Identifiability of deep generative models without auxiliary information

Bohdan Kivva[*]
University of Chicago
bkivva@uchicago.edu

Goutham Rajendran[*]
University of Chicago
goutham@uchicago.edu

Pradeep Ravikumar
Carnegie Mellon University
pradeepr@cs.cmu.edu

Bryon Aragam
University of Chicago
bryon@chicagobooth.edu

Abstract

We prove identifiability of a broad class of deep latent variable models that (a) have universal approximation capabilities and (b) are the decoders of variational autoencoders that are commonly used in practice. Unlike existing work, our analysis does not require weak supervision, auxiliary information, or conditioning in the latent space. Specifically, we show that for a broad class of generative (i.e. unsupervised) models with universal approximation capabilities, the side information u is not necessary: We prove identifiability of the entire generative model where we do not observe u and only observe the data x. The models we consider match autoencoder architectures used in practice that leverage mixture priors in the latent space and ReLU/leaky-ReLU activations in the encoder, such as VaDE and MFC-VAE. Our main result is an identifiability hierarchy that significantly generalizes previous work and exposes how different assumptions lead to different "strengths" of identifiability, and includes certain "vanilla" VAEs with isotropic Gaussian priors as a special case. For example, our weakest result establishes (unsupervised) identifiability up to an affine transformation, and thus partially resolves an open problem regarding model identifiability raised in prior work. These theoretical results are augmented with experiments on both simulated and real data.

1 Introduction

One of the key paradigm shifts in machine learning (ML) over the past decade has been the transition from handcrafted features to automated, data-driven representation learning, typically via deep neural networks. One complication of automating this step in the ML pipeline is that it is difficult to provide guarantees on what features will (or won't) be learned. As these methods are being used in high stakes settings such as medicine, health care, law, and finance where accountability and transparency are not just desirable but often legally required, it has become necessary to place representation learning on a rigourous scientific footing. In order to do this, it is crucial to be able to discuss ideal, target features and the underlying representations that define these features. As a result, the ML literature has begun to move beyond consideration solely of downstream tasks (e.g. classification, prediction, sampling, etc.) in order to better understand the structural foundations of deep models.

Deep generative models (DGMs) such as variational autoencoders (VAEs) [37, 56] are a prominent example of such a model, and are a powerful tool for unsupervised learning of latent representations, useful for a variety of downstream tasks such as sampling, prediction, classification, and clustering.

[*]Equal contribution

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Despite these successes, training DGMs is an intricate task: They are susceptible to posterior collapse and poor local minima [13, 24, 69, 72], and characterizing their latent space remains a difficult problem [e.g. 40, 67]. For example, does the latent space represent semantically meaningful or practically useful features? Are the learned representations stable, or are they simply artifacts of peculiar choices of hyperparameters? These questions have been the subject of numerous studies in recent years [e.g. 7, 12, 42, 45, 48, 60], and in order to better understand the behaviour of these models and address these questions, the machine learning literature has recently turned its attention to fundamental identifiability questions [14, 35, 69]. Identifiability is a crucial primitive in machine learning tasks that is useful for probing stability, consistency, and robustness. Without identifiability, the output of a model can be unstable and unreliable, in the sense that retraining under small perturbations of the data and/or hyperparameters may result in wildly different models.[1] In the context of deep generative models, the model output of interest is the latent space and the associated representations induced by the model.

In this paper, we revisit the identifiability problem in deep latent variable models and prove a surprising new result: Identifiability is possible under commonly adopted assumptions *and* without conditioning in the latent space, or equivalently, without weak supervision or side information in the form of auxiliary variables. This contrasts a recent line of work that has established fundamental new results regarding the identifiability of VAEs that requires conditioning on an auxiliary variable u that renders each latent dimension conditionally independent [35]. While this result has been generalized and relaxed in several directions [10, 22, 23, 36, 39, 43, 50, 61, 73], fundamentally these results still crucially rely on the side information u. We show that this is in fact unnecessary—confirming existing empirical studies [e.g 18, 70]—and do so without sacrificing any representational capacity. What's more, the model we analyze is closely related to deep architectures that have been widely used in practice [16, 18, 32, 33, 41, 41, 44, 71]: We show that there is good reason for this, and provide new insight into the properties of these models and support for their continued use.

Overview More specifically, we consider the following generative model for observations x:

$$x = f(z) + \varepsilon, \quad x = (x_1, \ldots, x_n) \in \mathbb{R}^n, \quad z = (z_1, \ldots, z_m) \in \mathbb{R}^m, \tag{1}$$

where the latent variable z follows a Gaussian mixture model (GMM),[2] $f : \mathbb{R}^m \to \mathbb{R}^n$ is a piecewise affine nonlinearity such as a ReLU network, and $\varepsilon \in \mathbb{R}^n$ is independent, random noise.[3] We do not assume that the number of mixture components, nor the architecture of the ReLU network, are known in advance, nor do we assume that z has independent components. Both the mixture model and neural network may be arbitrarily complex, and we allow for the discrete hidden state that generates the latent mixture prior to be high-dimensional and dependent. This includes both vanilla VAEs (i.e. with a standard isotropic Gaussian prior) and classical ICA models (i.e. for which the latent variables are mutually independent) as special cases. Since both z and f are allowed to be arbitrarily complex, the model (1) has universal approximation capabilities, which is crucial for modern applications.

This model has been widely studied in the literature from a variety of different perspectives:

- *Nonlinear ICA.* When the z_i are mutually independent, (1) recovers the standard nonlinear ICA model that has been extensively studied in the literature [1, 26–29, 74]. Although our most general results do not make independence assumptions, our results cover nonlinear ICA as a special case (see Section A.2 for more discussion).
- *VAE with mixture priors.* When the prior over z is a mixture model (e.g. such as a GMM), the model (1) is closely related to popular autoencoder architectures such as VaDE [32], SVAE [33], GMVAE [16], DLGMM [52], VampPrior [66], MFC-VAE [18], etc. Although such VAEs with mixture priors have been used extensively in applications, theoretical results are missing.
- *Warped mixtures.* Another closely related model is the warped mixture model of Iwata et al. [31], which is a Bayesian version of (1). Once again, theoretical guarantees for these models are lacking.

[1] Formally, identifiability means the parametrization of the model is injective. See Section 2 for details.

[2] See Remark 2.1 for extensions to more general mixture priors.

[3] Our results include the noiseless case $\varepsilon = 0$ as a special case.

Assumptions on f	Assumptions on Z	Theoretical guarantees	Result
(P1)	(F1), (F2)	$\mathbb{P}(Z)$ identifiable up to an affine transformation	Theorems 3.9(a), 3.10(a)
(P1)	(F1), (F4)	$\mathbb{P}(Z)$ and f up to identifiable an affine transformation	Theorems 3.9(c), 3.10(d)
(P1), (P2)	(F1), (F4)	$\mathbb{P}(Z)$ and f identifiable up to permutation, scaling and translation	Theorems 3.9(b), 3.10(b)
(P1), (P2), (P3)	(F1), (F4)	$\mathbb{P}(U, Z)$ and f are identifiable up to permutation, scaling and translation	Theorems 3.10(c), 3.10(d)

Table 1: Summary of results in this paper. The strength of the assumptions increases in each successive row, as do the strength of the guarantees. See Section 3.3 for formal statements.

- *iVAE.* Finally, (1) is also the basis of the iVAE model introduced by Khemakhem et al. [35], where identifiability (up to certain equivalences) is proved when there is an additional auxiliary variable u that is observed such that $z_i \perp\!\!\!\perp z_j \mid u$.

Contributions Driven by this recent interest from both applied and theoretical perspectives, our main results (Theorems 3.9, 3.10) show that the model (1) is identifiable up to various *linear* equivalences, without conditioning or auxiliary information in the latent space. In fact, we develop a hierarchy of results under progressively stronger assumptions on the model, beginning with affine equivalence and ending up with a much stronger equivalence up to permutations only. See Table 1 for a summary.

In order to develop this hierarchy, we prove several technical results of independent interest:

1. First, we establish a novel identifiability result for nonparametric mixtures (Theorem C.2);
2. Second, we show how to use the mixture prior to strengthen existing identifiability results for nonlinear ICA (Theorem D.1);
3. Third, we extend existing results [38] on the recovery of structured multivariate discrete latent variable models to recovery under an unknown affine transformation (Theorem F.2).

Our proof techniques—based on elementary tools from analytic function theory and mixture identifiability—are new and depart from existing work in this area. As a consequence, the analysis itself provides new insight into the structure and behaviour of deep generative models.

Related work This problem is widely studied, and has garnered significant recent interest, so we focus only on the most closely related work here.

Classical results on nonlinear ICA [28] establish the nonidentifiability of the general model (i.e. without restrictions on z and f); see also Darmois [15], Jutten et al. [34]. More recently, Khemakhem et al. [35] proved a major breakthrough by showing that given side information u, identifiability of the entire generative model is possible up to certain (nonlinear) equivalences. Since this pathbreaking work, many generalizations have been proposed [10, 22, 23, 36, 39, 43, 50, 61, 73], all of which require some form of auxiliary information. Other approaches to identifiability include various forms of weak supervision such as contrastive learning [75], group-based disentanglement [46], and independent mechanisms [20]. Non-identifiability has also been singled out as a contributing factor to practical issues such as posterior collapse in VAEs [69, 72].

Our approach is to avoid additional forms of supervision altogether, and enforce identifiability in a purely unsupervised fashion. Recent work along these lines includes Wang et al. [69], who propose to use Brenier maps and input convex neural networks, and Moran et al. [51] who leverage sparsity and an anchor feature assumption. Aside from different assumptions, the main difference between this line of work and our work is that their work only identifies the latent space $P(Z)$, whereas our focus is on jointly identifying *both* $P(Z)$ and f. In fact, we provide a decoupled set of assumptions

that allow f or $P(Z)$ or both to be identified. Thus, we partially resolve in the affirmative an open problem regarding model identifiability raised by the authors in their discussion.

Another distinction between this line of work and the current work is our focus on architectures and modeling assumptions that are *standard* in the deep generative modeling literature, specifically ReLU nonlinearities and mixture priors. As noted above, there is a recent tradition of training variational autoencoders with mixture priors [16, 18, 32, 33, 41, 41, 44, 71]. Our work builds upon this empirical literature, showing that there is good reason to study such models: Not only have they been shown to be more effective compared to vanilla VAEs, we show that they have appealing theoretical properties as well. In fact, recent work [18, 70] has observed precisely the identifiability phenomena studied in our paper, however, this work lacks rigourous theoretical results to explain these observations.

Another related line of work studies identification in graphical models with latent variables, albeit without any explicit connection to deep generative models [17, 38, 49, 54].

Finally, since a key step in our proof involves the analysis of a nonparametric mixture model (see Appendix C for details), it is worth reviewing previous work in mixture models. See Allman et al. [2] for an overview. Of particular use for the present work are Teicher [63] and Barndorff-Nielsen [8], wherein the identifiability of Gaussian and exponential family mixtures, respectively, are proved. Specifically for nonparametric mixtures, existing results consider product mixtures [21, 64], grouped observations [57, 68], symmetric measures [9, 25], and separation conditions [4]. For context, we note here that a discrete VAE can be interpreted as a mixture model in disguise: This is a perspective that we leverage in our proofs. We are not aware of previous work in the deep generative modeling literature that exploits this connection to prove identifiability results.

2 Preliminaries

We first introduce the main generative model that we study and its properties, and then proceed with a brief review of identifiability in deep generative models.

Generative model The observations $x \in \mathbb{R}^n$ are realizations of a random vector X, and are generated according to the generative model (1), where $z \in \mathbb{R}^m$ represents realizations of an unobserved random vector Z. We make the following assumptions on Z and f:[4]

(P1) $P(Z)$ is a (possibly degenerate) Gaussian mixture model with an unknown number of components $J \geq 1$, i.e.

$$p(z) = \sum_{j=1}^{J} \lambda_j \varphi(z; \mu_j, \Sigma_j), \quad \sum_{j=1}^{J} \lambda_j = 1, \quad \lambda_j > 0, \tag{2}$$

where $p(z)$ is the density of $P(Z)$ with respect to some base measure, and $\varphi(z; \mu_j, \Sigma_j)$ is the gaussian density with mean μ_j and covariance Σ_j.

(F1) f is a piecewise affine function, such as a multilayer perceptron with ReLU (or leaky ReLU) activations.

Recall that an affine function is a function $x \mapsto Ax + b$ for some matrix A. As already discussed, special cases of this model have been extensively studied in both applications and theory, and both (P1)-(F1) are quite standard in the literature on deep generative models and represent a useful model that is widely used in practice [e.g. 16, 18, 32, 33, 41, 41, 44, 71]. In particular, when $J = 1$ this is simply a classical VAE with an isotropic Gaussian prior (see Section A.2 for more discussion).

Remark 2.1. The assumption that $P(Z)$ is a GMM can be replaced with more general exponential family mixtures [8] as long as (a) the resulting mixture prior $p(z)$ is an analytic function and (b) the exponential family is closed under affine transformations.

Universal approximation Under assumptions (P1)-(F1), the model (1) has universal approximation capabilities. In fact, any distribution can be approximated by a mixture model (2) with sufficiently

[4] In the sequel, we will use (P#) to index assumptions on the prior $P(Z)$, and (F#) to index assumptions on the decoder f.

many components J [e.g. 53]. Alternatively, when J is bounded, by taking f to be a sufficiently deep and/or wide ReLU network, any distribution can be approximated by $f(Z)$ [e.g. 47, 65], even if f is invertible [30]. Thus, there is no loss in representational capacity in (P1)-(F1). To the best of our knowledge, our results are the first to establish identifiability of *both* the latent space and decoder for deep generative models *without* conditioning in the latent space or weak supervision. We note that Wang et al. [69] and Moran et al. [51] also propose deep architectures that identify the latent space, but not the decoder.

Identifiability A statistical model is specified by a (possibly infinite-dimensional, as in our setting) parameter space Θ, a family of distributions $\mathcal{P}$, and a mapping $\pi : \Theta \to \mathcal{P}$; i.e. $\pi(\theta) \in \mathcal{P}$ for each $\theta \in \Theta$. In more conventional notation, we define $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$, in which case $p_\theta = \pi(\theta)$. A statistical model is called *identifiable* if the parameter mapping π is one-to-one (injective). In practical applications, the strict definition of identifiability is too strong, and relaxed notions of identifiability are sufficient. Classical examples include identifiability up to permutation, re-scaling, or orthogonal transformation. More generally, a statistical model is *identifiable up to an equivalence relation* $\sim$ defined on Θ if $\pi(\theta) = \pi(\theta') \implies \theta \sim \theta'$. For more details on the different notions of identifiability in deep generative models, see [35, 36, 58].

More precisely, we use the following definition. Let $f_\sharp P$ denote the pushforward measure of P by f.

Definition 2.2. Let $\mathcal{P}$ be a family of probability distributions on $\mathbb{R}^m$ and $\mathcal{F}$ be a family of functions $f : \mathbb{R}^m \to \mathbb{R}^n$.

1. For $(P, f) \in \mathcal{P} \times \mathcal{F}$ we say that the prior P is *identifiable (from $f_\sharp P$) up to an affine transformation* if for any $(P', f') \in \mathcal{P} \times \mathcal{F}$ such that $f_\sharp P \equiv f'_\sharp P'$ there exists an invertible affine map $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $P' = h_\sharp P$ (i.e., P' is the pushforward measure of P by h).

2. For $(P, f) \in \mathcal{P} \times \mathcal{F}$ we say that the pair (P, f) is *identifiable (from $f_\sharp P$) up to an affine transformation* if for any $(P', f') \in \mathcal{P} \times \mathcal{F}$ such that $f_\sharp P \equiv f'_\sharp P'$ there exists an invertible affine map $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $f' = f \circ h^{-1}$ and $P' = h_\sharp P$.

If the noise ε has a known distribution, then $f_\sharp P$ is identifiable from the convolution $(f_\sharp P) * \varepsilon$. Hence, this definition can be automatically extended to the setup with known noise. This definition also can be extended to transformations besides affine transformations (e.g. permutations, translations, etc.) in the obvious way.

