	0.97±0.01	0.97±0.01
20	1	10	50 %	0	0	0.93±0.05	0.92±0.05
40	1	10	50 %	0	0	0.94±0.03	0.94±0.03
10	**0**	10	50 %	0	0	0.97±0.01	0.97±0.01
10	**1**	10	50 %	0	0	0.97±0.01	0.97±0.01
10	**2**	10	50 %	0	0	0.97±0.01	0.95±0.01
10	**3**	10	50 %	0	0	0.91±0.04	0.91±0.04
10	1	**3**	50 %	0	0	0.96±0.03	0.96±0.03
10	1	**10**	50 %	0	0	0.97±0.01	0.97±0.01
10	1	**20**	50 %	0	0	0.94±0.02	0.93±0.01
10	1	10	**1var**	0	0	0.46 ±0.04	0.45±0.04
10	1	10	**50 %**	0	0	0.97±0.01	0.97±0.01
10	1	10	**75 %**	0	0	0.93 ±0.03	0.93±0.04
10	1	10	50 %	**0**	0	0.97±0.01	0.97±0.01
10	1	10	50 %	**1**	0	0.87±0.04	0.88±0.04
10	1	10	50 %	**2**	0	0.61±0.04	0.61±0.04
10	1	10	50 %	**3**	0	0.49±0.02	0.49±0.02
10	1	10	50 %	0	**0**	0.96±0.03	0.97±0.01
10	1	10	50 %	0	**15**	0.84 ±0.01	0.83±0.02
10	1	10	50 %	0	**30**	0.61±0.01	0.60±0.02
10	1	10	50 %	0	**45**	0.63±0.01	0.62±0.01

D.4 Empirical study on scalability

As shown in Tab. 2, Stage 1 only need 5000 epoch whereas Stage 2 need 80000. Consequently, the second stage of optimization constitutes the primary computational bottleneck of our method. To quantify scalability, we report both the number of epochs needed for convergence and the average execution time per 1000 epochs in seconds for the second stage. We can observe an exponential complexity increase in the table:

Table 5: Number of epochs for convergence and training time

n	5	10	20	40	50
Epoch	2,500	20,000	125,000	300,000	800,000
Execution time per 1000 epoch in Stage2 (s)	13.4	20.2	45.6	145.4	219.7

In particular we report results for $n = 50$ and limited results for $n = 100$. For comparisons, few works in the CRL literature are able to scale reliably to this number of latent variables, and most of the methods typically focus on 10-20 variables. For $n = 50$, we got $MCC = 0.93 \pm 0.03$ averaged over 3 random seeds. For $n = 100$, full convergence of the second stage was not reached. However, after 15,000 epochs of the first stage, we observed an average $R^2 = 0.82 \pm 0.13$, indicating a meaningful affine identifiability which validates Theorem 3.7.

D.5 Sensitivity Analysis of Hyper-parameters

We performed a sensitivity analysis on two hyperparameters: the sparsity level ϵ in Equation (5), which controls the sparsity level, and the learning rate. Bold options in the tables are the ones we used in the main paper,

where $\epsilon = 1e-2$ is inspired by Xu et al. (2024), which shows that the results are not sensitive to ϵ, but they are sensitive to the learning rate.

Table 6: Sensitivity analysis of sparsity level ϵ in Eq. (5): Average MCCs over 5 random seeds (Setup: $n = 10$, $k = 1$, $m = 10$, $\rho = 50$, $\delta = 0$, $\theta = 0$)

ϵ	1000	100	10	1	1e-1
MCC	0.50±0.01	0.50±0.01	0.50±0.01	0.50±0.01	0.90±0.08
ϵ	**1e-2**	1e-3	1e-4	1e-5	
MCC	**0.97±0.01**	0.94±0.03	0.90±0.07	0.88±0.07	

The final hyperparameters we used in the experiments are chosen based on the MCC and R^2 computed with the ground truth variables. If we did not know the ground truth, tuning the parameters would be more difficult, but in line with other DL approaches. For example, we can choose the learning rate as usual in DL by observing the behavior of the loss, while for the sparsity level ϵ we can choose it based on potential violations of our assumptions in the reconstructed variables, e.g., if the reconstructed variables are non-Gaussian.

Table 7: Sensitivity analysis of learning rate: Average MCCs over 5 random seeds

Learning rate	1e-2	1e-3	**1e-4**	1e-5
MCC	0.25±0.03	0.53±0.08	**0.97±0.01**	0.92±0.02

D.6 Ablation study on sparsity constraint and comparison with close works

We ran the experiments on all settings without a sparsity constraint in the second stage. As summarized Appendix D.6, the significant drop in MCC shows the important role that sparsity plays in the disentanglement process.

Additionally, we compared two works (Xu et al., 2024; Kivva et al., 2022) which are closely related to our work, where Xu et al. (2024) serves as an oracle since it provides both group information and degenerate dimension, which is learned in our method. Xu et al. (2024) can be seen as a special case which can be explained in our theorem and since it provides stronger information so we do not aim to outperform it. Kivva et al. (2022) only works for non-degenerate cases, which shows the importance of our method to solve the unexplored problem in the degenerate latent space.

Table 8: Ablation study without sparsity constraints and baselines: Average MCCs over 5 random seeds. Ours = proposed method, WO = without sparsity constraints, Oracle (Xu et al., 2024), VaDE (Kivva et al., 2022)

n	5	10	20	40
Ours	0.96±0.01	0.97±0.01	0.92±0.05	0.94±0.03
WO	0.57±0.02	0.57±0.04	0.44±0.01	0.36±0.01
Oracle	0.91±0.05	0.98±0.00	0.96±0.01	0.94±0.02
VaDE	0.45±0.05	0.32±0.03	0.23±0.15	0.25±0.11

(a) Varying n

k	0	1	2	3
Ours	0.97±0.01	0.97±0.01	0.95±0.01	0.91±0.04
WO	0.56±0.04	0.57±0.04	0.54±0.01	0.55±0.02
Oracle	0.98±0.01	0.98±0.00	0.95±0.04	0.95±0.03
VaDE	0.33±0.05	0.32±0.03	0.31±0.03	0.27±0.03

(b) Varying k

m	3	10	20
Ours	0.96±0.03	0.97±0.01	0.93±0.01
WO	0.55±0.01	0.57±0.04	0.56±0.02
Oracle	0.99±0.00	0.98±0.00	0.95±0.01
VaDE	0.33±0.04	0.32±0.03	0.35±0.07

(c) Varying m

ρ	1	50	75
Ours	0.45±0.04	0.97±0.01	0.93±0.04
WO	0.42±0.03	0.57±0.04	0.56±0.03
Oracle	0.86±0.04	0.98±0.00	0.98±0.01
VaDE	0.28±0.03	0.32±0.03	0.31±0.04

(d) Varying ρ

δ	0	1	2	3
Ours	0.97±0.01	0.88±0.04	0.61±0.04	0.49±0.02
WO	0.57±0.04	0.54±0.02	0.55±0.04	0.55±0.03
Oracle	0.98±0.00	0.73±0.16	0.41±0.04	0.38±0.02
VaDE	0.32±0.03	0.30±0.02	0.27±0.03	0.28±0.03

(e) Varying δ

θ	0	15	30	45
Ours	0.97±0.01	0.83±0.02	0.60±0.02	0.62±0.01
WO	0.57±0.04	0.56±0.03	0.58±0.04	0.56±0.02
Oracle	0.98±0.00	0.85±0.01	0.62±0.02	0.63±0.01
VaDE	0.32±0.03	0.32±0.03	0.31±0.04	0.32±0.03

(f) Varying θ

D.7 Comparison with Baseline VaDE without misspecification

We provide results where we evaluate our method in a setting that is fair to the baseline VaDE, where we generate the latent with a non-degenerate GMM. In this setting, we can only compare the identifiability up to AT measured by metric R^2, but we cannot compare for PS for any of the two methods, because our method needs degeneracy, and VaDE needs an auxiliary variable and pairwise conditional independence. As shown below, our method outperforms VaDE for all settings.

Table 9: Results for the numerical experiments (Stage 1) on nondegenerate GMM data. The bold font indicates which parameters are varying in each block of rows.

n	k	m	ρ	δ	θ	**MCC** of VaDE	**MCC** of Our Method
5	1	10	50 %	0	0	0.66±0.09	0.94±0.01
10	1	10	50 %	0	0	0.78±0.05	0.94±0.01
20	1	10	50 %	0	0	0.86±0.04	0.93±0.01
40	1	10	50 %	0	0	0.84±0.07	0.93±0.01
10	**0**	10	50 %	0	0	0.86±0.05	0.95±0.01
10	**1**	10	50 %	0	0	0.78±0.05	0.94±0.01
10	**2**	10	50 %	0	0	0.74±0.12	0.94±0.01
10	**3**	10	50 %	0	0	0.69±0.11	0.93±0.02
10	1	**3**	50 %	0	0	0.78±0.10	0.99±0.00
10	1	**10**	50 %	0	0	0.78±0.05	0.94±0.01
10	1	**20**	50 %	0	0	0.70±0.10	0.89±0.01
10	1	10	**1var**	0	0	0.77 ±0.05	0.94±0.01
10	1	10	**50 %**	0	0	0.78±0.05	0.94±0.01
10	1	10	**75 %**	0	0	0.75 ±0.08	0.94±0.02
10	1	10	50 %	0	**0**	0.78±0.05	0.94±0.01
10	1	10	50 %	0	**15**	0.77 ±0.01	0.94±0.01
10	1	10	50 %	0	**30**	0.78±0.02	0.94±0.02
10	1	10	50 %	0	**45**	0.78±0.02	0.94±0.01

D.8 Over-parameterization of latent dimension n

In our previous analysis, the latent dimension n is treated as fixed. To the best of our knowledge, prior work typically assumes n is known in theoretical analysis, but then relies on over-parameterization strategies in empirical analysis. We also consider an over-parameterized choice and show empirically that our method performs well under these settings. While a formal proof of identifiability for remains open, our experimental findings suggest that moderate over-specification does not undermine identifiability guarantees in practice. We use our basic setup: $n = 10$, $k = 1$, $m = 10$, $\rho = 50$, $\delta = 0$, $\theta = 0$, varying estimation $\hat{n} \in \{12, 14, 16, 18, 20\}$.

Table 10: Average R2 and MCC over 5 random seeds

$\hat{n}$	12(n+2)	14(n+4)	16(n+6)	18(n+8)	20(n+10)
R2	0.91	0.89	0.88	0.88	0.83
MCC	0.96	0.96	0.93	0.96	0.96

D.9 Causal graph learning analysis

The aim of causal representation learning (CRL) is to recover latent causal variables from high-dimensional observations. Once these causal variables are recovered, one can learn the equivalence class of the causal graph on these variables, e.g. by applying a standard causal discovery method on top of these recovered variables.