Identifiability is a crucial property for a statistical model: Without identifiability, different training runs may lead to very different parameters, making training unpredictable and replication difficult. The failure of identifiability, also known as *underspecification* and *ill-posedness*, has recently been flagged in the ML literature as a root cause of many failure modes that arise in practice [14, 69, 72]. As a result, there has been a growing emphasis on identification in the deep learning literature, which motivates the current work. Finally, in addition to these reproducibility and interpretability concerns, identifiability is a key component in many applications of latent variable models including causal representation learning [59], independent component analysis [11], and topic modeling [3, 5]. See Ran and Hu [55] for additional discussion and examples.

Auxiliary information and iVAE It is well-known that assuming independence of the latent factors—i.e. $Z_i \perp\!\!\!\perp Z_j$—is insufficient for identifiability [28]. Recent work, starting with iVAE, shows identifiability by additionally assuming that a k-dimensional auxiliary variable u is observed such that $p(z \mid u)$ is conditionally factorial, i.e. $Z_i \perp\!\!\!\perp Z_j \mid U$. This extra information serves to break symmetries in the latent space and is crucial to existing proofs of identifiability.

To make the connection with this work clear, observe that assumption (P1) is equivalent to assuming that there is an additional hidden state $U \in \{1, \dots, J\}$ such that $P(Z = z \mid U = j) = p_j(z)$ and $P(U = j) = \lambda_j$. More generally, $U = (U_1, \dots, U_k)$ may be multivariate. In this way, a direct parallel between our work and previous work is evident, with several crucial caveats:

- We do *not* assume that U is observed—even partially—or known in any way;
- We allow for the Z_i to be arbtrarily dependent even after conditioning on U, and this dependence need not be known;

- We do not even require the number of states J to be known, and we do not require any bounds on J (e.g. iVAE requires $J \geq m+1$).
- In the case where U is multivariate (i.e $k := \dim(U) > 1$), we do not require the number of latent dimensions k, the state spaces, or their dependencies to be known.
- The original iVAE paper only proves identifiability of f up to a nonlinear transformation (see Lemma G.2 in Appendix G for details). By contrast, we will show identifiability of f up to an affine transformation, without knowing U.

In order to break the symmetry without knowing anything about U or its dependencies, we develop fundamentally new insights into nonparametric identifiability of latent variable models.

3 Main results

For any positive integer d, let $[d] = \{1, \ldots, d\}$. By (P1), we can write the model (1) as follows. Let $U = (U_1, \ldots, U_k) \in [d_1] \times \cdots [d_k]$ where $d_i := \dim(U_i)$ and $k := \dim(U)$; we allow U to be multivariate ($k > 1$) and dependent—i.e., we do not assume that the U_i are marginally independent. It follows trivially from (P1) that $P(U_1 = u_1, \ldots, U_k = u_k) \in \{\lambda_1, \ldots, \lambda_J\}$ and $J = \prod_i d_i$, where we recall that J is the *unknown* number of mixture components in $P(Z)$. Denote the marginal distribution of U, which depends on λ_j, by P_λ. The variables (U, Z) are unobserved and encode the underlying latent structure:

$$\left.\begin{aligned} U = u &\sim P_\lambda(U = u) \\ [Z \,|\, U = u] &\sim N(\mu_u, \Sigma_u) \\ [X \,|\, Z = z] &\sim f(z) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2) \end{aligned}\right\} \implies U \to Z \to X. \tag{3}$$

Here, P_λ is the distribution on U described above. Our goal is to identify the latent distribution $P(U, Z)$ and/or the nonlinear decoder f from the marginal distribution $P(X)$ induced by (3). We will additionally assume throughout that $m \leq n$; see Remark 3.6 for a discussion of the overcomplete case with $m > n$.

Our main results (Theorems 3.9-3.10) provide a hierarchy of progressively stronger conditions under which $P(U, Z)$, f, or both, can be identified in progressively stronger ways. The idea is to illustrate explicitly what conditions are sufficient to identify the latent structure up to affine equivalence (the weakest notion of identifiability we consider), equivalence up to permutation, scaling, and translation, and permutation equivalence (the strongest notion of identifiability we consider, and the strongest possible for any latent variable model).

We defer the statement of the main results to Section 3.3, after the main conditions have been described. As a preview to the main results, we first present the following corollary:

Corollary 3.1. *Suppose $k = \dim(U) = 1$, $J \geq 1$, (U, Z) are unobserved, and X is observed. (a) If f is an invertible ReLU network, then both $P(U, Z)$ and f are identifiable up to an affine transformation. (b) If f is only weakly injective (cf. (F2)), then $P(U, Z)$ is still identifiable up to an affine transformation.*

For comparison, Corollary 3.1 already strengthens existing results, since U is not required to be known and we are able to identify f. In fact, the latter answers an open question raised by Wang et al. [69]. What's more, this is just the *weakest* result implied by our main results: Under stronger assumptions on the latent structure, the affine equivalence presented above can be strengthened further.

Taken together, the results in this section have the following concrete implication for practitioners: For stably training variational autoencoders, there is now compelling justification to work with a GMM prior and deep ReLU/Leaky-ReLU networks. As we saw above, this is commonly done in practice already.

3.1 Possible assumptions on f

To distinguish cases where f is and is not identifiable, we require the following technical definition. Recall that for sets A, B, $f^{-1}(A) = \{x : f(x) \in A\}$ and $f(B) = \{f(x) : x \in B\}$.

Definition 3.2. Let $m \leq n$ (see Remark 3.6) and $f : \mathbb{R}^m \to \mathbb{R}^n$.

(F2) We say that f is *weakly injective* if (i) there exists $x_0 \in \mathbb{R}^n$ and $\delta > 0$ s.t. $|f^{-1}(\{x\})| = 1$ for every $x \in B(x_0, \delta) \cap f(\mathbb{R}^m)$, and (ii) $\{x \in \mathbb{R}^n : |f^{-1}(\{x\})| = \infty\} \subseteq f(\mathbb{R}^m)$ has measure zero with respect to the Lebesgue measure on $f(\mathbb{R}^m)$.

(F3) We say that f is *observably injective* if $\{x \in \mathbb{R}^n : |f^{-1}(\{x\})| > 1\} \subseteq f(\mathbb{R}^m)$ has measure zero with respect to the Lebesgue measure on $f(\mathbb{R}^m)$. In other words, f is injective for almost every x in its image $f(\mathbb{R}^m)$ (i.e. almost every "observable" x).

(F4) We say that f is *injective* if $|f^{-1}(\{x\})| = 1$ for every $x \in f(\mathbb{R}^m)$.

Remark 3.3. For piecewise affine functions assumption (F2) is weaker than assumption (F3), which in turn is weaker than (F4). Therefore, for piecewise affine functions we have the chain of implications:

$$\text{(F4)} \implies \text{(F3)} \implies \text{(F2)}.$$

In the sequel, we mostly focus on (F2) and (F4) for simplicity; although we prove results for (F3) in Appendix D.1. See also Remarks 3.5, 3.11.

Example 3.4. In general, a deep ReLU network may be either injective or observably injective, or neither (e.g. $\mathrm{ReLU}(-\mathrm{ReLU}(x)) = 0$). For example, although $x \mapsto \mathrm{ReLU}(x)$ is not injective, it is observably injective, where $\mathrm{ReLU}(x) = \max\{0, x\}$ is the usual rectified linear unit. To see this, note that image of ReLU is the set $\mathbb{R}_{\geq} = \{y \mid y \geq 0\}$, and ReLU has the unique preimage for every $y \in \mathbb{R}_{>} = \{y \mid y > 0\}$. Clearly, $(\mathbb{R}_{\geq} \setminus \mathbb{R}_{>}) = \{0\}$ has measure zero inside $\mathbb{R}_{\geq}$.

At the same time, $x \mapsto 0$ and $x \mapsto |x|$ are not even weakly injective.

Remark 3.5. In Appendix H, we show that ReLU networks or Leaky ReLU networks are generically observably injective (and hence also weakly injective) under simple assumptions on their architecture.

Remark 3.6. We restrict attention to the case $m \leq n$, which is a standard assumption, as it is common to think of a latent space to be a low-dimensional representation of the observed space. In the overcomplete case, i.e. when $m > n$, we believe that identifiability is unlikely unless stronger assumptions are made, or weaker notions of identifiability are considered. To see this, consider the projection $f(x, y) = x$, which is trivially affine. Then we can arbitrarily transform the y-coordinate without changing P, i.e. $(f \circ g)_\sharp P = f_\sharp P$, where $g(x, y) = (x, h(y))$ for any h. As an example of identifiability in the overcomplete regime under stronger assumptions, when the auxiliary variable u is known, [36] show that the feature maps f and g in conditional energy-based models (for which $p(x \mid u) \propto \exp(f(x)^T g(u))$) can be identified up to an affine transformation.

3.2 Possible assumptions on Z

Our weakest result requires no additional assumptions on Z beyond (P1); see Corollary 3.1. Under stronger assumptions, more can be concluded. As with the previous section, the assumptions presented here are not necessary, but may be imposed in order to extract stronger results.

The first condition is a mild condition that allows us to strengthen affine identifiability:

(P2) $Z_i \perp\!\!\!\perp Z_j \mid U$ for all $i \neq j$ and there exist a pair of states $U = u_1$ and $U = u_2$ such that all $((\Sigma_{u_1})_{tt} / (\Sigma_{u_2})_{tt} \mid t \in [m])$ are distinct. (Note that this implies $J \geq 2$).

The second condition is more technical, and is only necessary if $k > 1$ and we wish to identify $P(U)$ in addition to $P(Z)$. In fact, not only will we recover $P(U)$, but also the (unknown) number of hidden variables (i.e. k) and their state spaces (i.e. d_j). Note that $P(U)$ is not needed to sample from (1), as long as we have $P(Z)$. Before introducing this condition, we need a preliminary definition.

Definition 3.7. Let U_{-i} denote $\{U_j : j \neq i\}$. We define $\mathrm{ne}(U_i) = [m] \setminus \{t : Z_t \perp\!\!\!\perp U_i \mid U_{-i}\}$ and $\mathrm{ne}(Z_i) = \{t : Z_i \in \mathrm{ne}(U_t)\}$. For a subset $Z' \subset Z$, $\mathrm{ne}(Z') = \cup_{Z_i \in Z'} \mathrm{ne}(Z_i)$.

The neighborhood $\mathrm{ne}(U_i)$ collects the variables Z_t that depend on U_i directly.

(P3) The following conditions hold:

(a) For all $Z' \subset Z$ and $u_1 \neq u_2$, $P(Z' \mid \mathrm{ne}(Z') = u_1) \neq P(Z' \mid \mathrm{ne}(Z') = u_2)$;

(b) If $P(U', Z, X) = P(U, Z, X)$, then $\dim(U') \leq \dim(U)$; and

(c) For any $U_i \neq U_j$ the set $\mathrm{ne}(U_i)$ is not a subset of $\mathrm{ne}(U_j)$.

Condition (P3) is a "maximality" condition that is adapted from Kivva et al. [38]: We are interested in identifying the most complex latent structure with the most number of hidden variables. This is in fact necessary since we can always merge two (or more) hidden variables into a single hidden variable without changing the joint distribution. Moreover, if two distinct hidden variables $U_i \neq U_j$ have the same neighborhood (or one is a subset of another), then it is known that $P(U)$ cannot be identified [17, 38, 54]. Evidently, if we seek to learn $P(U)$ in addition to $P(Z)$, then this must be avoided. Finally, as the proof will indicate, this condition is slightly stronger than what is needed (see Remark F.3 for details).

Remark 3.8. Condition (P3) should be contrasted with the stronger "anchor words" assumption that has appeared in prior work [5, 6, 51]: In fact, the existence of an anchor word for each U_j automatically implies that $\mathrm{ne}(U_i)$ is not a subset of $\mathrm{ne}(U_j)$ for $i \neq j$. Thus, anchor words are a sufficient but not necessary condition for identifiability, whereas Condition (P3) is indeed necessary as described above.

More details and discussion on these assumptions can be found in Appendix F.

3.3 Main identifiability results

When $\dim(U) = 1$, there is no additional structure in U to learn, and so the setting simplifies considerably. We begin with this special case before considering the case of general multivariate U.

Theorem 3.9. *Assume* $\dim(U) = 1$. *Under (P1)-(F1), we have the following:*

(a) (F2) $\implies$ *$P(U, Z)$ is identifiable from $P(X)$ up to an affine transformation of Z.*

(b) (F2)+(P2) $\implies$ *$P(U, Z)$ is identifiable from $P(X)$ up to permutation, scaling, and/or translation of Z.*

(c) In either (a) or (b), if additionally (F4) holds and f is continuous, then f is also identifiable from $P(X)$ up to an affine transformation.

The next result generalizes Theorem 3.9 to arbitrary (possibly multivariate) discrete U. This is an especially challenging case: Unlike previous work such as iVAE that assumes U (and hence its structure) is known, we do not assume anything about U is known. Thus, everything about U must be reconstructed based on $P(X)$ alone, hence the need for (P3) to identify $P(U)$ below.

Theorem 3.10. *Under (P1)-(F1), we have the following:*

(a) (F2) $\implies$ *$P(Z)$ is identifiable from $P(X)$ up to an affine transformation.*

(b) (F2)+(P2) $\implies$ *$P(Z)$ is identifiable from $P(X)$ up to permutation, scaling, and/or translation.*

(c) (F2)+(P2)+(P3) $\implies$ *$(k, d_1, \ldots, d_k, P(U))$ are identifiable from $P(X)$ up to a permutation of U, and $P(Z)$ is identifiable up to permutation, scaling, and/or translation.*

(d) In any of (a), (b), or (c), if additionally (F4) holds and f is continuous, then f is also identifiable from $P(X)$ up to an affine transformation.

Without (P3), Kivva et al. [38] have shown that it is not possible to recover the high-dimensional latent state U, however, we can still identify the continuous latent state Z, which is enough to generate random samples from the model (1). In order to have fine-grained control over the individual variables in U, however, it is necessary to assume (P3).

Remark 3.11. If (F4) is relaxed to (F3) f may not be identifiable up to an affine transformation, but it is "essentially" identifiable in the following sense. Let $S = \{x : |f^{-1}(\{x\})| > 1\}$. On every connected component of $\mathbb{R}^m \setminus f^{-1}(S)$, f is identifiable up to an affine transformation (which may depend on the connected component). Note, for f defined by a ReLU NN, points of S are atoms of $P(X)$.

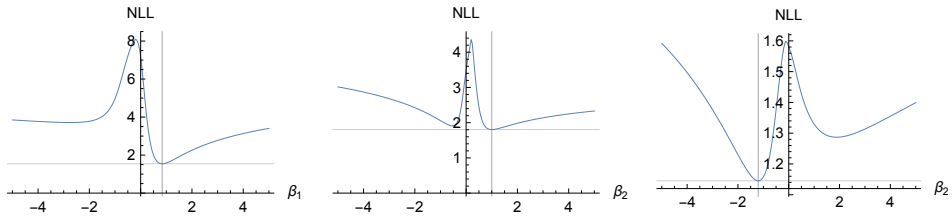


Figure 1: Selected examples of the negative log-likelihood for different runs. In each figure, one parameter from a model (e.g. β_j is a weight in the neural network defining f) is selected, and the value of the negative log-likelihood is visualized as a function of this parameter. Vertical lines indicate the ground truth and (global) minimizer, which always coincide. Three particularly interesting, nonconvex examples are shown here. See Appendix J.3.1 for details.

Remark 3.12. If the assumption (F2) that f is weakly injective is removed, then the claim of Theorem 3.9 is not true anymore. Consider $g(x) = f(x) = |x|$ and

$$\begin{aligned} P &= \frac{1}{3}N(-2,\sigma^2) + \frac{1}{3}N(-1,\sigma^2) + \frac{1}{3}N(3,\sigma^2) \quad \text{and} \\ P' &= \frac{1}{3}N(-2,\sigma^2) + \frac{1}{3}N(1,\sigma^2) + \frac{1}{3}N(3,\sigma^2). \end{aligned} \tag{4}$$

It is easy to verify that P cannot be transformed into P' by an affine transformation, but $f_\sharp P$ and $g_\sharp P'$ are equally distributed.

Remark 3.13. In Theorems 3.9(a) and 3.10(a), the identifiability up to an affine transformation is the best possible if no additional assumptions on Z are made (i.e. beyond (P1)). Indeed, for an arbitrary invertible affine map $h : \mathbb{R}^m \to \mathbb{R}^m$, $h(Z)$ has a GMM distribution, $f \circ h^{-1}$ is an invertible piecewise affine map, and (U, Z, f) and $(U, h(Z), f \circ h^{-1})$ in model (3) generate the same distribution.