We provide results in which we apply a standard causal discovery method, PC (Spirtes et al., 2000), on our representations. We consider the same setting as in Table 3. We generate 5000 samples for each graph. We use the implementation from the pcalg package and use partial correlation tests with a significance threshold $\alpha = 0.01$.

In Tab. 11, we report the average SHD on the CPDAG learned over our $\hat{\mathbf{z}}$ w.r.t. the oracle CPDAG (the ground truth CPDAG with oracle tests). We report the average SHD on the CPDAG learned over the ground truth variables $\mathbf{z}$ w.r.t. the oracle CPDAG. This is the best result we could obtain with a perfect reconstruction of the ground truth variables. We report the delta between them as an alternative measure of the accuracy of our reconstruction of the causal variables.

Table 11: The average SHD on the CPDAG learned over our $\hat{\mathbf{z}}$ and $\mathbf{z}$ w.r.t. the oracle CPDAG.

n	k	m	ρ	δ	θ	**SHD** ($\mathbf{z}$)	**SHD** ($\hat{\mathbf{z}}$)	**ΔSHD**
5	1	10	50 %	0	0	1.4±1.67	3.8±1.79	2.4
10	1	10	50 %	0	0	1.8±1.79	4.2±1.10	2.4
20	1	10	50 %	0	0	7.8±2.86	22±12.90	14.2
40	1	10	50 %	0	0	13.6±4.04	39.8±17.75	26.2
10	**1**	10	50 %	0	0	1.8±1.79	4.2±1.10	2.4
10	**2**	10	50 %	0	0	13.2±2.17	16±2.64	2.8
10	**3**	10	50 %	0	0	25±1.87	26.6±3.85	1.6
10	1	**3**	50 %	0	0	1.8±1.79	3±1.22	1.2
10	1	**10**	50 %	0	0	1.8±1.79	4.2±1.10	2.4
10	1	**20**	50 %	0	0	1.8±1.79	8.8±6.05	7
10	1	10	**1var**	0	0	1.8±1.79	21.4±3.05	19.6
10	1	10	**50 %**	0	0	1.8±1.79	4.2±1.10	2.4
10	1	10	**75 %**	0	0	1.8±1.79	7.4±4.04	5.6
10	1	10	50 %	**0**	0	1.8±1.79	4.2±1.10	2.4
10	1	10	50 %	**1**	0	1.8±1.79	16.8±5.12	15
10	1	10	50 %	**2**	0	1.8±1.79	21.8±2.59	20
10	1	10	50 %	**3**	0	1.8±1.79	22.2±2.68	20.4
10	1	10	50 %	0	**0**	1.8±1.79	4.2±1.10	2.4
10	1	10	50 %	0	**15**	2.2±1.24	18±2.83	15.8
10	1	10	50 %	0	**30**	2.6±2.41	20.8±1.64	18.2
10	1	10	50 %	0	**45**	4.5±0.71	21.4±2.30	16.9

These results show that our learned causal graph is similar to the learned graphs on the ground truth variables.

D.10 Image dataset: Multiple Balls

Data generation process We use PyGame to render images with size $64 \times 64 \times 3$ as shown in Figure 10. The number of balls $\mathbf{b}$ varies from 2 to 6. In this setting, the latent variables we consider are the (x, y) coordinates of each ball with support $[0, 1]^2$. We sample the (x, y) coordinates of each ball $i \in [\mathbf{b}]$ independently from a truncated 2-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ bounded within the square $(0.05, 0.95)^2$ with randomized $\boldsymbol{\mu}_i$ and fixed $\boldsymbol{\Sigma}_i$: $\boldsymbol{\mu}_i \sim \text{Unif}(0.3, 0.7)^2$, and $\boldsymbol{\Sigma}_i = ((0.01, 0.00)(0.00, 0.01))$. To avoid the high ratio of occultation, we use rejection sampling to sample $\boldsymbol{\mu}_i$ to ensure the Euclidean distance between any pair of $\boldsymbol{\mu}_i$ values is at least 0.2. To model the degeneracy in this setting, we set each ball at a fixed position $\boldsymbol{\mu}_i$ with probability $p = 0.1$ by sampling a binomial variable to indicate the degeneracy status.

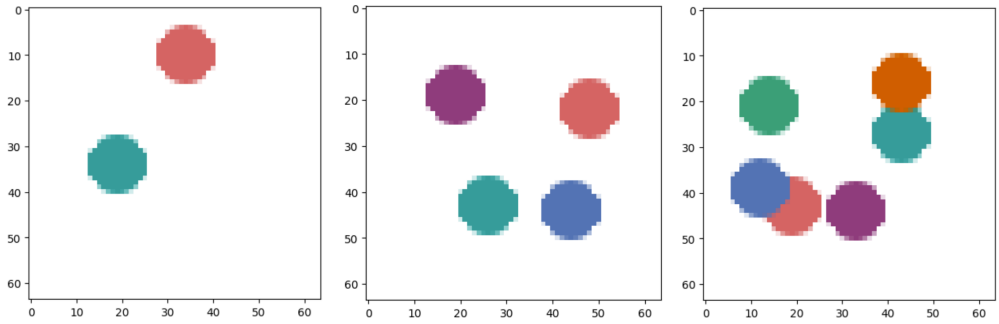

Figure 10: Example images of multiple ball dataset, modified based on the rendering code provided by Ahuja et al. (2022). Balls can move freely inside the frame, and each ball's coordinates can be fixed randomly to a constant value, which causes their latent dimensions to be degenerate; Varying the number of balls from 2 to 6.