4 Experiments

There has been extensive work already to verify empirically that the model (1) under (P1)-(F1) is identifiable. For example, [70] observe that deep generative models with clustered latent spaces are empirically identifiable, and compared this directly to models that rely on side information, and [18] show that meaningful latent variables can be learned consistently in a fully unsupervised manner even when U has high-dimensional structure. Moreover, [18] indicate that high-dimensional structure is important for improved performance. Beyond these, it is well-known that VAEs with mixture priors such as VaDE [32] achieve competitive performance on many benchmark tasks; see [16, 18, 33, 41, 41, 44, 71] for additional experiments and verification. Building upon the established success of these methods, we augment these experiments as follows: 1) We use simple examples to verify that the likelihood indeed has a unique minimizer at the ground truth parameters; 2) We train VaDE on (misspecified) simulated toy models; and 3) We measure stability (up to affine transformations) of the learnt latent spaces on real data. To measure this, we report the Mean Correlation Coefficient [36, Appendix A.2] metric, which is standard, and an L^2-based alignment metric (denoted by $\mathrm{dist}_{\mathrm{Aff},L2}$). Definitions of these metrics and additional details on the experiments can be found in Appendix J.

Maximum likelihood We simulated models satisfying (P1)-(F1) by randomly choosing weights and biases for a single-layer ReLU network and randomly generating a GMM with $J = 2$ or 3 components. These models are simple enough that exact computation of the MLE along the likelihood surface is feasible via numerical integration (Figure 1). In all our simulations (50 total), the ground truth was the unique minimizer of the negative log-likelihood, as predicted by the theory. These examples also illustrate a small-scale test of misspecification in the theoretical model: We include cases where J is misspecified and f fails to satisfy (F4), but the MLE succeeds anyway.

Simulated data In our experiments on synthetic datasets we consider, to obtain an experimental evidence of identifiability of model (3) we fit VaDE to observed data 5 times (see Figure 2). Let

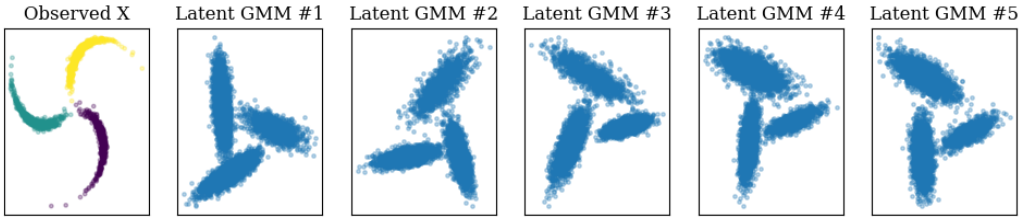


Figure 2: Recovered latent spaces for 5 runs of VaDE on pinwheel dataset with 3 clusters

$Z^{(1)}, Z^{(2)}, \ldots, Z^{(5)}$ be the learned latent spaces. For every pair $Z^{(i)}, Z^{(j)}$ we evaluate the MCC and $\mathrm{dist}_{\mathrm{Aff},L2}$ loss. For instance, for the pinwheel dataset with three clusters as in Figure 2, the average $\mathrm{dist}_{\mathrm{Aff},L2}(p_1, p_2)$ across 20 pairs $Z^{(i)}, Z^{(j)}$ is 0.113 with standard deviation 0.065. The average weak MCC is 0.87 and the average strong MCC is 1.0. This shows strong evidence of recovery of the latent space up to affine transformations.

Real data We measure stability of the learnt latent space by training MFCVAE [18] on MNIST 10 times with different initializations and then comparing the latent representations learnt. It becomes computationally infeasible to compute $\mathrm{dist}_{\mathrm{Aff},L2}$ therefore we report only MCC. The strong MCCs are computed to be 0.7 (ReLU), 0.69 (LeakyReLU) and the weak MCCs are computed to be 0.91 (ReLU), 0.94 (LeakyReLU). These observations validate the observations first made in [70], who ran extensive experiments on VaDE and iVAE on several large datasets including MNIST, SVHN and CIFAR10. These strong correlations confirm our theory and are of particular importance to practitioners for whom stability of learning is of the essence.

5 Conclusion

We have proved a general series of results describing a hierarchy of identifiability for deep generative models that are currently used in practice. Our experiments confirm both on exact and approximate simulations that identifiability indeed holds in practice. An obvious direction for future work is to study finite-sample identifiability problems such as sample complexity and robustness (i.e. how many samples are needed to ensure that the global minimizer of the likelihood is reliably close to the ground truth?). Theoretical questions aside, developing a better understanding of the ELBO and its effect on optimization is an important practical question. For example, an important limitation of the current set of results is that they apply only to the likelihood, which is known to be nonconvex and intractable to optimize (see Figure 1 for concrete examples). It is an important open question to use these insights to develop better algorithms and optimization techniques that work on finite-samples with misspecified models (i.e. real data).

More generally, although our assumptions map onto architectures and priors that are widely used in practice, it is important to emphasize the relevant distinction between models and estimators. That is, the architectures used in practice represent the *estimators* used, and may not reflect realistic assumptions on the *model* itself (which is typically misspecified). For example, the piecewise affine assumption may not accurately reflect valid assumptions about real-world problems. Given the lack of purely unsupervised, nonparametric identifiability results in the literature, we view our results as an important technical step towards understanding practical identifiability for deep generative models. Thus, an important future direction is to replace our assumptions with more appropriate modeling assumptions that are relevant for practical applications.

6 Acknowledgements

We thank anonymous reviewers for useful comments and suggestions. G.R. was partially supported by NSF grants CCF-1816372 and CCF-200892. B.A. was supported by NSF IIS-1956330, NIH R01GM140467, and the Robert H. Topel Faculty Research Fund at the University of Chicago Booth School of Business. P.R. was supported by ONR via N000141812861, and NSF via IIS-1909816, IIS-1955532, IIS-2211907.

References

[1] S. Achard and C. Jutten. Identifiability of post-nonlinear mixtures. *IEEE Signal Processing Letters*, 12(5):423–426, 2005.

[2] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, pages 3099–3132, 2009.

[3] A. Anandkumar, D. J. Hsu, M. Janzamin, and S. M. Kakade. When are overcomplete topic models identifiable? uniqueness of tensor tucker decompositions with structured sparsity. *Advances in neural information processing systems*, 26, 2013.

[4] B. Aragam, C. Dan, E. P. Xing, and P. Ravikumar. Identifiability of nonparametric mixture models and bayes optimal clustering. *Ann. Statist.*, 48(4):2277–2302, 2020. ISSN 0090-5364. doi: 10.1214/19-AOS1887. arXiv:1802.04397.

[5] S. Arora, R. Ge, and A. Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

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

[7] Y. Bansal, P. Nakkiran, and B. Barak. Revisiting model stitching to compare neural representations. *Advances in Neural Information Processing Systems*, 34, 2021.

[8] O. Barndorff-Nielsen. Identifiability of mixtures of exponential families. *Journal of Mathematical Analysis and Applications*, 12(1):115–121, 1965.

[9] L. Bordes, S. Mottelet, and P. Vandekerkhove. Semiparametric estimation of a two-component mixture model. *Annals of Statistics*, 34(3):1204–1232, 2006.

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

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

[12] A. Csiszárik, P. Kőrösi-Szabó, Á. Matszangosz, G. Papp, and D. Varga. Similarity and matching of neural network representations. *Advances in Neural Information Processing Systems*, 34, 2021.

[13] B. Dai, Z. Wang, and D. Wipf. The usual suspects? reassessing blame for vae posterior collapse. In *International Conference on Machine Learning*, pages 2313–2322. PMLR, 2020.

[14] A. D'Amour, K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

[15] G. Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951.

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

[17] R. J. Evans. Graphs for margins of bayesian networks. *Scandinavian Journal of Statistics*, 43 (3):625–648, 2016.

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

[19] É. Gassiat, A. Cleynen, and S. Robin. Inference in finite state space non parametric hidden markov models and applications. *Statistics and Computing*, 26(1):61–71, 2016.

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

[21] P. Hall and X.-H. Zhou. Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, pages 201–224, 2003.

[22] H. Hälvä and A. Hyvarinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence*, pages 939–948. PMLR, 2020.

[23] H. Hälvä, S. L. Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ica. *arXiv preprint arXiv:2106.09620*, 2021.

[24] J. He, D. Spokoyny, G. Neubig, and T. Berg-Kirkpatrick. Lagging inference networks and posterior collapse in variational autoencoders. In *International Conference on Learning Representations*, 2018.

[25] D. R. Hunter, S. Wang, and T. P. Hettmansperger. Inference for mixtures of symmetric distributions. *Annals of Statistics*, pages 224–251, 2007.

[26] A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in Neural Information Processing Systems*, 29, 2016.

[27] A. Hyvarinen and H. Morioka. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017.

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

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

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

[31] T. Iwata, D. Duvenaud, and Z. Ghahramani. Warped mixtures for nonparametric cluster shapes. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 311–320, 2013.

[32] Z. Jiang, Y. Zheng, H. Tan, B. Tang, and H. Zhou. Variational deep embedding: An unsupervised and generative approach to clustering. *arXiv preprint arXiv:1611.05148*, 2016.

[33] M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. R. Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in neural information processing systems*, 29, 2016.

[34] C. Jutten, J. Karhunen, et al. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pages 245–256, 2003.

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

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

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

[38] B. Kivva, G. Rajendran, P. Ravikumar, and B. Aragam. Learning latent causal graphs via mixture oracles. *Advances in Neural Information Processing Systems*, 34, 2021.

[39] D. A. Klindt, L. Schott, Y. Sharma, I. Ustyuzhaninov, W. Brendel, M. Bethge, and D. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2020.

[40] J. Klys, J. Snell, and R. Zemel. Learning latent subspaces in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

[41] D. B. Lee, D. Min, S. Lee, and S. J. Hwang. Meta-gmvae: Mixture of gaussian vae for unsupervised meta-learning. In *International Conference on Learning Representations*, 2020.

[42] K. Lenc and A. Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999, 2015.

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

[44] X. Li, Z. Chen, L. K. Poon, and N. L. Zhang. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. *arXiv preprint arXiv:1803.05206*, 2018.

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

[46] F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

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

[48] G. Luise, M. Pontil, and C. Ciliberto. Generalization properties of optimal transport gans with latent distribution learning. *arXiv preprint arXiv:2007.14641*, 2020.

[49] A. Markham and M. Grosse-Wentrup. Measurement dependence inducing latent causal models. In *Conference on Uncertainty in Artificial Intelligence*, pages 590–599. PMLR, 2020.

[50] G. Mita, M. Filippone, and P. Michiardi. An identifiable double vae for disentangled representations. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7769–7779. PMLR, 18–24 Jul 2021.

[51] G. E. Moran, D. Sridhar, Y. Wang, and D. M. Blei. Identifiable variational autoencoders via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021.

[52] E. Nalisnick, L. Hertel, and P. Smyth. Approximate inference for deep latent gaussian mixtures. In *NIPS Workshop on Bayesian Deep Learning*, volume 2, page 131, 2016.

[53] H. D. Nguyen and G. McLachlan. On approximations via convolution-defined mixture models. *Communications in Statistics-Theory and Methods*, 48(16):3945–3955, 2019.

[54] J. Pearl and T. S. Verma. A statistical semantics for causation. *Statistics and Computing*, 2(2): 91–95, 1992.

[55] Z.-Y. Ran and B.-G. Hu. Parameter identifiability in statistical machine learning: a review. *Neural Computation*, 29(5):1151–1203, 2017.

[56] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In E. P. Xing and T. Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR.

[57] A. Ritchie, R. A. Vandermeulen, and C. Scott. Consistent estimation of identifiable nonparametric mixture models from grouped observations. *arXiv preprint arXiv:2006.07459*, 2020.

[58] G. Roeder, L. Metz, and D. Kingma. On linear identifiability of learned representations. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 9030–9039. PMLR, 18–24 Jul 2021.

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

[60] L. Schott, J. von Kügelgen, F. Träuble, P. Gehler, C. Russell, M. Bethge, B. Schölkopf, F. Locatello, and W. Brendel. Visual representation learning does not generalize strongly within the same domain. *arXiv preprint arXiv:2107.08221*, 2021.

[61] P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (GIN). In *International Conference on Learning Representations*, 2019.

[62] P. Stock and R. Gribonval. An embedding of relu networks and an analysis of their identifiability. *arXiv preprint arXiv:2107.09370*, 2021.

[63] H. Teicher. Identifiability of finite mixtures. *The annals of Mathematical statistics*, pages 1265–1269, 1963.

[64] H. Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

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

[66] J. Tomczak and M. Welling. Vae with a vampprior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR, 2018.

[67] A. Van Den Oord, O. Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

[68] R. A. Vandermeulen, C. D. Scott, et al. An operator theoretic approach to nonparametric mixture models. *Annals of Statistics*, 47(5):2704–2733, 2019.

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

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

[71] M. Willetts, S. Roberts, and C. Holmes. Disentangling to cluster: Gaussian mixture variational ladder autoencoders. *arXiv preprint arXiv:1909.11501*, 2019.

[72] Y. Yacoby, W. Pan, and F. Doshi-Velez. Failure modes of variational autoencoders and their effects on downstream tasks. In *ICML Workshop on Uncertainty and Robustness in Deep Learning (UDL)*, 2020.

[73] X. Yang, Y. Wang, J. Sun, X. Zhang, S. Zhang, Z. Li, and J. Yan. Nonlinear ica using volume-preserving transformations. In *International Conference on Learning Representations*, 2021.

[74] K. Zhang and L. Chan. Minimal nonlinear distortion principle for nonlinear independent component analysis. *Journal of Machine Learning Research*, 9(Nov):2455–2487, 2008.

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 5 and Broader Impacts statement in the appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See (P1)-(P3), (F1)-(F4)
   (b) Did you include complete proofs of all theoretical results? [Yes] See Supplementary materials.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplement.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A] All assets are public and open-source.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Broader impacts The ability to identify the true latent generative model for data has both positive and negative implications. On the one hand, it could be used to interpret data generative processes which can be scrutinized for to reduce implicit bias and improve algorithmic fairness. On the other hand, it may have negative consequences if privacy is a concern. Environmental concerns related to training large generative models is also worth considering.

Outline This supplement is organized as follows:

- Appendix A provides additional comparisons to related work, examples, and counterexamples.
- Appendix B provides an outline of the overall proof.
- Appendices C-F contain the main technical proofs.
- Appendix G compares the notion of equivalence in iVAE to affine equivalence.
- Appendix H provides conditions that guarantee a ReLU NN satisfies (F3) and/or (F4).
- Appendix J provides details on the experiments.

Appendices G-H are not required for the main proofs, but are included for completeness and reference.

A Additional details from Section 3

In this appendix, we provide additional detailed comparisons against previous work as well as additional examples and counterexamples to further illustrate our theory.

A.1 Detailed comparisons

Since the original iVAE paper [35], there have been many generalizations and extensions proposed. We pause here to provide a more detailed comparison of our results against this developing literature. For a comparison against iVAE, see Section 2.

We first discuss related work that assumes auxiliary information is available (i.e. U is known), then discuss more recent work that does not assume any auxiliary information; the ensuing comparisons are then presented in alphabetical order.

Assuming auxiliary information is available.

1. [22] achieves identifiability in the fully unsupervised regime for the model in which the latent state is defined by a Hidden Markov Model (HMM). The proof of identifiability in [22] invokes [19] to essentially recover the HMM transition matrix and the auxiliary variable U from X, reducing the problem to [35]. Our Theorem C.1 shows that identifiability in fully unsupervised regime is possible even without additional structure given here by the time-dependency according to Markov dynamics.
2. [36] extend [35] by observing that the conditional independence $Z_i \perp\!\!\!\perp Z_j \mid U$ is not required for identifiability, so they propose a more general IMCA framework for conditional energy-based models. However, identifiability in [36] still critically relies on observing an auxiliary variable (in their setting, this is a dependent variable Y). Our Theorem C.1 achieves same type of identifiability as [36] (up to affine transformation) without relying on conditional independence or an auxiliary variable.
3. [61] extends the iVAE identifiability theory of [35] by showing that a stronger notion of identifiability can be achieved if Z is distributed according to factorial GMM (instead of a general exponential family as in [35]). More specifically, given the auxiliary information U, they show that Z can be recovered up to permutation and scaling of the variables Z_i. By contrast, in Theorem E.4, we show that under similar assumptions Z_i are identifiable up to permutation and scaling and importantly, we do this only from X, without using U in any way. We also do not require the GMM to be factorial. Finally, our proof technique is different: While [61] relies on [35] (and hence, for instance, require $J \geq m+1$), our proof is independent of [35].