CNN architecture Before we train the model described in Sec. 4, we pre-train a CNN model to reduce the input image dimension from $64 \times 64 \times 3$ to a vector of size 128×1. To control the information loss, we use a symmetric convolutional autoencoder to reconstruct the images. All convolutional layers use kernel size 3 with

stride 1 and padding 1, and a ReLU activation follows each convolutional or transposed convolutional layer, except for the final output layer which uses a Sigmoid activation function. We list the concrete network structure in Appendix D.10.

Table 12: Architecture of the convolutional autoencoder used in Sec. 5 (input size: $3 \times 64 \times 64$).

Layer Type	**Configuration**	**Output Shape**
Input	RGB image	(3, 64, 64)
	Encoder	
Conv2D + ReLU	$3 \to 32$, kernel=3, stride=1, pad=1	(32, 64, 64)
MaxPool2D	2×2	(32, 32, 32)
Conv2D + ReLU	$32 \to 64$, kernel=3, stride=1, pad=1	(64, 32, 32)
MaxPool2D	2×2	(64, 16, 16)
Conv2D + ReLU	$64 \to 64$, kernel=3, stride=1, pad=1	(64, 16, 16)
MaxPool2D	2×2	(64, 8, 8)
Conv2D + ReLU	$64 \to 128$, kernel=3, stride=1, pad=1	(128, 8, 8)
MaxPool2D	2×2	(128, 4, 4)
Conv2D + ReLU	$128 \to 128$, kernel=3, stride=1, pad=1	(128, 4, 4)
MaxPool2D	4×4	(128, 1, 1)
	Decoder	
ConvTranspose2D + ReLU	$128 \to 128$, kernel=2, stride=2	(128, 2, 2)
ConvTranspose2D + ReLU	$128 \to 64$, kernel=2, stride=2	(64, 4, 4)
ConvTranspose2D + ReLU	$64 \to 64$, kernel=2, stride=2	(64, 8, 8)
ConvTranspose2D + ReLU	$64 \to 32$, kernel=2, stride=2	(32, 16, 16)
ConvTranspose2D + Sigmoid	$32 \to 3$, kernel=4, stride=4	(3, 64, 64)

To evaluate the performance, we perform linear regression from each representation separately to the corresponding ground-truth ball (x,y)-position. For $\mathbf{b} = 2, 4, 6$, we provide the average R^2 over 5 random seeds for each ball in Tables 13 to 15. The results show that for the settings $\mathbf{b} = 2, 4$, the position of each ball can be well recovered through pairs of representations, which is consistent with our theoretical results. However, as shown in the example images, when $\mathbf{b} = 6$, even though we use rejection sampling to ensure the distance between the mean values of any pair of balls is at least 0.2, due to the limited space of the frame, a high portion of occultation cannot be avoided, which also cause a performance drop as shown in Table 15.

Table 13: Results for Multiple Balls with $\mathbf{b} = 2$: averaged R^2 of linear regression between ground truth ball positions and learned representations.

	Ball 1	Ball 2
MCC	0.958 ± 0.052	0.914 ± 0.148

Table 14: Results for Multiple Balls with $\mathbf{b} = 4$: averaged R^2 of linear regression between ground truth ball positions and learned representations.

	Ball 1	Ball 2	Ball 3	Ball 4
MCC	0.953 ± 0.033	0.853 ± 0.194	0.932 ± 0.041	0.880 ± 0.137

Table 15: Results for Multiple Balls with $\mathbf{b} = 6$: averaged R^2 of linear regression between ground truth ball positions and learned representations.

	Ball 1	Ball 2	Ball 3
MCC	0.743 ± 0.106	0.105 ± 0.046	0.293 ± 0.229
	Ball 4	Ball 5	Ball 6
MCC	0.407 ± 0.375	0.474 ± 0.183	0.595 ± 0.241

D.11 Identify $x-$ and $y-$positions with fined-grained masks

we can define finer-grained sparsity patterns. In particular, our method can disentangle the $x-$ and $y-$positions of each ball, if the dataset allows freezing $x-$ and $y-$positions individually.

We ran experiments on 5 random seeds with 2 balls (in total 4 latent variables), and got an average MCC $= 0.97 \pm 0.03$. In Appendix D.11, we show that the R^2 of linear regression between ground truth ball positions $\mathbf{b}_1^x$, $\mathbf{b}_1^y$, $\mathbf{b}_2^x$, $\mathbf{b}_2^y$, and learned representations $\hat{\mathbf{z}}_1$,...,$\hat{\mathbf{z}}_4$. As we can see, both $x-$ and $y-$ can be disentangled well.

Table 16: Example results of $\hat{z}_i$ assignments across blocks

	$\mathbf{b}_1^x$	$\mathbf{b}_1^y$	$\mathbf{b}_2^x$	$\mathbf{b}_2^y$
$\hat{\mathbf{z}}_1$	0.03	0.00	**0.98**	0.00
$\hat{\mathbf{z}}_2$	**0.95**	0.01	0.01	0.00
$\hat{\mathbf{z}}_3$	0.00	0.00	0.00	**0.97**
$\hat{\mathbf{z}}_4$	0.02	**0.99**	0.00	0.01

D.12 Compute information

The experiments in Section 5 are implemented on NVIDIA A100 GPU. The experiments in Section 5 on multiple balls are implemented on NVIDIA A40 GPU. Each job used 1 GPU per node and 20 GB of memory

E Discussion on Assumptions

In this section, we assess the applicability and robustness of our assumptions and discuss to what extent they hold or may fail in practical scenarios. We begin with empirical evidence on violations of Gaussianity and piecewise affine mixing, followed by a closer examination of each theoretical assumption, supported by proofs and illustrative examples.

E.1 Parametric assumptions: Gaussian latent and piecewise affine mixing

Mixture models capture a significantly broader range of latent structures than standard single-component (e.g., standard Gaussian) priors used in many prior works (e.g. VAE and β VAE). In fact, mixture models, especially GMMs, are commonly used to approximate real-world latent spaces in applications. Our use of a degenerate mixture is especially relevant in high-dimensional sparse regimes, where not all latent variables are active in every sample.

The piecewise affine mixing function assumption aligns with a broad class of practical transformations (e.g., ReLU networks). Moreover, it can approximate any diffeomorphism functions up to arbitrary precision given a sufficient number of pieces.

In addition, to test the robustness of our method, we report some results for violations of these assumptions in Table 17. We experiment with the violation of 1) the Gaussian assumption (replaced by Exponential and Gumbel), and 2) the piecewise affine (replace LeakyReLU by 1. smooth LeakyReLU 2. Sigmoid). Except for the Sigmoid setting, we did not observe a performance drop, indicating the robustness of our method under certain violations of our assumptions.

Table 17: Results for the numerical experiments (Stage 2) on misspecified settings.

n	k	m	ρ	MCC			
				Exponential	**Gumbel**	**Smooth LeakyReLU**	**Sigmoid**
5	1	10	50 %	0.92±0.06	0.95±0.01	0.97±0.01	0.36±0.02
10	1	10	50 %	0.96±0.01	0.97±0.01	0.98±0.01	0.32±0.06
20	1	10	50 %	0.94±0.02	0.95±0.02	0.94±0.09	0.29±0.04
10	**0**	10	50 %	0.95±0.01	0.97±0.01	0.95±0.08	0.37±0.06
10	**1**	10	50 %	0.96±0.01	0.97±0.01	0.98±0.01	0.32±0.06
10	**2**	10	50 %	0.93±0.04	0.96±0.02	0.97±0.01	0.29±0.06
10	**3**	10	50 %	0.91±0.04	0.91±0.06	0.93±0.03	0.36±0.09
10	1	**3**	50 %	0.98±0.01	0.99±0.01	0.99±0.00	0.60±0.03
10	1	**10**	50 %	0.96±0.01	0.97±0.01	0.98±0.01	0.32±0.06
10	1	**20**	50 %	0.91±0.04	0.94±0.02	0.97±0.01	0.30±0.08
10	1	10	**1var**	0.66±0.08	0.59±0.07	0.70 ±0.07	0.36±0.01
10	1	10	**50 %**	0.96±0.01	0.97±0.01	0.98±0.01	0.32±0.06
10	1	10	**75 %**	0.92±0.04	0.94±0.04	0.95 ±0.03	0.32±0.07

E.2 Assumption 3.4: Genericity of pdGMMs

In this section, we clarify the genericity of Assumption 3.4. We begin by providing a proof that the set of parameters violating this assumption is a measure-zero subset of the parameter space under consideration, i.e., even when we consider only covariance matrices that are degenerate. Then, we introduce an example in two-dimensional space to support and illustrate the intuition behind it. We also extend this example to higher dimensions and conduct an ablation study to show the necessity of this assumption. Finally, we discuss the applicability of this assumption in practice.

Assumption 3.4 (Genericity of pdGMMs)**.** Consider a pdGMM in reduced form $\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. Let $\mathcal{J}_k := \{j \in [J] : rank(\boldsymbol{\Sigma}_j) = k)\}$ be the indices of the components with rank k. For any $k \leq n$ and any subset $\mathcal{J}_k^0 \subseteq \mathcal{J}_k$, suppose that the intersection of the supports of the components $\mathcal{Z}_j$ is non-empty, i.e., $\cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j \neq \varnothing$. We assume there exists at least one point $\mathbf{z}_0 \in \cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j$ s.t. the Mahalanobis distance [4] of $\mathbf{z}_0$ to the distributions of any two distinct Gaussian components $i, j \in \mathcal{J}_k^0$ is not the same.

Lemma E.1. *Consider a pdGMM in reduced form $\mathbf{Z} \sim \sum_{j=1}^{J} \lambda_j N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$. The set of parameters $(\lambda_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), J \in [J]$ violating Assumption 3.4 is measure-zero in the parameter space under consideration, i.e., when we consider $\boldsymbol{\Sigma}_j$ to be degenerate covariance matrices.*

Proof. A violation of Assumption 3.4 means for some $0 < k \leq n$, there exists an subset $\mathcal{J}_k^0 \in \mathcal{J}_k$, such that all points $\mathbf{z}_0 \in \cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j$, $\exists\, i, j \in \mathcal{J}_k^0$ such that the Mahalonobis distance are the same, i.e., $\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_i)} = \sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_j)}$.