4. [73] studies identifiability of the model (3) under the assumption that f is volume preserving and Z comes from a conditionally factorial exponential family, similar to iVAE. They prove that if U is known, (P2) holds, and f is twice differentiable, then Z is identifiable up to permutation and non-linear functions applied to each Z_i (i.e., $Z_i = h_i(Z_{\tau(i)})$). If additionally Z is a GMM, then Z can be recovered up to permutation, scaling, and translation. In comparison, we do not require U to be known, and we do not require f to be volume preserving or even differentiable everywhere. We show that under the same assumption (P2) the latent variables Z_i can be recovered up to permutation, scaling, and translation if f is only assumed to be piecewise affine. Additionally, we show that a weaker notion of identifiability holds if Z is not assumed to be conditionally factorial.
5. [75] considers a contrastive model in which samples arrive in pairs, which is a type of weak supervision. Additionally, it is assumed that the latent variables are sampled uniformly from a convex body, and that f is differentiable and injective. By comparison, our model allows for more general non-uniform mixture priors, non-injective and non-smooth f, and is fully unsupervised.

No auxiliary information.

1. [18] propose a novel Multifacet VAE (MFCVAE) model for unsupervised deep clustering. Their model has the following form

$$p(x, z, u) = p(x|z) \prod_{j=1}^{k} p(z_j|u_j)p(u_j), \quad z_j|u_j \sim \mathcal{N}(\mu_{u_j}, \Sigma_{u_j}) \tag{5}$$

   Through empirical experiments, [18] emphasizes the importance of high-dimentional structure of U and shows how it results in improved clustering performance. The key idea is that while the number of meaningful clusters in the data may be very large, there may be meaningful individual categorical variables U_i ("facets") with a much smaller number of states, which may be easier to learn. In this way, by simultaneously performing clustering for each "facet" U_i one can learn meaningful fine-grained clusters in the data. Note that k binary variables U_i result in $J = 2^k$ fine-grained clusters in the data.

   Compared to our work, [18] is focused on practical implementation details, and lacks a formal identifiability theory. In fact, our results provide precisely such a formal identifiability theory in a more general setting. If $p(x|z)$ is modeled by ReLU/leaky-ReLU NN, MFCVAE is a special case of our model (3) with high-dimensional U. More specifically, the MFCVAE model (5) restricts our model (3) to the case when u_i are independent and $\mathrm{ne}(U_i) = \{i\}$. In particular, it satisfies assumption (P3). Therefore, Theorem 3.10 implies that for MFCVAE with diagonal covariances Σ_{u_j}, $\dim(U), \dim(U_j)$, $P(U)$ are identifiable from $P(X)$ up to a permutation of U, and $P(Z)$ is identifiable up to permutation, scaling, and/or translation.
2. [38] establishes the identifiability of latent representations for non-parametric measurement models $U \to X$. Their result crucially relies on the fact that observed variables are conditionally independent $X_i \perp\!\!\!\perp X_j \mid U$. Our Theorem F.2 significantly generalizes this result, by showing the same guarantees for the model (3) that allows arbitrarily complex dependencies between the observed variables X.
3. [51] propose a sparse VAE and prove that the latent space of this model is identifiable. Similar to [69], identifiability of f is not addressed. Their identifiability results also assume an anchor feature assumption, which we do not require. Even our strongest assumption (P3) is weaker compared to the anchor feature assumption (see Remark 3.8). Moreover, we do not require any sparsity assumptions.
4. [69] propose LIDVAE as a way to identify the latent space of a VAE without auxiliary information, however, their approach only guararantees identifiability of $P(Z)$, and does not address f (this is acknowledged by the authors in their discussion as an open question). By restricting f to be a Brenier map, they guarantee that the likelihood is injective, which leads to identifiability of $P(Z)$. Compared to [69] our work restricts f in a different way (i.e. by an injective ReLU network), which matches common practice. Moreover, we show that both f and the multivariate U structure (i.e. in addition to $P(Z)$) are identifiable under mild additional assumptions.

A.2 Special cases

Our main results contain some notable special cases that warrant additional discussion.

Classical VAE The classical, vanilla VAE [37, 56] with an isotropic Gaussian prior is equivalent to (3) with $J = 1$. In this case, U is trivial and the Gaussian distribution $P(Z)$ can be transformed by an affine map to a standard isotropic Gaussian $\mathcal{N}(0, I)$. In this case, Theorem 3.9(c) shows that f is identifiable from $P(X)$ up to an orthogonal transformation. In fact, this case can readily be deduced from known results on the identifiability of ReLU networks, e.g. [62].

Although the $J = 1$ case is already identifiable, there are clear reasons to prefer a clustered latent space: It is natural to model data that has several clusters by a latent space that has similar clusters (e.g. Figure 2). Although in principle any distribution can be approximated by $f(Z)$ where $Z \sim \mathcal{N}(0, I)$ and f is piecewise affine, such f is likely to be extremely complex. At the same time, the same distribution may have a representation with Z being a simple GMM and f being a simple piecewise affine function. Clearly, the latter representation is preferable to the former and can likely be more robustly learned in practice. This is consistent with previous empirical work [16, 18, 32, 33, 41, 44, 71].

Linear ICA In classical linear ICA [11], we observe $X = AZ$, where Z is assumed to have independent components. Compared to the general model (1), this corresponds to the special case where f is linear and $\varepsilon = 0$. In our most general setting under (F2) only, our results imply that $P(Z)$ can be recovered up to an affine transformation *without* assuming independent components, which might seem surprising at first. This is, however, easily explained: In this case, X is also a GMM, and hence $P(Z)$ can already be trivially recovered up to the affine transformation $z \mapsto Az$. This follows from well-known identifiability results for GMMs [63]. This provides some intuition to how the mixture prior assumption (P1) helps to achieve identifiability.

Nonlinear ICA In classical nonlinear ICA, one assumes the model (1) with (a) no assumptions on f and (b) independence assumptions in the latent space. It is well-known that this model is nonidentifiable [28]. Our problem setting is distinguished from the classical nonlinear ICA model via assumptions (P1)-(F1). While we do not require the Z_i to be mutually independent, we impose assumptions on the form of f. It is precisely this inductive bias that allows us to recover identifiability. As a result, our identifiability theory does not contradict known results such as the Darmois construction [15] discussed in [28].

A.3 Counterexamples

A natural question is whether or not the mixture prior (P1) or the piecewise affine nonlinearity (F1) can be relaxed while still maintaining identifiability. In fact, it is not hard to show this is not possible: If either (P1) or (F1) is broken, then the model (1) becomes nonidentifiable. Of course, this is entirely expected given known negative results on nonlinear ICA [28].

Example A.1. If f is allowed to be arbitrary, but (P1) is still enforced, then (1) is no longer identifiable: Pick any two GMMs $P = \sum_{j=1}^{J} \lambda_j N(\mu_j, \Sigma_j)$ and $P' = \sum_{j=1}^{J'} \lambda'_j N(\mu'_j, \Sigma'_j)$. Then we can always find a function g such that $g_\sharp P' = f_\sharp P$ (e.g. use the inverse CDF transform), and $g \neq f$.

Example A.2. If $P(Z)$ is allowed to be arbitrary, but (F1) is still enforced, then (1) is no longer identifiable: Consider any two arbitrary piecewise affine, injective functions $f, g : \mathbb{R}^m \to \mathbb{R}^m$. Then almost surely the preimages $f^{-1}(\{x\})$ and $g^{-1}(\{x\})$ will not be equivalent up to an affine transformation. In other words, fixing $P(X)$, we can find models (f, P) and (g, P') such that $f_\sharp P = P(X) = g_\sharp P'$, but f is not equivalent to g (i.e. up to any affine transformation).

B Proof outline

We will prove the main results by breaking the argument into four phases:

1. (Appendix C) First, we show that if f is weakly injective, then $\mathbb{P}(Z)$ is identifiable (Theorem C.1). The proof involves a novel result on identifiability of a nonparametric mixture model (Theorem C.2) that may be of independent interest.

2. (Appendix D) Second, we show that if f is continuous and injective, then f is identifiable up to an affine transformation (Theorem D.1). This result strengthens existing identifiability results in nonlinear ICA by exploiting the mixture prior, which is crucial in the sequel.
3. (Appendix E) Next, we show that if Z is conditionally factorial GMM, then under mild generic assumptions, the individual variables Z_i can be recovered (up to permutation, scaling and translation) (Theorem E.1).
4. (Appendix F) Finally, since for conditionally factorial Z we are able to recover the individual variables Z_i, we show how we can apply the theory developed in [38] to recover the multivariate discrete latent variable U, its dimension, domain sizes of each U_i and $\Pr(U, Z)$ (Theorem F.2). Since we can only recover Z up to permutation, scaling and translation, the results from [38] cannot be applied directly, and we show how to perform this recovery under an unknown affine transformation.

Each of these phases tackles a particular level of the identifiability hierarchy described in the main theorems. A detailed proof outline of each main theorem is provided below; technical proofs can be found in the subsequent appendices.

A notable difference between Theorems 3.9 ($k = 1$) and 3.10 ($k > 1$) is the conclusion in the latent space: Theorem 3.9 identifies $P(U, Z)$ jointly whereas Theorem 3.10 identifies $P(Z)$ and $P(U)$ separately. The reason is simple: If U is 1-dimensional, i.e., $k = 1$, then $P(U, Z)$ for (3) is trivially identifiable from $P(Z)$, since $P(Z)$ is assumed to be a GMM by (P1). Indeed, since finite mixture of Gaussians are identifiable, we can recover $P(U = u)$ and $P(Z \mid U = u)$ as mixture weights and corresponding Gaussian components. This extends to more general exponential mixtures as in Remark 2.1, see Barndorff-Nielsen [8] for details.

When $k > 1$, the situation is considerably more nontrivial, as one also needs to learn the high-dimensional structure of U.

Proof of Theorem 3.9. We assume $\varepsilon = 0$ without any loss of generality; i.e. it is sufficient to consider the noiseless case. This follows from a standard deconvolution argument as in Khemakhem et al. [36] (see Step I of the proof of Theorem 1).

(a) By Theorem C.1, $\mathbb{P}(Z)$ is identifiable up to an affine transformation. Moreover, as described above, we can identify $\mathbb{P}(U, Z)$ from $\mathbb{P}(Z)$.

(b) Since $P(Z)$ is identifiable up to an affine transformation by part a), claim follows from Theorem E.1.

(c) By Theorem D.1, f is identifiable. □

Proof of Theorem 3.10. As with Theorem 3.9, we assume $\varepsilon = 0$ without loss of generality.

(a) By Theorem C.1, $\mathbb{P}(Z)$ is identifiable up to an affine transformation.

(b) Since $P(Z)$ is identifiable up to an affine transformation by part a), by Theorem E.1, Z_i are identifiable up to permutation, scaling and translation.

(c) Follows from Theorem F.2.

(d) By Theorem D.1, f is identifiable. □

C Identifiability of Z up to an affine transformation via nonparametric mixtures

In this section we prove that if in model (3) the function f is weakly injective, then Z is identifiable up to an affine transformation. More specifically, we prove the following:

Theorem C.1. *Assume that (U, Z, X) are distributed according to model (3). If f is weakly injective (see (F2) in Definition 3.2), then $\mathbb{P}(U, Z)$ is identifiable from $\mathbb{P}(X)$ up to an affine transformation.*

We will prove this result by first proving a result on identifiability of nonparametric mixtures that may be of independent interest.

Theorem C.2. *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be piecewise affine functions satisfying (F2). Let $Y \sim \sum_{i=1}^{J} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$ and $Y' \sim \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j)$ be a pair of GMMs (in reduced form). Suppose that $f(Y)$ and $g(Y')$ are equally distributed.*

Then there exists an invertible affine transformation $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $h(Y) \equiv Y'$, i.e., $J = J'$ and for some permutation $\tau \in S_J$ we have $\lambda_i = \lambda'_{\tau(i)}$ and $h_\sharp \mathcal{N}(\mu_i, \Sigma_i) = \mathcal{N}(\mu'_{\tau(i)}, \Sigma'_{\tau(i)})$.

In other words, a mixture model whose components are piecewise affine transformations of a Gaussian is identifiable. To see this more clearly, observe that

$$\sum_{j=1}^{J} \lambda_k f_\sharp \mathcal{N}(\mu_k, \sigma_k) \sim f_\sharp \Big(\sum_{j=1}^{J} \lambda_k \mathcal{N}(\mu_k, \sigma_k) \Big).$$

To the best of our knowledge, this identifiability result for a nonparametric mixture model is new to the literature. In Theorem C.2, the transformation and number of components is allowed to be unknown and arbitrary, and no separation or independence assumptions are needed.

C.1 Technical lemmas

We recall that a m-dimensional Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ with covariance Σ and mean μ has the following density function

$$p(x) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp\Big((-1/2)(x-\mu)^T \Sigma^{-1} (x-\mu) \Big). \tag{6}$$

We assume that all Gaussian components are *non-degenerate* in the sense that Σ is positive definite. We also recall that if $Y \sim \mathcal{N}(\mu, \Sigma)$ and $Y' = AY + b$ for an invertible $A \in \mathbb{R}^{m \times m}$ and $b \in \mathbb{R}^m$, then $Y' \sim \mathcal{N}(A\mu + b, A\Sigma A^T)$.

Definition C.3. We say that a Gaussian mixture distribution

$$P = \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j) \tag{7}$$

is in reduced form if $\lambda_j > 0$ for every $j \in [J]$ and for every $i \neq j \in [J]$ we have $(\mu_i, \Sigma_i) \neq (\mu_j, \Sigma_j)$.

In the proofs we use the notion of real analytic functions. We remind the definition for reader's convenience.

Definition C.4. Let $D \subseteq \mathbb{R}^n$ be an open set. A function $f : D \to \mathbb{R}$ is called a (real) analytic function if for every compact $K \subset D$ there exists a constant $C > 0$ such that for any $\alpha \in \mathbb{N}^n$ we have

$$\sup_{x \in K} \left| \frac{\partial^\alpha f}{\partial x^\alpha}(x) \right| \leq \alpha! C^{|\alpha|+1}. \tag{8}$$

Alternatively, a real analytic function $f : D \to \mathbb{R}$ can be defined as a function that has a Taylor expansion convergent on D.

It is a standard fact that a linear combination and a product of analytic functions are analytic, and it is well-known that the density of the multivariate Gaussian is a real analytic function on $\mathbb{R}^m$. We will also need the standard notion of analytic continuation:

Definition C.5. Let $D_0 \subseteq D \subseteq \mathbb{R}^n$ be open sets. Let $f_0 : D_0 \to \mathbb{R}$. We say that an analytic function $f : D \to \mathbb{R}$ is an *analytic continuation* of f_0 onto D if $f(x) = f_0(x)$ for every $x \in D_0$.

Definition C.6. Let $x_0 \in \mathbb{R}^m$ and $\delta > 0$. Let $p : B(x_0, \delta) \to \mathbb{R}$. Define

$$\mathrm{Ext}(p) : \mathbb{R}^m \to \mathbb{R} \tag{9}$$

to be the unique analytic continuation of p on the entire space $\mathbb{R}^m$ if such a continuation exists, and to be 0 otherwise.

Definition C.7. Let $D_0 \subset D$ and $p : D \to \mathbb{R}$ be a function. We define $p|_{D_0} : D_0 \to \mathbb{R}$ to be a restriction of p to D_0, namely a function that satisfies $p|_{D_0}(x) = p(x)$ for every $x \in D_0$.

Theorem C.8. *Consider a pair of finite GMMs (in reduced form) in $\mathbb{R}^m$*

$$P = \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j) \quad \textit{and} \quad P' = \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j). \tag{10}$$

Assume that there exists a ball $B(x_0, \delta)$ such that P and P' induce the same measure on $B(x_0, \delta)$. Then $P \equiv P'$, i.e., $J = J'$ and for some permutation τ we have $\lambda_i = \lambda'_{\tau(i)}$ and $(\mu_i, \Sigma_i) = (\mu'_{\tau(i)}, \Sigma'_{\tau(i)})$.