To constrain the parameter space of $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Sigma}_j$ to the ones with rank k, we parameterize $\boldsymbol{\Sigma}_i = A_i A_i^T$ and $\boldsymbol{\Sigma}_j = A_j A_j^T$ via full-column-rank decompositions with $\text{rank}(A_i) = \text{rank}(A_j) = k$. Therefore, we have

$$\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_i)} = ||(A_i^T A_i)^{-1} A_i^T (\mathbf{z}_0 - \boldsymbol{\mu}_i)|| \tag{161}$$

$$\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_j)} = ||(A_j^T A_j)^{-1} A_j^T (\mathbf{z}_0 - \boldsymbol{\mu}_j)||. \tag{162}$$

In this way, we can define the subset of parameter space that violates the assumption for each pair of components i and j.

$$\mathcal{V}_{ij}^{\mathcal{J}_k^0} := \{(A_i, A_j, \boldsymbol{\mu}_i, \boldsymbol{\mu}_j) : ||(A_i^T A_i)^{-1} A_i^T (\mathbf{z}_0 - \boldsymbol{\mu}_i)|| = ||(A_j^T A_j)^{-1} A_j^T (\mathbf{z}_0 - \boldsymbol{\mu}_j)|| \quad \forall\, \mathbf{z}_0 \in \cap_{j\in\mathcal{J}_k^0} \mathcal{Z}_j\} \tag{163}$$

[4] Mahalanobis distance of a point $\mathbf{z}_0$ from a distribution $N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ is $\sqrt{(\mathbf{z}_0 - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{z}_0 - \boldsymbol{\mu}_j)}$. When $\boldsymbol{\Sigma}_j$ is singular, we use the Moore-Penrose inverse for $\boldsymbol{\Sigma}_j^{-1}$

The total set of violating parameters for the pdGMM is $\mathcal{V} = \cup_{k=1}^{n} \cup_{\mathcal{J}_k^0 \subseteq \mathcal{J}_k} \cup_{i,j \in \mathcal{J}_k^0} \mathcal{V}_{ij}^{\mathcal{J}_k^0}$. Since a finite union of measure zero sets is still measure zero, if we can show that the measure of $\mathcal{V}_{ij}^{\mathcal{J}_k^0}$ is 0 for all pairs of components $i, j \in \mathcal{J}_k^0$, the proof is complete.

Let $\mathbf{f}_{ij}^{\mathbf{z}_0} := ||(A_i^T A_i)^{-1} A_i^T (\mathbf{z}_0 - \boldsymbol{\mu}_i)|| - ||(A_j^T A_j)^{-1} A_j^T (\mathbf{z}_0 - \boldsymbol{\mu}_j)||$. By Mityagin (2020), a real analytic function from $\mathbb{R}^m$ to $\mathbb{R}$ for $m \geq 1$ is either zero everywhere or non-zero almost everywhere. We can apply this result to our setting and show that $\mathbf{f}_{ij}^{\mathbf{z}_0}$, which is a real analytic function in $(A_i, A_j, \boldsymbol{\mu}_i, \boldsymbol{\mu}_j)$, is non-zero almost everywhere, which means that for each point $\mathbf{z}_0$ we can distinguish each pair of components. This means that once we can find one nonzero solution, we can conclude that the function is all nonzero almost everywhere. We construct the nonzero solution in this way: we can put the first term in the definition of $\mathbf{f}_{ij}^{\mathbf{z}_0}$ to zero by taking $\boldsymbol{\mu}_i = \mathbf{z}_0$ and using an arbitrary full-column matrix A_i. For the second term, we want to show that there exists a non-zero solution. We fix A_j and show that if $\mathbf{z}_0 - \boldsymbol{\mu}_j \notin Im((A_j^T A_j)^{-1} A_j^T)$, then $\mathbf{f}_{ij}^{\mathbf{z}_0}$ is nonzero.

Thus, we conclude based on the result by Mityagin (2020) that $\mathbf{f}_{ij}^{\mathbf{z}_0}$ is non-zero almost everywhere for any $\mathbf{z}_0$. This means that the set of violations $\mathcal{V}_{ij}^{\mathcal{J}_k^0}$ is a measure-zero set for any pair i, j and therefore its finite union $\mathcal{V}$ is a measure-zero set. □

Example E.2. Consider $n = 2$ and $J = 2$ with components $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Suppose

$$\boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}. \tag{164}$$

There are only two ways to decompose $\boldsymbol{\Sigma}_1$ into the form $A_1 A_1^T$: $A_1 = (1, 0)^T$ or $(-1, 0)^T$. Let $\boldsymbol{\Sigma}_2 = A_2 A_2^T$ where $A_2 = (a_1, a_2)^T$, and $\boldsymbol{\mu}_2 = (\mu_1, \mu_2)$. Consider $\boldsymbol{\varepsilon}_1 = \boldsymbol{\varepsilon} \in \mathbb{R}$, then $\boldsymbol{\varepsilon}_2 = \boldsymbol{\varepsilon}$ or $-\boldsymbol{\varepsilon}$. To violate Ass. 3.4, $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ can must satisfy:

$$\begin{cases} \boldsymbol{\varepsilon} = a_1 \boldsymbol{\varepsilon} + \mu_1 \\ 0 = a_2 \boldsymbol{\varepsilon} + \mu_2 \end{cases} \text{ or } \begin{cases} -\boldsymbol{\varepsilon} = a_1 \boldsymbol{\varepsilon} + \mu_1 \\ 0 = a_2 \boldsymbol{\varepsilon} + \mu_2 \end{cases} \text{ or } \begin{cases} \boldsymbol{\varepsilon} = -a_1 \boldsymbol{\varepsilon} + \mu_1 \\ 0 = -a_2 \boldsymbol{\varepsilon} + \mu_2 \end{cases} \text{ or } \begin{cases} -\boldsymbol{\varepsilon} = -a_1 \boldsymbol{\varepsilon} + \mu_1 \\ 0 = -a_2 \boldsymbol{\varepsilon} + \mu_2 \end{cases} \tag{165}$$

By solving the equations, we can conclude if and only if $(a_1 - 1)\mu_2 = a_2 \mu_1$ or $|(a_1 - 1)\mu_2| = |a_2 \mu_1|$, Ass. 3.4 is not satisfied, which is a measure-zero subset of parameter space.

We extend this example to dimensions $n = 2, 5, 10$ and show the results of our method on it in Appendix E.2. As you can see, these results are much worse than the comparable results in Table 1 for $n = 5, 10$, where the R^2 was ~ 0.93, and the MCC was ~ 0.96, showing why this assumption is needed.

Table 18: Average R^2 and MCC over 5 random seeds

n	2	5	10
R^2	0.87	0.70	0.64
MCC	0.92	0.63	0.47

In summary, we have shown that the violations of this assumption occur only on a measure-zero subset, supporting its genericity. The two-dimensional example further illustrates this result by explicitly characterizing the exceptional cases, which correspond to highly fine-tuned parameter choices. Together, these arguments support the statement that Assumption 3.4 is a mild and reasonable assumption, holding almost everywhere in real-world settings.

E.3 Assumption 3.6: Common basis and translation vector

This assumption states that the supports of all components should intersect and that a shared basis can represent them. This always holds for non-degenerate GMMs. For degenerate components, a violation can happen because the components are not intersecting, or if their individual bases do not form a global shared basis.

To show the necessity of Assumption 3.6, we provide an example that violates this assumption, showing that such cases only local affine identifiability can be achieved, whereas global identifiability fails.

Example E.3. Consider $n = 2$ with $\mathcal{Z}_1 := \text{span}\{\mathbf{e}^1\}$, $\mathcal{Z}_2 := \text{span}\{\mathbf{e}^2\}$ and $\mathcal{Z}_3 = \text{span}\{\mathbf{e}^1 + \mathbf{e}^2\}$. In this case, all three subspaces are one-dimensional and the issue is that the family of vectors $\{\mathbf{e}^1, \mathbf{e}^2, \mathbf{e}^1 + \mathbf{e}^2\}$ is linearly dependent and thus do not form a basis. One can construct a function $f : \cup_{j=1}^3 \mathcal{Z}_j \to \mathbb{R}$ that is affine on each $\mathcal{Z}_j$ individually, but not on the union:

$$f(\mathbf{z}) := \begin{cases} 0, \text{ if } \mathbf{z} \in \mathcal{Z}_1 \cup \mathcal{Z}_2 \\ \mathbf{z}_1, \text{ if } \mathbf{z} \in \mathcal{Z}_3 \end{cases} . \tag{166}$$

Clearly this function is not affine on $\cup_{j=1}^3 \mathcal{Z}_j$ since

$$f(\mathbf{e}^1 + \mathbf{e}^2) = 1 \neq 0 = f(\mathbf{e}^1) + f(\mathbf{e}^2) . \tag{167}$$

Intuitively, Assumption 3.6 requires that there exists a unique shared basis for all Gaussian components. We conduct an ablation study to show the necessity. To show a violation of this, we generate the data with a mixture of two different bases: a standard basis and one with rotations. As shown in Appendix E.3, these results are worse than the comparable results in Table 1, which shows why this assumption is needed for AT.

Table 19: Average R^2 over 5 random seeds

k	0	1	2	3
R^2	0.83	0.82	0.84	0.85

One concrete setting in which this assumption holds is the partially observable scenario in Xu et al. (2024). In that setting, there are n causal variables, but only a subset of them are "active" in each observation sample, e.g., they are captured in an image, while others are "inactive". The sets of "active" and "inactive" variables can change for each observation. For example, one could consider a camera that captures objects moving in and out of frame, or are occluded by another object. We consider the "inactive" objects as contributing to the observation with a specific constant value, i.e., their inputs to the mixing functions are constant (e.g. $\mathbf{0}$). In this setting, the intersection of the supports is the vector of all the constant values for each variable and since we are masking the inputs for the "inactive" variables, the individual basis of the supports will always be orthogonal and hence form a well-defined global shared basis, satisfying Assumption 3.6.

Other possible applications could include other high-dimensional applications that rely on sparse representations, such as language models (Cunningham et al., 2024; Gao et al., 2025; Marks et al., 2025), where we also have "active" and "inactive" concepts.

If it does not hold strictly in a setting, e.g., it only holds for a subset of components, for this subset we can still achieve the same result as Theorem 3.7, but for the other components we will learn a local linear identified representation individually as stated in Theorem 3.5, but not a global one.

E.4 Assumption 3.8a: Common Standard Basis

The only difference between the Common Standard Basis Assumption (Assumption 3.8 (a)) and Assumption 3.6 is the interception term, which can be canceled out easily by shifting the modeling by a constant in practice.

Table 1 already shows violations of Ass. 3.8a when δ (the masking value) and θ (the rotation of the standard basis) are non-zero, showcasing a decrease in performance when the assumption does not hold.

E.5 Assumption 3.8b: Sufficient Support Basis Index Variability

The Sufficient Support Basis Index Variability Assumption (Assumption 3.8 (b)), it is more likely to hold when the sparsity patterns across samples are sufficiently diverse. For example, in the partially observable CRL setting, this would mean that two variables are not always "inactive" and "active" at the same time, since we would otherwise not be able to disentangle them.