Proof. Follows from the identity theorem for real analytic functions and the identifiability of finite GMMs. □

Definition C.9. Let $f : \mathbb{R}^m \to \mathbb{R}^n$ be a piecewise affine function. We say that a point $x \in f(\mathbb{R}^m) \subseteq \mathbb{R}^n$ is *generic with respect to* f if the preimage $f^{-1}(\{x\})$ is finite and there exists $\delta > 0$, such that $f : B(z, \delta) \to \mathbb{R}^n$ is affine for every $z \in f^{-1}(\{x\})$.

Lemma C.10. *If $f : \mathbb{R}^m \to \mathbb{R}^n$ is a piecewise affine function such that $\{x \in \mathbb{R}^n : |f^{-1}(\{x\})| = \infty\} \subseteq f(\mathbb{R}^m)$ has measure zero with respect to the Lebesgue measure on $f(\mathbb{R}^m)$, then $\dim(f(\mathbb{R}^m)) = m$ and almost every point in $f(\mathbb{R}^m)$ (with respect to the Lebesgue measure on $f(\mathbb{R}^m)$) is generic with respect to f.*

Proof. Let $g_i(z) = Az + b$, $g : D \to \mathbb{R}^n$ be one of the affine pieces defining piecewise affine function f. If A does not have full column rank, then every $x \in g(D)$ has an infinite number of preimages. Therefore, the assumption of the lemma implies that for at least one of the affine pieces g_i, A has full column rank. Thus, $\dim(f(\mathbb{R}^m)) = m$.

Let $S = \{x \in \mathbb{R}^n : |f^{-1}(\{x\})| = \infty\}$ then by assumption S has measure zero in $f(\mathbb{R}^m)$. Let E be the set of points $z \in \mathbb{R}^m$ such that for every $\delta > 0$, f is not affine on $B(z, \delta)$. Since f is piecewise affine, E can be covered by a locally-finite union of $(m-1)$-dimensional subspaces, i.e. every compact set intersects only finitely many of these (potentially infinite) $(m-1)$-dimensional subspaces. Thus E has measure zero. Moreover, since $\dim(f(\mathbb{R}^m)) = m$, $f(E)$ has measure zero in $f(\mathbb{R}^m)$.

Finally, by definition, every $x \in f(\mathbb{R}^m) \setminus \big(S \cup f(E)\big)$ is generic. □

We make the following useful observation.

Lemma C.11. *Consider a random variable Z distributed according to the GMM $\sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j)$. Consider the random variable $X = f(Z)$, where $f : \mathbb{R}^m \to \mathbb{R}^m$ is a piecewise affine function, such that $\dim(f(\mathbb{R}^m)) = m$. Let $x_0 \in \mathbb{R}^m$ be a generic point with respect to f. Let p be the density function of X. Then the number of points in the preimage $f^{-1}(\{x_0\})$ can be computed as*

$$|f^{-1}(\{x_0\})| = \lim_{\delta \to 0} \int_{x \in \mathbb{R}^m} \mathrm{Ext}(p|_{B(x_0,\delta)})(x) dx. \tag{11}$$

Proof. Since x_0 is generic with respect to f, the preimage of x_0 consists of finitely many points, $f^{-1}(\{x_0\}) = \{z_1, z_2, \ldots, z_s\}$, and there exists $\varepsilon > 0$ such that for every $i \in [s]$ there is a well-defined invertible affine function $g_i : B(z_i, \varepsilon) \to \mathbb{R}^m$ such that $g_i(z) = f(z)$ for all $z \in B(z_i, \varepsilon)$.

We can write $g_i(z) = A_i z + b_i$ for some $A_i \in \mathbb{R}^{m \times m}$ and $b_i \in \mathbb{R}^m$. Let $\delta_0 > 0$ be such that

$$B(x_0, \delta_0) \subseteq \bigcap_{i=1}^{s} g_i(B(z_i, \varepsilon)). \tag{12}$$

Let $0 < \delta < \delta_0$. Then, for $\mu'_{ij} = A_i \mu_k + b_i$ and $\Sigma_{ij} = A_i \Sigma_j A_i^T$, and every $x \in B(x_0, \delta)$ we have

$$p|_{B(x_0,\delta)}(x) = \sum_{i=1}^{s} \sum_{j=1}^{J} \frac{\lambda_j}{\sqrt{(2\pi)^m \det \Sigma_{ij}}} \exp\left((-1/2)(x - \mu'_{ij})^T \Sigma_{ij}^{-1} (x - \mu'_{ij})\right). \tag{13}$$

The RHS of (13) is a real analytic function defined on all of $\mathbb{R}^m$ (i.e. it is an entire function) that equals p on an open neighborhood, hence it defines $\mathrm{Ext}(p|_{B(x_0,\delta)})$ on the entire space $\mathbb{R}^m$. Therefore,

$$\begin{aligned}
&\int_{x\in\mathbb{R}^m} \mathrm{Ext}(p|_{B(x_0,\delta)})(x)dx = \\
&= \int_{x\in\mathbb{R}^m} \sum_{i=1}^{s}\sum_{j=1}^{J} \frac{\lambda_j}{\sqrt{(2\pi)^m \det \Sigma_{ij}}} \exp\left((-1/2)(x-\mu'_{ij})^T \Sigma_{ij}^{-1}(x-\mu'_{ij})\right) = \\
&= \sum_{i=1}^{s} \int_{x\in\mathbb{R}^m} \sum_{j=1}^{J} \frac{\lambda_j}{\sqrt{(2\pi)^m \det \Sigma_{ij}}} \exp\left((-1/2)(x-\mu'_{ij})^T \Sigma_{ij}^{-1}(x-\mu'_{ij})\right) = \\
&= s = |f^{-1}(\{x_0\})|. \qquad \square
\end{aligned}$$

We can deduce the following corollary.

Corollary C.12. *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be piecewise affine functions that satisfy (F2).*

Let $Z \sim \sum_{i=1}^{J} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$ and $Z' \sim \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j)$. Suppose that $f(Z)$ and $g(Z')$ are equally distributed. Assume that for $x_0 \in \mathbb{R}^n$ and $\delta > 0$, f is invertible on $B(x_0, 2\delta) \cap f(\mathbb{R}^m)$.

Then there exists $x_1 \in B(x_0, \delta)$ and $\delta_1 > 0$ such that both f and g are invertible on $B(x_1, \delta_1) \cap f(\mathbb{R}^m)$.

Proof. Since f is piecewise affine and f is invertible on $B(x_0, 2\delta) \cap f(\mathbb{R}^m)$, then $\dim f(\mathbb{R}^m) = m$. Note that since $f(Z)$ and $g(Z')$ are equally distributed and since regular GMMs have positive density at every point, we have

$$f(\mathbb{R}^m) = \mathrm{supp}(f(Z)) = \mathrm{supp}(g(Z')) = g(\mathbb{R}^m).$$

Therefore, $\dim(g(\mathbb{R}^m)) = \dim(f(\mathbb{R}^m)) = m$ and, by Lemma C.10, almost every point $x \in B(x_0, \delta) \cap f(\mathbb{R}^m)$ is generic with respect to f and w.r.t to g. Let $x_1 \in B(x_0, \delta)$ be such a point. Since f is invertible on $B(x_1, \delta)$, we have that $|f^{-1}(\{x_1\})| = 1$. Since x_1 is generic with respect to f and with respect to to g, by Lemma C.11, we deduce that $|g^{-1}(\{x_1\})| = 1$. Therefore, since x_1 is generic, there exists $0 < \delta_1 < \delta$ such that on $\big(B(x_1, \delta_1) \cap f(\mathbb{R}^m)\big) \subset \big(B(x_0, 2\delta) \cap f(\mathbb{R}^m)\big)$ the function g is invertible. $\square$

C.2 Identifiability of nonparametric mixtures

First we prove our identifiability theorem under the assumption that f and g are invertible in the neighborhood of the same point.

Theorem C.13. *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be piecewise affine. Let $Z \sim \sum_{i=1}^{J} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$ and $Z' \sim \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j)$ be a pair of GMMs (in reduced form). Suppose that $f(Z)$ and $g(Z')$ are equally distributed.*

Assume that there exists $x_0 \in \mathbb{R}^n$ and $\delta > 0$ such that f and g are invertible on $B(x_0, \delta) \cap f(\mathbb{R}^m)$. Then there exists an invertible affine transformation $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $h(Z) \equiv Z'$, i.e., $J = J'$ and for some permutation τ we have $\lambda_i = \lambda'_{\tau(i)}$ and $h_\sharp \mathcal{N}(\mu_i, \Sigma_i) = \mathcal{N}(\mu'_{\tau(i)}, \Sigma'_{\tau(i)})$.

Proof. Since f and g are piecewise affine and both f and g are invertible on $B(x_0, \delta) \cap f(\mathbb{R}^m)$, then $\dim f(\mathbb{R}^m) = m$ and the inverse functions are piecewise affine. Hence, moreover, there exist x_1 and $\delta_1 > 0$ with $B(x_1, \delta_1) \subseteq B(x_0, \delta)$ such that f^{-1} and g^{-1} on $B(x_1, \delta_1) \subseteq B(x_0, \delta)$ are defined by affine functions.

Let $L \subseteq \mathbb{R}^n$ be an m-dimensional affine subspace, such that $B(x_1, \delta_1) \cap f(\mathbb{R}^m) = B(x_1, \delta_1) \cap L$.

Let $h_f, h_g : \mathbb{R}^m \to L$ be a pair of invertible affine functions such that h_f^{-1} coincides with f^{-1} on $B(x_1, \delta_1) \cap L$ and h_g^{-1} coincides with g^{-1} on $B(x_1, \delta_1) \cap L$. This means that distributions $h_f(Y)$

and $h_g(Y')$ coincide on $B(x_1, \delta_1) \cap L$. Moreover, since h_f and h_g are affine transformations, then $h_f(Y)$ and $h_g(Y')$ are finite GMMs. Therefore, by Theorem C.8, $h_f(Y) \equiv h_g(Y')$. The claim of the theorem holds for $h = h_g^{-1} \circ h_f$. □

Combining this identifiability result with results of Section C.1, we obtain the proof of our main identifiability result for non-parametric mixtures.

Proof of Theorem C.2. By Corollary C.12 there exists $x_0 \in f(\mathbb{R}^m)$ that is generic with respect to to both f and g and $\delta > 0$ such that f and g are invertible on $B(x_0, \delta) \cap f(\mathbb{R}^m)$. Therefore, the result follows from Theorem C.13. □

C.3 Proof of Theorem C.1

We give a proof by contradiction. Assume that there exists another model (U', Z', X') and a piecewise affine function g in model 3 that generates the same distribution, i.e., $\mathbb{P}(X) = \mathbb{P}(X')$.

By Corollary C.12 there exists $x_0 \in f(\mathbb{R}^m)$ that is generic with respect to to both f and g and $\delta > 0$ such that f and g are invertible on $B(x_0, \delta) \cap f(\mathbb{R}^m)$. Therefore, by Theorem C.13, there exists $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $Z' = h(Z)$. In other words, $P(U, Z)$ is identifiable up to an affine transformation.

D Identifiability of f

In this section we show that if f is continuous piecewise affine and injective then it is identifiable from $P(X)$ up to an affine transformation.

Theorem D.1. *Assume that (U, Z, X) are distributed according to model (3). Assume that f is continuous piecewise affine and satisfies (F4) (i.e., f is injective).*

Then $(\mathbb{P}(U, Z), f)$ is identifiable from $\mathbb{P}(X)$ up to an affine transformation.

Before proving this theorem, we provide an example that shows that assumption (F2) does not guarantee that f can be recovered uniquely up to an affine transformation in Theorem C.1.

Example D.2. Consider

$$Y \sim \frac{1}{2}\mathcal{N}(-2, 1) + \frac{1}{2}\mathcal{N}(2, 1) \tag{14}$$

Define a pair of piecewise affine functions (see also Figure 3)

$$f(x) = \begin{cases} x - 4, & \text{for } x \geq 2, \\ -x, & \text{for } -2 \leq x < 2, \\ x + 4, & \text{for } -4 \leq x < -2, \\ (x+4)/5, & \text{for } x < -4. \end{cases} \qquad g(x) = \begin{cases} x - 4, & \text{for } x \geq 4, \\ -x + 4, & \text{for } 2 \leq x < 4, \\ x, & \text{for } -2 \leq x < 2, \\ -x - 4, & \text{for } -4 \leq x < -2, \\ (x+4)/5, & \text{for } x < -4. \end{cases} \tag{15}$$

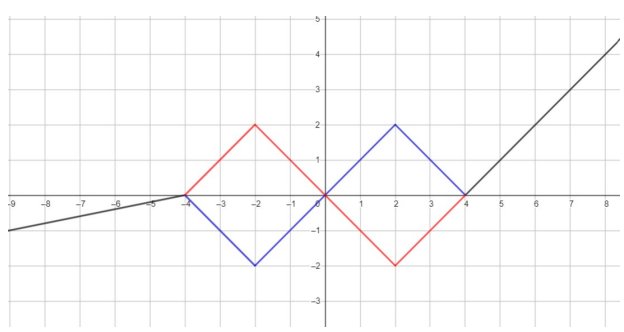


Figure 3: Graphs of f (black and red) and g (black and blue) in Example D.2.

Then it is easy to see that $f(Y)$ and $g(Y)$ have the same distribution, but f cannot be transformed into g by an affine transformation.

In order to prove Theorem D.1 we need to show that for a mixture of Gaussians P and a pair of piecewise affine functions f, g if $f_\sharp P = g_\sharp P$, then $f = h \circ g$ for some invertible affine h. We first consider the case when g is the identity.

Lemma D.3. *Let $Z \sim \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j)$. Assume that $f : \mathbb{R}^m \to \mathbb{R}^m$ is a continuous piecewise affine function such that $f(Z) \sim Z$. Then f is affine.*

Proof. Since Z has positive density at every point and $f(Z) \sim Z$ we must have $\dim f(\mathbb{R}^m) = m$.

If f is not affine, then there exist an $(m-1)$-dimensional affine subspace L, $z_0 \in L$ and $\delta > 0$ such that the following holds: The subspace L divides $B(z_0, \delta)$ into two sets (formally, these are "half-balls") B^+ and B^- such that $f_+(z) := f|_{B^+}(z) = A_1 z + b_1$ and $f_-(z) := f|_{B^-}(z) = A_2 z + b_2$, where $(A_1, b_1) \neq (A_2, b_2)$ and A_1, A_2 are invertible.

Since $f(Z) \sim Z$ we have

$$\{f_+(\mu_1), \dots, f_+(\mu_J)\} = \{\mu_1, \mu_2, \dots, \mu_K\} = \{f_-(\mu_1), \dots, f_-(\mu_J)\}$$

as multisets (i.e. including repetitions). Let $\mu_* = \frac{1}{J} \sum_{j=1}^{J} \mu_j$. Then, since f_+ and f_- are affine we get $f_+(\mu_*) = f_-(\mu_*) = \mu_*$. By translating Y and adjusting f accordingly, we may assume that $\mu^* = 0$. In this case, $b_1 = b_2 = 0$. Moreover, since $f_+(z) = f_-(z)$ for $z \in L$, we get

$$(A_1^{-1} A_2)(z) = z \quad \text{for all } z \in L. \tag{16}$$

Finally, since $f(Y) \sim Y$, we have

$$\{A_1 \Sigma_1 A_1^T, \dots, A_1 \Sigma_J A_1^T\} = \{\Sigma_1, \Sigma_2, \dots \Sigma_J\} = \{A_2 \Sigma_1 A_2^T, \dots, A_2 \Sigma_J A_2^T\},$$

as multisets (i.e. including repetitions). This implies that

$$\prod_{j=1}^{J} \det\left(A_1 \Sigma_j A_1^T\right) = \prod_{j=1}^{J} \det\left(\Sigma_j\right) = \prod_{j=1}^{J} \det\left(A_2 \Sigma_j A_2^T\right).$$

Hence, $\det(A_1)^2 = \det(A_2)^2 = 1$, and $\det(A_1^{-1} A_2)^2 = 1$. By (16), $A_1^{-1} A_2$ is the identity map on L. Let v be a unit vector orthogonal to L (in the direction of B^+). Then we get that either $A_1^{-1} A_2 v = v$, or $A_1^{-1} A_2 v = -v$. In the latter case $A_1(y_0 + (\delta/2)v) = A_2(y_0 - (\delta/2)v)$, which means that f is not injective. This contradicts Lemma C.11. Therefore, we must have $A_1^{-1} A_2 v = v$, and so, by (16), $A_1 = A_2$.