Even if this assumption does not hold strictly in a setting, our framework still supports a weaker form of identifiability: block-wise identifiability, which allows disentanglement across blocks of latent variables, even if

variables within each block may remain entangled. This relaxation still enables meaningful interpretability and structure in learned representations.

We ran experiments with different blocks of variables that were always active/inactive together. We use our basic setup: $n = 10$, $k = 1$, $m = 10$, $\rho = 50$, $\delta = 0$, $\theta = 0$ with three types of block to see whether the performance varies with different block sizes: $[[\mathbf{z}_0, \mathbf{z}_1], [\mathbf{z}_2, \mathbf{z}_3], [\mathbf{z}_4, \mathbf{z}_5], [\mathbf{z}_6, \mathbf{z}_7], [\mathbf{z}_8, \mathbf{z}_9]]$, $[[\mathbf{z}_0], [\mathbf{z}_1], [\mathbf{z}_2], [\mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5], [\mathbf{z}_6, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_9]]$, $[[\mathbf{z}_0], [\mathbf{z}_1], [\mathbf{z}_2, \mathbf{z}_3], [\mathbf{z}_4, \mathbf{z}_5, \mathbf{z}_6, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_9]]$, and we show R^2 of the linear regression between each block and learned representations $\hat{\mathbf{z}}_1, \ldots, \hat{\mathbf{z}}_10$. The results show that even though the variables are entangled within blocks, disentanglement can be achieved across blocks.

Table 20: Example results of $\hat{\mathbf{z}}_i$ assignments across blocks

	$[\mathbf{z}_0, \mathbf{z}_1]$	$[\mathbf{z}_2, \mathbf{z}_3]$	$[\mathbf{z}_4, \mathbf{z}_5]$	$[\mathbf{z}_6, \mathbf{z}_7]$	$[\mathbf{z}_8, \mathbf{z}_9]$
$\hat{\mathbf{z}}_0$	**0.95**	0.01	0.01	0.02	0.01
$\hat{\mathbf{z}}_1$	**0.90**	0.00	0.00	0.00	0.00
$\hat{\mathbf{z}}_2$	0.01	**0.84**	0.01	0.01	0.00
$\hat{\mathbf{z}}_3$	0.00	**0.84**	0.00	0.00	0.01
$\hat{\mathbf{z}}_4$	0.00	0.00	**0.91**	0.00	0.00
$\hat{\mathbf{z}}_5$	0.00	0.00	**0.91**	0.00	0.01
$\hat{\mathbf{z}}_6$	0.00	0.00	0.00	**0.86**	0.00
$\hat{\mathbf{z}}_7$	0.01	0.01	0.00	**0.96**	0.01
$\hat{\mathbf{z}}_8$	0.01	0.00	0.00	0.01	**0.91**
$\hat{\mathbf{z}}_9$	0.00	0.00	0.01	0.02	**0.88**

Table 21: Example results of $\hat{\mathbf{z}}_i$ assignments across blocks

	$[\mathbf{z}_0]$	$[\mathbf{z}_1]$	$[\mathbf{z}_2]$	$[\mathbf{z}_3, \mathbf{z}_4, \mathbf{z}_5]$	$[\mathbf{z}_6, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_9]$
$\hat{\mathbf{z}}_0$	**0.95**	0.00	0.00	0.01	0.03
$\hat{\mathbf{z}}_1$	0.00	**0.89**	0.00	0.00	0.00
$\hat{\mathbf{z}}_2$	0.01	0.00	**0.92**	0.01	0.02
$\hat{\mathbf{z}}_3$	0.00	0.00	0.00	**0.91**	0.02
$\hat{\mathbf{z}}_4$	0.01	0.00	0.00	**0.94**	0.01
$\hat{\mathbf{z}}_5$	0.00	0.00	0.00	**0.92**	0.00
$\hat{\mathbf{z}}_6$	0.00	0.00	0.00	0.00	**0.89**
$\hat{\mathbf{z}}_7$	0.01	0.00	0.01	0.01	**0.96**
$\hat{\mathbf{z}}_8$	0.01	0.00	0.00	0.01	**0.92**
$\hat{\mathbf{z}}_9$	0.00	0.00	0.00	0.02	**0.93**

Table 22: Example results of $\hat{\mathbf{z}}_i$ assignments across blocks

	$[\mathbf{z}_0]$	$[\mathbf{z}_1]$	$[\mathbf{z}_2, \mathbf{z}_3]$	$[\mathbf{z}_4, \mathbf{z}_5, \mathbf{z}_6, \mathbf{z}_7, \mathbf{z}_8, \mathbf{z}_9]$
$\hat{\mathbf{z}}_0$	**0.92**	0.00	0.01	0.05
$\hat{\mathbf{z}}_1$	0.00	**0.89**	0.00	0.01
$\hat{\mathbf{z}}_2$	0.01	0.00	**0.91**	0.04
$\hat{\mathbf{z}}_3$	0.00	0.00	**0.88**	0.01
$\hat{\mathbf{z}}_4$	0.01	0.00	0.00	**0.94**
$\hat{\mathbf{z}}_5$	0.01	0.00	0.01	**0.91**
$\hat{\mathbf{z}}_6$	0.02	0.00	0.00	**0.92**
$\hat{\mathbf{z}}_7$	0.00	0.00	0.00	**0.88**
$\hat{\mathbf{z}}_8$	0.01	0.00	0.00	**0.92**
$\hat{\mathbf{z}}_9$	0.00	0.00	0.00	**0.82**