Therefore, $f_+ = f_-$, which contradicts $(A_1, b_1) \neq (A_2, b_2)$. It follows that f must be affine. □

Theorem D.4. *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be continuous invertible piecewise affine functions. Let $Z \sim \sum_{i=1}^{J} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$ and $Z' \sim \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j)$ be a pair of GMMs (in reduced form). Suppose that $f(Z)$ and $g(Z')$ are equally distributed.*

Then there exists an affine transformation $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $h(Z) \equiv Z'$ and $g = f \circ h^{-1}$.

Proof. By Theorem C.13, there exists an invertible affine transformation $h_0 : \mathbb{R}^m \to \mathbb{R}^m$ such that $h_0(Z) = Z'$. Then, $f(Z) \sim g(h_0(Z))$, and since g and h_0 are invertible, we can rewrite this as $Z \sim (h_0^{-1} \circ g^{-1} \circ f)(Z)$. By Lemma D.3, $(h_0^{-1} \circ g^{-1} \circ f)$ is affine, i.e. there exists an invertible affine map h_1 such that

$$h_0^{-1} \circ g^{-1} \circ f = h_1 \quad \Leftrightarrow \quad f = g \circ (h_0 \circ h_1)$$

Hence the claim of the theorem holds for $h = h_0 \circ h_1$. □

Proof of Theorem D.1. Immediately follows from Theorems C.1 and D.4. □

D.1 Identifiability under assumption (F3)

In this section we discuss the case (F3). In particular, show that in (3) under the weaker assumption (F3), f is identifiable up to an affine transformation on the preimage of every connected open set onto which f is injective.

Theorem D.5. *Let $f, g : \mathbb{R}^m \to \mathbb{R}^n$ be continuous piecewise affine functions satisfying (F3).*

Let $Z \sim \sum_{i=1}^{J} \lambda_i \mathcal{N}(\mu_i, \Sigma_i)$ and $Z' \sim \sum_{j=1}^{J'} \lambda'_j \mathcal{N}(\mu'_j, \Sigma'_j)$ be a pair of variables with GMM distribution (in reduced form). Suppose that $f(Z)$ and $g(Z')$ are equally distributed.

Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a connected open set such that f and g are injective onto $\mathcal{D}$. Then there exists an affine transformation $h : \mathbb{R}^m \to \mathbb{R}^m$ such that $h(Z) \equiv Z'$ and $g(z) = (f \circ h^{-1})(z)$ for every $z \in g^{-1}(\mathcal{D})$.

Proof. Similarly, as in the proof of Theorem D.4, by Theorem C.13, there exists an invertible affine transformation $h_0 : \mathbb{R}^m \to \mathbb{R}^m$ such that $h_0(Z) = Z'$. Then, $f(Z) \sim g(h_0(Z))$, and since g is invertible on $\mathcal{D}$ and h_0 is invertible, we can rewrite this as $Z \sim (h_0^{-1} \circ g^{-1} \circ f)(Z)$ on $f^{-1}(\mathcal{D})$. Since f is invertible and continuous piecewise affine, $f^{-1}(\mathcal{D})$ is an open connnected set. Therefore, applying Lemma D.3 on $f^{-1}(\mathcal{D})$, we deduce that $(h_0^{-1} \circ g^{-1} \circ f)$ is affine on $f^{-1}(\mathcal{D})$, i.e. there exists an invertible affine map h_1 such that

$$h_0^{-1} \circ g^{-1} \circ f = h_1 \quad \Leftrightarrow \quad f = g \circ (h_0 \circ h_1) \quad \text{on } f^{-1}(\mathcal{D})$$

Therefore, for $h = (h_0 \circ h_1)$, we have $g(y) = (f \circ h^{-1})(z)$ for every $z \in g^{-1}(\mathcal{D})$. □

Remark D.6. Let f be a continuous piecewise affine function that satisfies (F3). Denote

$$S = \{x \in \mathbb{R}^n \,:\, |f^{-1}(\{x\})| > 1\} \subseteq f(\mathbb{R}^m).$$

Recall that assumption (F3) says that S has measure zero in $f(\mathbb{R}^m)$.

We claim that (F3) implies that for every $x \in S$ in fact $|f^{-1}(\{x\})| = \infty$. Indeed, if for all sufficiently small $\delta > 0$ we have $\dim\big(B(x,\delta) \cap f(\mathbb{R}^m)\big) < m$, then $|f^{-1}(\{x\})| = \infty$ since f is continuous piecewise affine. Otherwise, using Corollary C.10, we get that for every $\delta > 0$ there exists a generic with respect to f point $x_\delta \in B(x,\delta) \cap f(\mathbb{R}^m)$. Assumption (F3) implies that $|f^{-1}(\{x_\delta\})| = 1$ for every x_δ. Therefore, since f is continuous piecewise affine we get that either $|f^{-1}(\{x\})| = 1$ or $|f^{-1}(\{x\})| = \infty$.

E Identifiability of Z up to a permutation, scaling and translation

Under (P2), we have

$$Z \sim \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j), \tag{17}$$

where Σ_j is diagonal for every $j \in [J]$. In the setup of model (3) this just means that $Z_i \perp\!\!\!\perp Z_j \mid U$.

Let $Y = AZ + b$, where $A : \mathbb{R}^m \to \mathbb{R}^m$ is an invertible linear map and $b \in \mathbb{R}^m$. Then Y is also a GMM. We next show how Z may be recovered from Y up to a permutation, scaling, and translation.

Theorem E.1. *Let $J \geq 2$, and $\lambda_j > 0$ for all $j \in [J]$. Let $Z = (Z_1, Z_2, \ldots, Z_m)$ be given by*

$$Z \sim \sum_{j=1}^{J} \lambda_j \mathcal{N}(\mu_j, \Sigma_j) \tag{18}$$

Assume that Σ_j is diagonal for every $j \in [J]$. Let $Y = AZ + b$, where $A : \mathbb{R}^m \to \mathbb{R}^m$ is an invertible linear map and $b \in \mathbb{R}^m$. Moreover, assume that there exist indices $i_1, i_2 \in [J]$, such that all numbers $\left((\Sigma_{i_1})_{tt} \,/\, (\Sigma_{i_2})_{tt} \mid t \in [m]\right)$ are distinct. Given Y, one can recover an invertible linear map $A' : \mathbb{R}^m \to \mathbb{R}^m$, such that $(A')^{-1}A = QD$, where Q is a permutation matrix and D is a diagonal matrix with positive entries.

Remark E.2. The translation b is impossible to recover without stronger assumptions, as b corresponds to an arbitrary translation in the Z space. In other words, choice of b determines the origin in the coordinate space of Z and it can be completely arbitrary.

Remark E.3. A slightly different version of Theorem E.1 under different assumptions appeared in [73]. The main difference is that [73] assumed that f is volume-preserving but nonlinear, whereas we restrict to the general (i.e. not necessarily volume-preserving) linear case.

Proof. Without loss of generality assume $i_1 = 1$ and $i_2 = 2$.

Let Σ_i be the covariance matrices of Z_i and let $\widetilde{\Sigma}_i$ be the covariance matrices of Y_i for $i \in [J]$. Clearly

$$\widetilde{\Sigma}_i = A\Sigma_i A^T \quad \text{for each } i \in [J]. \tag{19}$$

The matrices $\widetilde{\Sigma}_i$ are PSD. Therefore, using SVD we can find PSD matrices V_i, such that for every $i \in [J]$,

$$\widetilde{\Sigma}_i = V_i V_i^T. \tag{20}$$

Moreover, such a decomposition is unique up to an orthogonal matrix, i.e., for every pair of such decompositions $\widetilde{\Sigma}_i = V_i V_i^T = V_i'(V_i')^T$ there exists a unitary matrix R such that $V_i R = V_i'$. Therefore, for every $i \in [J]$ there exists a matrix R_i, such that

$$V_i R_i = A\Sigma_i^{1/2} \tag{21}$$

In particular,

$$V_1 R_1 \Sigma_1^{-1/2} = V_2 R_2 \Sigma_2^{-1/2} \quad \Rightarrow \quad R_1 \left(\Sigma_1^{-1/2}\Sigma_2^{1/2}\right) R_2^{-1} = \left(V_1^{-1} V_2\right) \tag{22}$$

Since R_1 and R_2^{-1} are unitary and $\left(\Sigma_1^{-1/2}\Sigma_2^{1/2}\right)$ is diagonal, they can be determined from the SVD of $V_1^{-1}V_2$. Moreover, they can be determined uniquely up to a permutation matrix since all diagonal entries of $\Sigma_1^{-1/2}\Sigma_2^{1/2}$ are distinct. In other words, using SVD for $\left(V_1^{-1}V_2\right)$ we can find R_1' such that for some permutation matrix P we have

$$V_1 R_1' Q = A\Sigma_1^{1/2}, \quad \text{so, for} \quad A' := V_1 R_1 \quad \text{we have} \quad (A')^{-1} A = Q\Sigma_1^{-1/2}. \tag{23}$$

This concludes the proof. □

As an immediate corollary we can deduce the following theorem from Theorem C.1.

Theorem E.4. *Assume that (U, Z, X) are distributed according to model (3) and that f is weakly injective. Suppose that $Z_i \perp\!\!\!\perp Z_j \mid U$ for all $i \neq j$. Moreover, assume that there exist a pair of states $U = u_1$ and $U = u_2$ such that all $((\Sigma_{u_1})_{tt} \,/\, (\Sigma_{u_2})_{tt} \mid t \in [m])$ are distinct.*

Then $P(U, Z)$ is identifiable from $P(X)$ up to permutation, scaling ans translation of Z_i.

Proof. By Theorem C.1, $\mathbb{P}(Z)$ is identifiable from $\mathbb{P}(X)$ up to an affine transformation. That is, we can reconstruct a random variable Y from $\mathbb{P}(X)$ which satisfies $Y = AZ + b$ for some invertible $A \in \mathbb{R}^{m\times m}$.

Now, by Theorem E.1, we can find A' such that $Z' = (A')^{-1}Y = QDZ + (A')^{-1}b$, where Q is a permutation matrix and D is a diagonal matrix. This means, that we can recover Z up to permutation, shift and scaling of individual variables Z_i. □

F Identifiability of multivariate U structure

When $k = 1$, $P(Z)$ contains all the information about $P(U, Z)$, however, when $k > 1$ (i.e. U is multivariate), this may not be true anymore. It is not even obvious that $P(Z)$ must contain information about the true dimension of U. The distribution $P(U, Z)$ may contain interesting dependencies between individual variables U_i and Z_j.

Previously, [38] studied necessary and sufficient conditions for identifiability of $P(U)$ when Z is observed under the so-called *measurement model*. A key limitation of Kivva et al. [38] is that it

requires the observed variables to be conditionally independent, which is not the case in our setting. Ultimately, this is a consequence of Z being unobserved: Previous work such as Kivva et al. [38] assumes there is only a single layer of hidden variables connected to the observations. In our setting, under (3), we need to recover U from Z, the latter of which is unobserved. As a result, if we can only identify Z up to an affine transformation (e.g., like in Theorem C.1); i.e. we can only recover $Z' = AZ + b$, then it almost surely will not be conditionally factorial. Hence, the results from Kivva et al. [38] cannot be applied directly for weak (e.g., up to affine transformation, or as in 35) notions of identifiability of Z.

Luckily, in Section E, we showed how to recover the true Z from $Z' = AZ + b$. This will enable us to identify $P(U)$ in Theorem 3.10(c). In the remainder of this appendix, we outline these details.

We say that a distribution $\mathbb{P}(U, Z)$ satisfies the *Markov property* with respect to the neighborhoods $\mathrm{ne}(Z_i)$ (cf. Definition 3.7) if

$$\mathbb{P}(U, Z) = \mathbb{P}(U) \prod_i \mathbb{P}(Z_i \mid \mathrm{ne}(Z_i)). \tag{24}$$

Remark F.1. The neighborhoods $\mathrm{ne}(Z_i)$ define a bipartite graph between $(U_1, \ldots, U_k)$ and $(Z_1, \ldots, Z_m)$ that is described in Kivva et al. [38]. Since this graph is not needed for our purposes, we proceed without further mention of this graph. The assumptions below have been re-phrased accordingly.

[38] show that assumptions (L1)-(L4) below are necessary for identifiability of U.

(L1) (No twins) For any $U_i \neq U_j$ we have $\mathrm{ne}(U_i) \neq \mathrm{ne}(U_j)$.

(L2) (Maximality) There is no U' such that:

- (a) $\mathbb{P}(U', Z)$ is Markov with respect to the neighborhoods $\mathrm{ne}(Z_i)$ defined by U';
- (b) U' is obtained from U by splitting a hidden variable (equivalently, U is obtained from U' by merging a pair of vertices);
- (c) U' satisfies Assumption (L1).

(L3) (Nondegeneracy) The distribution over (U, Z) satisfies:

- (a) $\mathbb{P}(U = u) > 0$ for all u.
- (b) For all $Z' \subset Z$ and $u_1 \neq u_2$, $\mathbb{P}(Z' | \mathrm{ne}(Z') = u_1) \neq \mathbb{P}(Z' | \mathrm{ne}(Z') = u_2)$, where u_1 and u_2 are distinct configurations of $\mathrm{ne}(Z')$.

(L4) (Subset condition) For any pair of distinct variables U_i, U_j the set $\mathrm{ne}(U_i)$ is not a subset of $\mathrm{ne}(U_j)$.

We prove the following identifiability result.

Theorem F.2. *Assume that (U, Z, X) are distributed as in (3) and that f satisfies (F2). Assume further that (P2)-(P3) hold and $P(U = u) > 0$ for all u in the domain of U.*

Then $\dim(U) = k$, $\dim(U_j)$, $\mathbb{P}(U, Z)$ are identifiable from $P(X)$ up to a permutation of variables U_i and permutation, scaling and translation of variables Z_i.

Proof. The assumptions of Theorem F.2 are stronger than those of Theorem E.4, so by Theorem E.4, $P(Z)$ is identifiable up to a permutation, scaling and translation of Z.

Combined with the positivity assumption $P(U = u) > 0$, the assumptions (L1)-(L4) are weaker than assumption (P3). Indeed, (P3) (a) is equivalent to (L3) (b); (P3) (c) is equivalent to (L4) and implies (L1); and, finally, (P3) (b) and (c) together imply (L2).

Since Z is identifiable up to a permutation, scaling and translation, $Z_i \perp\!\!\!\perp Z_j \mid U$, and assumptions (L1)-(L4) hold, using [38, Thm 3.2], we deduce that $\dim(U) = k$, $\dim(U_j)$, $\mathbb{P}(U)$, and $\mathrm{ne}(U_i)$ are identifiable up to a permutation of the variables U_i. Finally, by the Markov Property, $\mathbb{P}(U)$, $\mathrm{ne}(U_i)$ for all i, and the fact that $P(Z)$ is a finite GMM (that is identifiable) are sufficient to recover $\mathbb{P}(U, Z)$. □

Remark F.3. As the proof indicates, assumptions (L1)-(L4) are weaker than (P3), so Theorem F.2 implies part (c) of Theorem 3.10.

G Equivalence in iVAE

In this section we compare the equivalence relation up to which iVAE [35] guarantees identifiability and equivalence up to an affine transformation. While iVAE achieves the best possible identifiability under the assumptions they make, we show that identifiability up to an affine transformation is considerably stronger.

G.1 iVAE equivalence relation

Recall that iVAE [35] considers the following model, which differs from (3) by assuming that Z has conditionally factorial exponential family distribution:

$$\left.\begin{aligned} U = u &\sim p(u) \\ [Z \mid U = u] &\sim \prod_{i=1}^{m} \frac{Q_i(z_i)}{C(u)} \exp\left(\sum_{j=1}^{t} T_{i,j}(z_i)\lambda_{i,j}(u)\right) \\ [X \mid Z = z] &\sim f(z) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(v, \sigma^2) \end{aligned}\right\} \implies U \to Z \to X. \tag{25}$$

Here $T_i = (T_1, T_2, \dots T_t)$ are sufficient statistics, Q_i is the base measure and $\lambda_{i,j}$ parameters depending on u. iVAE defines the following equivalence relation:

Definition G.1.

$$(f, T, \sigma) \sim (f', T', \sigma') \quad \Leftrightarrow \quad \exists A, c: \quad T(f^{-1}(\{x\})) = A(T'((f')^{-1}(x) + c, \tag{26}$$

where $A : \mathbb{R}^{mt} \to \mathbb{R}^{mt}$ is an invertible linear map, and $c \in \mathbb{R}^N$.

This type of identifiability allows for essentially any (synchronized) changes to Z and f:

Lemma G.2. *Let $\varphi : \mathbb{R}^m \to \mathbb{R}^m$ be any invertible map. Let $f' = f \circ \varphi$, and $T' = T \circ \varphi$. Then $(f, T, \sigma) \sim (f', T', \sigma)$.*

Moreover, if Z has exponential family distribution with statistics T, then $Z' = \varphi^{-1}(Z)$, has an exponential family distribution with statistics T', and $f(Z) \sim f'(Z')$.

Proof. We have $(f')^{-1} = \varphi^{-1} \circ f^{-1}$, so $T' \circ (f')^{-1} = T \circ f$. Hence $(f, T, \sigma) \sim (f', T', \sigma')$, where in (26) A is the identity map and $c = 0$.

Since Z comes from an exponential family distribution, we can write

$$\mathbb{P}(Z \mid U) = h(Z)g(U)\exp(\lambda(U)T(Z)). \tag{27}$$

Let $Z' = \varphi^{-1}(Z)$. Then by the change of variable formula

$$\mathbb{P}(Z' \mid U) = \Big(h(\varphi(Z)) \det |Jac(\varphi(\bullet))|_{\bullet = \varphi^{-1}(Z)}\Big)\, g(U)\exp(\lambda(U)T(\varphi(Z))), \tag{28}$$

where $Jac(\varphi)$ is the Jacobian of φ. Hence Z' indeed has an exponential family distribution with statistics T'. Clearly, $f'(Z') = (f \circ \varphi \circ \varphi^{-1})(Z) \equiv f(Z)$. □

Remark G.3. In other words, the equivalence relation (26) allows an *arbitrary* (possibly highly nonlinear) change of basis in the latent Z space. In principle, this may indicate, that any meaningful analysis of the Z space in this setup may be challenging.

Remark G.4. As in Khemakhem et al. [35], the additional assumption that Z has a conditionally factorial distribution imposes additional restrictions on φ. In this case, $\varphi : \mathbb{R}^m \to \mathbb{R}^m$ can be any invertible coordinatewise function $\varphi(Z') = (\varphi_1(z'_1), \varphi_2(z'_2), \dots \varphi_2(z'_m))$.

G.2 GMMs give more robust identifiability

The next result was also observed in [61]. We present a slightly simplified proof for completeness.

If $\mathbb{P}(Z|U)$ is a multivariate Gaussian distribution, then the sufficient statistics are given by

$$T_m = (z_1, \dots, z_m, z_1 z_1, z_1 z_2, \dots z_m z_m). \tag{29}$$

Remark G.5. For product measures, there are no cross-terms $z_i z_j$.

Proposition G.6 (61, Appendix B)**.** *Assume that $(T_m, f, \sigma) \sim (T_m, f', \sigma')$, where T_m is defined by (29). Then there exists an invertible linear map $M : \mathbb{R}^m \to \mathbb{R}^m$ and a vector $c \in \mathbb{R}^m$ such that $f^{-1}(\{x\}) = M(f')^{-1}(x) + c$ for every x.*

Proof. Let $z = f^{-1}(\{x\})$ and $z' = (f')^{-1}(x)$. By an assumption of the proposition there exists an invertible matrix $A : \mathbb{R}^{m+m^2} \to \mathbb{R}^{m+m^2}$ such that

$$\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \\ z_1 z_1 \\ z_1 z_2 \\ \vdots \\ z_m z_m \end{pmatrix} = A \begin{pmatrix} z'_1 \\ z'_2 \\ \vdots \\ z'_n \\ z'_1 z'_1 \\ z'_1 z'_2 \\ \vdots \\ z'_m z'_m \end{pmatrix} + b \tag{30}$$

This means that for every i there exists a polynomial p_i of degree at most 2 such that $z_i = p_i(z'_1, \dots, z'_m)$. Assume that for some i, we have $\deg(p_i) = 2$. Then it is easy to verify (say, by using lexicographical order on monomials) that $\deg(p_i^2) = 4$. If z' is defined on an open neighbourhood, we get a contradiction with (30) as z_i^2 can be written as a degree-2 polynomial over variables z'_j. Therefore, every p_i is a polynomial of degree at most 1. But this means that that $z = Mz' + c$ for some matrix M and a vector c. Moreover, since A is invertible, M is invertible as well. □

H Conditions on ReLU Neural Network that guarantee that it is an observable injection

For completeness, in this section we provide simple sufficient conditions on ReLU architectures that guarantee that it is an observable injection (cf. (F3)) and simple sufficient conditions on leaky-ReLU architectures which guarantee that it is injection (cf. (F4)). For a more comprehensive account of identifiability in ReLU networks, see [62].

We recall the definitions of ReLU and leaky-ReLU (with parameter $a > 0,\ a \neq 1$) activation functions

$$\mathrm{ReLU}(x) = \begin{cases} x, & \text{for } x > 0, \\ 0, & \text{for } x \leq 0, \end{cases} \qquad \mathrm{LReLU}(x) = \begin{cases} x, & \text{for } x > 0, \\ a \cdot x, & \text{for } x \leq 0. \end{cases} \tag{31}$$

A standard choice of a for leaky-ReLU is $a = 0.01$.

Definition H.1. Let $\mathrm{Aff}(n_1, n_2)$ denote the set of affine maps $h : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$.

Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a general activation function. For a vector $x \in \mathbb{R}^t$, $\sigma(x)$ is a vector obtained from x by applying σ coordinatewise.

Definition H.2. Let $n_1, n_2, \dots, n_t \geq n_0 = m$ and σ be an activation function. Define

$$\mathcal{F}_\sigma^{n_0,\dots,n_t} = \{h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \dots \sigma \circ h_1 \mid h_i \in \mathrm{Aff}(n_{i-1}, n_i)\} \tag{32}$$

$$\mathcal{F}_\sigma^{m \hookrightarrow n} = \bigcup_{t=1}^{\infty} \bigcup_{n_1, n_2, \dots, n_t \geq n_0,\ n_0 = m,\ n_t = n} \mathcal{F}_\sigma^{n_0,\dots,n_t} \tag{33}$$

Remark H.3. The function families $\mathcal{F}_{\mathrm{ReLU}}^{m \hookrightarrow n}$, $\mathcal{F}_{\mathrm{LReLU}}^{m \hookrightarrow n}$ are genuinely nonparametric: There is no bound on the number of layers.

Remark H.4. In the arguments below we do not rely on the fact that the activation function is the same on every layer, or even the same across the nodes of the same layer. However, we will give proofs only in this case, to simplify the presentation.

Remark H.5. ReLU networks under similar assumptions were also studied in [36].

Lemma H.6. *Let $f = h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \ldots \sigma \circ h_1 \in \mathcal{F}_{\text{ReLU}}^{m \hookrightarrow n}$. Assume that $m = n_0 \leq n_1 \leq \ldots \leq n_t = n$, and $\dim(f(\mathbb{R}^m)) = m$. Then for almost all $y \in f(\mathbb{R}^m)$ there exists δ_y such that f^{-1} is a well-defined affine function on $B(y, \delta_y) \cap f(\mathbb{R}^m)$.*

Proof. We prove the claim by induction on the depth of the NN. If $t = 1$, we have $f = h_1$ and the claim is trivial. Assume that we already proved the lemma for all $t \leq s - 1$. We prove the claim for $t = s$. We can write f as $f = h_t \circ \sigma \circ g$ where $g \in \mathcal{F}_{\text{ReLU}}^{m \hookrightarrow n_{t-1}}$.

Since $\dim(f(\mathbb{R}^m)) = m$, the map h_t has full column rank. Additionally, denoting by $\mathcal{D} = \{x \in \mathbb{R}^{n_{t-1}} \mid x_i > 0, \ \forall i \in [n_{t-1}]\}$ the domain on which σ is injective, we get $g(\mathbb{R}^m) \cap \mathcal{D}$ has positive measure in $g(\mathbb{R}^m)$. Moreover, by the induction assumption, g satisfies conclusion of the lemma, i.e., there exists a set S of measure 0 in $g(\mathbb{R}^m)$ such that for any $y \in g(\mathbb{R}^m) \setminus S$ there exists a $\delta_y > 0$ such that g^{-1} is a well-defined affine function on $B(y, \delta_y) \cap g(\mathbb{R}^m)$. Since h_t has full column rank, f^{-1} is a well-defined affine function on $B(x, \delta_x) \cap f(\mathbb{R}^m)$ for every $x = (f \circ \sigma)(y)$ where $y \in \big(g(\mathbb{R}^m) \setminus S\big) \cap \mathcal{D}$. Clearly, such x form a set of full measure in $f(\mathbb{R}^m)$. □

Corollary H.7. *Let $f = h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \ldots \sigma \circ h_1 \in \mathcal{F}_{\text{ReLU}}^{m \hookrightarrow n}$. Assume that $m = n_0 \leq n_1 \leq \ldots \leq n_t = n$, and $\dim(f(\mathbb{R}^m)) = m$, then f satisfies (F3).*

Proof. Immediately follows from Lemma H.6. □

Lemma H.8. *Let $f = h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \ldots \sigma \circ h_1 \in \mathcal{F}_{\text{LReLU}}^{m \hookrightarrow n}$. Assume that $m = n_0 \leq n_1 \leq \ldots \leq n_k = n$ and every h_i is invertible. Then for almost all $y \in f(\mathbb{R}^m)$ there exists δ_y such that f^{-1} is a well-defined affine function on $B(y, \delta_y) \cap f(\mathbb{R}^m)$.*

Proof. Clearly, any $f = h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \ldots \sigma \circ h_1 \in \mathcal{F}_{\text{LReLU}}$ is a piecewise affine function. The LReLU activation function is invertible, so f is invertible. Finally, since f is a piecewise affine transformation, for almost all $y \in B(x_0, \delta)$ there exists δ_y such that f^{-1} is an affine function on $B(y, \delta_y)$. □

Corollary H.9. *Let $f = h_t \circ \sigma \circ h_{t-1} \circ \sigma \circ \ldots \sigma \circ h_1 \in \mathcal{F}_{\text{LReLU}}^{m \hookrightarrow n}$. Assume that $m = n_0 \leq n_1 \leq \ldots \leq n_t = n$, then generically f satisfies (F4).*

Proof. Generically, every h_i has full column rank, and so is injective. Since LReLU is injective, we get that f is injective. □

We conclude with an example of a very simple LReLU NN that is not even weakly injective.

Example H.10. Let $\sigma(x) = x$ for $x \geq 0$ and $\sigma(x) = x/2$ for $x < 0$. Let $h_1 : \mathbb{R} \to \mathbb{R}^2$ defined as $h_1(x) = (x, -x)$. Then $\sigma \circ h_1(x) = (x, -x/2)$ if $x \geq 0$ and $\sigma \circ h_1(x) = (x/2, -x)$ if $x < 0$. Let $h_2 : \mathbb{R}^2 \to \mathbb{R}^2$ given by

$$h_2 = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

Then $(h_2 \circ \sigma \circ h_1)(x) = (3x/2, x/2)$ for $x \geq 0$ and $(h_2 \circ \sigma \circ h_1)(x) = (3x/2, -x/2)$ for $x < 0$ (see Figure 4). Let $h_3(x, y) = y$. Then $f(x) := (h_3 \circ \sigma \circ h_2 \circ \sigma \circ h_1)(x) = |x|/2$. By Remark 3.12, this implies that f is not invertible at every point except 0.

J Experiment details

J.1 Metrics

Previous work has relied on the Mean Correlation Coefficient (MCC) as a metric to quantify identifiability. For consistency with previous work, we report this metric, but also propose a new metric to quantify identifiability up to an affine transformation. There are two challenges in designing such a metric: Firstly, for two Gaussian mixtures, standard distance metrices such as TV-distance or KL-divergence do not have a closed form. Secondly, we need to find an affine map A that best aligns a pair of Gaussian mixtures. Therefore, developing a metric to quantify identifiability up to an affine transformation has natural challenges. We propose $\text{dist}_{\text{Aff},L2}$, defined below, as an additional metric in this setting.

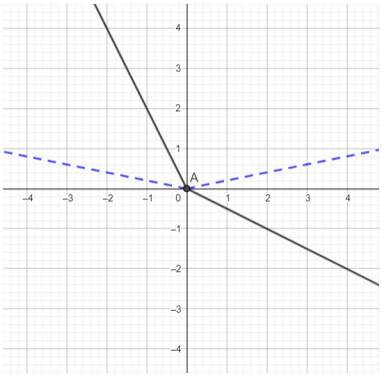


Figure 4: Graphs of $\sigma \circ h_1$ (black) and $h_2 \circ \sigma \circ h_1$ (blue) in Example H.10

.

Measuring loss In this work, we consider two different metrics. For a pair of distributions p_1, p_2, we define $\mathrm{dist}_{\mathrm{Aff},L2}$ loss as

$$\mathrm{dist}_{\mathrm{Aff},L2}(p_1, p_2) = \min_{\substack{A:\mathbb{R}^m \to \mathbb{R}^m, \\ \text{affine}}} \Delta_{L_2}(A_\sharp p_1, p_2), \quad \text{where} \quad \Delta_{L_2}(p_1, p_2) = \frac{\|p_1 - p_2\|_{L_2}}{\|p_1\|_{L_2}^{1/2} \|p_2\|_{L_2}^{1/2}} \tag{34}$$

The other metric we consider is the Mean Correlation Coefficient (MCC) metric which had been used in prior works [36, 70]. See Khemakhem et al. [36, Appendix A.2] for a detailed discussion. There are two versions of MCC that have been used:

- The *strong* MCC is defined to be the MCC before alignment via the affine map A.
- The *weak* MCC is defined to be the MCC after alignment.

In our experiments, we report both the strong MCC and weak MCC. Moreover, all reported MCCs are out-of-sample, i.e. the optimal affine map A is computed over half the dataset and then reused for the other half of the dataset.

Alignment To find the affine map A that best aligns the two GMMs, we use two approaches. One approach is to use Canonical Correlation Analysis (CCA) as was done in prior works in computing MCC.

We describe an alternative approach now. Given two GMMs, we iterate over all permutations of the components and for each fixed permutation, we find the best map A that maps the components accordingly. In an ideal setting, we would want to find A to align not just the means but also the covariance matrices but unfortunately this is a challenging optimization problem. Therefore, we instead find A that maps the means of the first GMM to the means of the second GMM. The map A can be found by solving a least-squares optimization problem which is straightforward using a Singular Value Decomposition (SVD). In practice, we find that this technique of matching the means works well.

J.2 Implementation

For VaDE [32], we use the implementation available at `https://github.com/mperezcarrasco/Pytorch-VaDE`. For MFCVAE [18], we use the author implementation available at `https://github.com/FabianFalck/mfcvae`. For iVAE [35], we use the implementation available at `https://github.com/MatthewWilletts/algostability`. Experiments were performed on an NVIDIA Tesla K80 GPU with 12GB memory.

J.3 Setup

Our experiments consist of three different setups, designed to probe different aspects of identifiability. First, we checked the exact log-likelihood for a unique global minimizer on simple toy models

(Appendix J.3.1). We then used VaDE [32] to train a practical VAE on a simulated dataset where the ground truth latent space is known (Appendix J.3.2). Finally, we compared the performance of MFCVAE [18] against iVAE on MNIST (Appendix J.3.3). The last experiment is based on previous work by [70] that compares iVAE to VaDE; we successfully replicated these experiments using MFCVAE as an additional baseline that closely aligns with our assumptions.

The fact that our theory closely aligns with and replicates existing empirical work illustrates that the model (3) is not merely a theoretical curiosity, but in fact practically relevant in modern applications. In our view, this is a significant advantage compared to related work.

J.3.1 Maximum likelihood

We simulated random models of the form (1) as follows:

1. Fix $J = 2$ or $J = 3$;
2. Randomly select $(\lambda_1, \ldots, \lambda_J)$ from a uniform grid by discretizing the simplex;
3. Randomly select $(\mu_1, \ldots, \mu_J)$ from a uniform grid on the hypercube;
4. Randomly select coefficients (α_1, α_2), weights (β_1, β_2), and biases (π_1, π_2) from a uniform grid on the hypercube.

Given these parameters, the prior $P(Z)$ is defined as in (2) and the decoder f is defined to be the following single-layer ReLU network

$$f(z) = \alpha_1 \operatorname{ReLU}(\beta_1 z + \pi_1) + \alpha_2 \operatorname{ReLU}(\beta_2 z + \pi_2).$$

As a result of the simulation mechanism, the following important cases of misspecification naturally arise:

- We allow $\lambda_j = 0$, i.e. the model allows for $J = 3$ components, but the true model only has two nontrivial components.
- We allow $\alpha_j = 0$ and $\beta_j = 0$, i.e. the model allows for up to two neurons in the hidden layer, but the true model only has one nontrivial neuron.
- f is not forced to be injective or even weakly injective, i.e. assumptions (F2)-(F4) are not checked explicitly.

After generating a pair $(f, P(Z))$, the exact negative log-likelihood is approximated via numerical integration. An exhaustive grid search is performed over all parameters to identify the global minimizers. The computational cost of this step limited the complexity of the models that could be tested, hence the restriction to simple toy models in this experiment. In all runs, the ground truth was the unique global minimizer of the negative log-likelihood, as predicted by our theory. Since the problem is nonconvex, there often exist additional (non-global) local minima (see e.g. Figure 1), however, the global minimizer is always unique up to affine equivalence. That is, due to affine equivalence, in some cases there is more than one global minimizer, but in all such cases it is easy to check that the different minimizers are indeed affinely equivalent. Multiple minimizers also arise when certain parameters (e.g. λ_j or α_j) vanish, again, these are easily checked.

J.3.2 Simulated data

We consider 4 synthetic datasets described below: Pinwheel and three different copies of the “Random parallelograms" dataset

See Section 4 for results of the simulated experiments on the “pinwheels” dataset (see 33). In those experiments we use 5000 samples and set $m = n = 2$. In that experiment we used the same neural network architecture as discussed below for “Random parallelograms".

We simulate an artificial dataset “Random parallelograms” as follows: We generate 3 randomly oriented parallelograms in the plane. After that, an n-dimensional observed distribution is obtained by sampling points uniformly at random from these parallelograms and by adding Gaussian noise to every sampled point.

We fit VaDE to each (observed) dataset 5 times (see Figures 2, 5-7). Let $Z^{(1)}, Z^{(2)}, \ldots, Z^{(5)}$ be the learned latent spaces. For every pair $Z^{(i)}, Z^{(j)}$ we evaluate the MCC and $\mathrm{dist}_{\mathrm{Aff},L2}$ loss. We report means of the MCCs/losses and their standard deviations in Table 2.

For the VaDE training, we use a sequential neural network architecture with LeakyReLU activations for the encoder, with four fully connected layers of the following dimentions: $n \to 64 \to 512 \to 64 \to m$. For the decoder, we use a sequential neural network architecture with LeakyReLU activations, with four fully connected layers of the following dimentions: $m \to 64 \to 512 \to 512 \to n$. We pretrain the autoencoder for 15 epochs and then run VaDE training for 20 epochs.

In all experiments with simulated data we set $m = 2$. We set the number of observed samples to be 5000.

Dataset	$\mathrm{dist}_{\mathrm{Aff},L2}$	Strong MCC	Weak MCC
Random parallelograms #1	0.1542 (0.150)	0.86 (0.09)	0.99 (0.003)
Random parallelograms #2	0.1231 (0.076)	0.83 (0.12)	0.99 (0.003)
Random parallelograms #3	0.578 (0.301)	0.91 (0.08)	0.99 (0.001)

Table 2: Mean (std) $\mathrm{dist}_{\mathrm{Aff},L2}$ distance (lower is better) and Mean (std) MCC (higher is better) for synthetic data

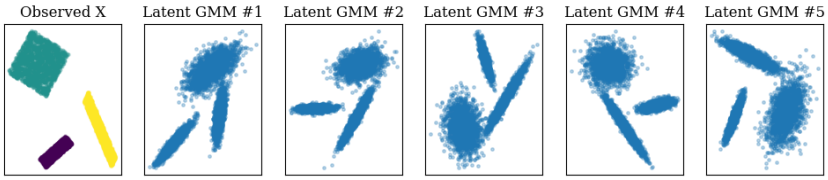


Figure 5: Recovered latent spaces for 5 runs of VaDE on "Random parallelograms" dataset #1 with 3 clusters

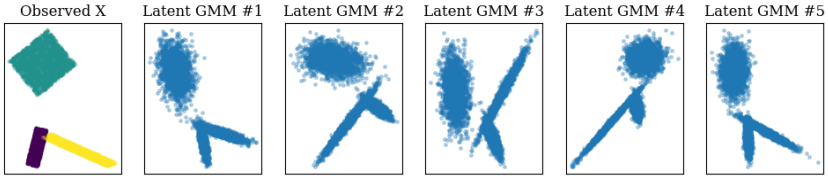


Figure 6: Recovered latent spaces for 5 runs of VaDE on "Random parallelograms" dataset #2 with 3 clusters

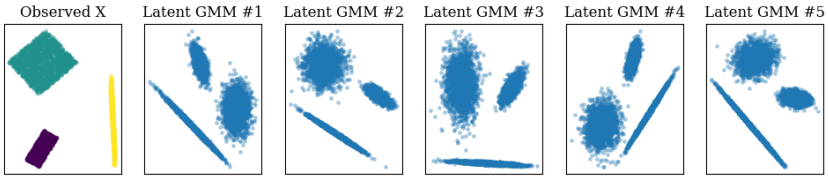


Figure 7: Recovered latent spaces for 5 runs of VaDE on "Random parallelograms" dataset #3 with 3 clusters

J.3.3 Real data

We run MFCVAE [18] on the MNIST dataset 10 times with different initializations. For all the 45 pairs of runs, we compute the strong MCC (before alignment) and weak MCC (after alignment with CCA of dimension 5). For these experiments, we omit the $\mathrm{dist}_{\mathrm{Aff},L2}$ metric since it's computationally infeasible with a large number of components. The mean and standard deviation of the MCCs are

reported in Table 3. As a baseline, we also report the same metrics for 10 runs of iVAE [35] on identical architecture and latent dimension, but recall that iVAE has additional access to the true digit labels U.

Architecture	Model	Activation	Strong MCC	Weak MCC
Arch1	MFCVAE	ReLU	0.7 (0.07)	0.91 (0.05)
	MFCVAE	LeakyReLU	0.69 (0.06)	0.94 (0.02)
	iVAE	LeakyReLU	0.65 (0.07)	0.88 (0.07)
Arch2	MFCVAE	ReLU	0.69 (0.07)	0.89 (0.08)
	MFCVAE	LeakyReLU	0.69 (0.06)	0.92 (0.03)
	iVAE	LeakyReLU	0.64 (0.07)	0.87 (0.04)
Arch3	MFCVAE	ReLU	0.69 (0.07)	0.86 (0.08)
	MFCVAE	LeakyReLU	0.70 (0.05)	0.92 (0.03)
	iVAE	LeakyReLU	0.67 (0.06)	0.87 (0.05)

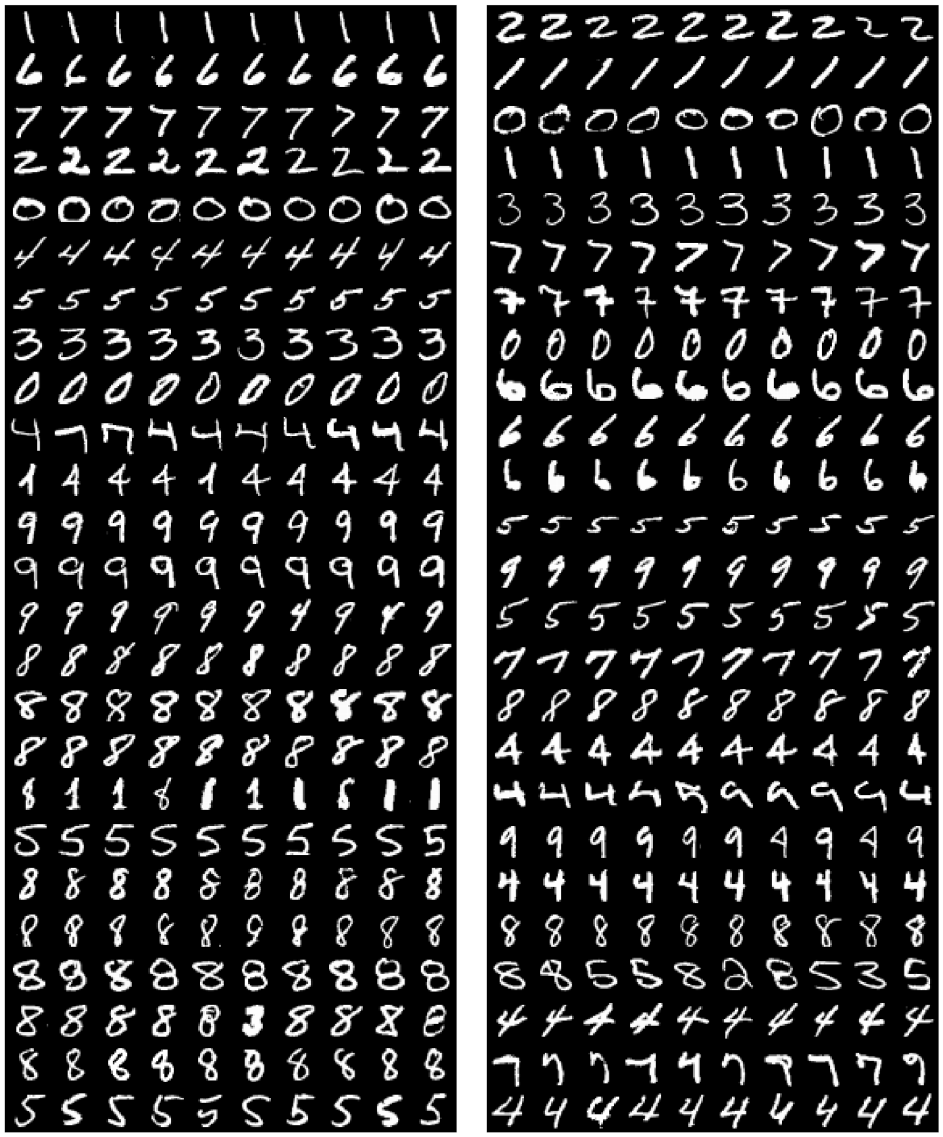
Table 3: Mean and standard deviation of the MCCs (higher is better) across various models, architectures and activations

As recommended in [18], we set the dimension of the latent space to be 5 and number of components to be 25. No hyperparameter tuning was done. The architectures we use are as follows:

- Arch1: The encoder is a sequential neural network architecture with fully connected layers of dimensions $n \to 500 \to 1000 \to m$. The decoder is also a sequential neural network architecture with fully connected layers of dimensions $m \to 500 \to 500 \to n$.
- Arch2: The encoder is a sequential neural network architecture that is fully connected with dimensions $n \to 256 \to 512 \to 512 \to m$. The decoder is similarly a sequential neural network architecture with fully connected layers of dimensions $m \to 512 \to 256 \to n$.
- Arch3: The encoder is a sequential neural network architecture that is fully connected with dimensions $n \to 128 \to 256 \to 128 \to 128 \to m$. The decoder is again a sequential neural network architecture with fully connected layers of dimensions $m \to 128 \to 128 \to n$.

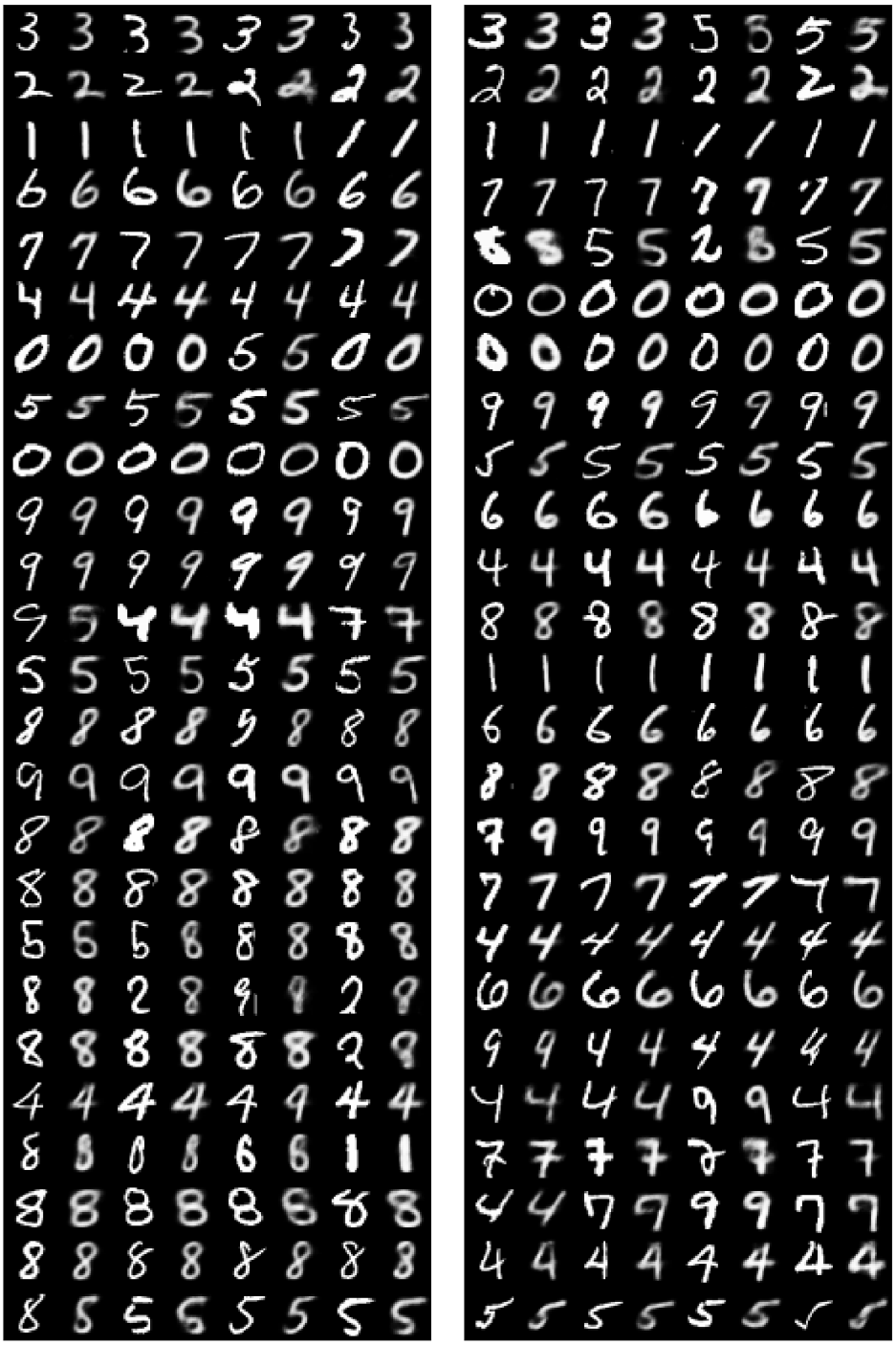
The work [70] ran extensive experiments comparing VaDE and iVAE. We augment these experiments by using MFCVAE instead of VaDE. We observe that even without access to U, MFCVAE has competitive performance (stability) in recovering the latent space as compared to iVAE which has full access to U. This offers strong evidence for stability of training, as predicted by our theory.

For purely illustrative purposes, we also show the output of MFCVAE on MNIST. In Figure 8, we show samples synthetically generated from each learnt cluster. In Figure 9, we visualize the true datapoint x and the corresponding reconstructed $\widehat{x}$ for four different datapoints in each cluster. For similar experiments on other datasets and other architectures, we refer the reader to [18].



(a) Arch1 (b) Arch2

Figure 8: Output of MFCVAE on MNIST data: Synthetically generated samples. Each row corresponds to a different learnt component. The columns are samples generated from the component. The rows are sorted by average confidence.



(a) Arch1 (b) Arch2

Figure 9: Output of MFCVAE on MNIST data: Reconstruction accuracy. Each row corresponds to a different learnt component, the columns correspond to 4 different pairs of x and $\widehat{x}$ in that order.