Embrace the Gap: VAEs Perform Independent Mechanism Analysis

Patrik Reizinger*[1], Luigi Gresele*[2], Jack Brady*[1], Julius von Kügelgen[2,3], Dominik Zietlow[2,4], Bernhard Schölkopf[2], Georg Martius[2], Wieland Brendel[1], and Michel Besserve†[2]

[1]University of Tübingen, Germany
[2]Max Planck Institute for Intelligent Systems, Tübingen, Germany
[3]University of Cambridge, Cambridge, United Kingdom
[4]Amazon Web Services, Tübingen, Germany
{patrik.reizinger,jack.brady,wieland.brendel}@uni-tuebingen.de,
{luigi.gresele,jvk,bs,gmartius,besserve}@tue.mpg.de, zietld@amazon.de

Abstract

Variational autoencoders (VAEs) are a popular framework for modeling complex data distributions; they can be efficiently trained via variational inference by maximizing the evidence lower bound (ELBO), at the expense of a gap to the exact (log-)marginal likelihood. While VAEs are commonly used for disentangled representation learning, it is unclear why ELBO maximization would yield such representations, since unregularized maximum likelihood estimation generally cannot invert the data-generating process without additional assumptions. Yet, VAEs often succeed at this task. We seek to elucidate this apparent paradox by studying nonlinear VAEs in the limit of near-deterministic decoders. We first prove that, in this regime, the optimal encoder approximately inverts the decoder—a commonly used but unproven conjecture—which we refer to as *self-consistency*. Leveraging self-consistency, we show that the ELBO converges to a regularized log-likelihood. This allows VAEs to perform what has recently been termed independent mechanism analysis (IMA): it adds an inductive bias towards decoders with column-orthogonal Jacobians, which helps recovering the true latent factors. The gap between ELBO and log-likelihood is therefore welcome, since it bears unanticipated benefits for nonlinear representation learning. In experiments on synthetic and image data, we show that VAEs uncover the true latent factors when the data generating process satisfies the IMA assumption.

1 Introduction

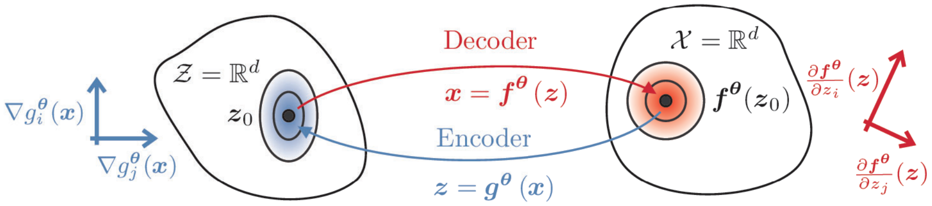
Latent Variable Models (LVMs) allow to effectively approximate a complex data distribution and to sample from it [3, 48]. Deep LVMs employ a neural network (the *decoder* or *generator*) to parameterize the conditional distribution of the observations given latent variables, which are typically assumed to be independent. However, Maximum Likelihood Estimation (MLE) of the model parameters is computationally intractable. In *Variational Autoencoders (VAEs)* [35, 56], the exact log-likelihood is substituted with a tractable lower bound, the evidence lower bound (ELBO). This objective introduces an approximate posterior of the latents given the observations (the *encoder*) from a suitable variational distribution whose mean and covariance are parametrized by neural networks. The encoder is introduced to efficiently train a deep LVM: however, it is not explicitly designed to extract useful representations [17, 58].

[*] Equal contribution. Code available at: github.com/rpatrik96/ima-vae

[†] Senior author

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

https://doi.org/10.52202/068431-0875


Decoder
$\boldsymbol{x} = \boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})$
Encoder
$\boldsymbol{z} = \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$
$\mathcal{Z} = \mathbb{R}^d$
$\boldsymbol{z}_0$
$\mathcal{X} = \mathbb{R}^d$
$\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}_0)$
$\nabla g_i^{\boldsymbol{\theta}}(\boldsymbol{x})$
$\nabla g_j^{\boldsymbol{\theta}}(\boldsymbol{x})$
$\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_i}(\boldsymbol{z})$
$\frac{\partial \boldsymbol{f}^{\boldsymbol{\theta}}}{\partial z_j}(\boldsymbol{z})$

$$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \xrightarrow{\text{Cov}} \text{diag}\left(\sigma_1^{\boldsymbol{\phi}}(\boldsymbol{x})^2, \dots, \sigma_d^{\boldsymbol{\phi}}(\boldsymbol{x})^2\right) = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x}) \frac{1}{\gamma^2} \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})^T \xleftarrow{\text{Cov}} \boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})] \xleftarrow{\boldsymbol{g}^{\boldsymbol{\theta}}} p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$$

Figure 1: **Modeling choices in VAEs promote *Independent Mechanism Analysis* (IMA) [23].** We assume a Gaussian VAE (3), and prove that in the near-deterministic regime the mean encoder approximatetely inverts the mean decoder, $\boldsymbol{g}^{\boldsymbol{\theta}} \approx \boldsymbol{f}^{\boldsymbol{\theta}-1}$ (*self-consistency*, Prop. 1). **Bottom:** Closing the gap requires matching the covariances of the variational (LHS, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$) and the true posterior (RHS, approximated by $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$, cf. § 3.2 for details). Under self-consistency, an encoder with diagonal covariance enforces a row-orthogonal encoder Jacobian $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})$—or equivalently, a column-orthogonal decoder Jacobian $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$. This regularization was termed Independent Mechanism Analysis (IMA) [23] and shown to be beneficial for learning the true latent factors. The connection elucidates unintended benefits of using the ELBO for representation learning.

Nonetheless, VAEs and their variants are widely used in representation learning [25, 1], where they often recover semantically meaningful representations [39, 8, 34, 5]. Our understanding of this empirical success is still incomplete, since (deep) LVMs with independent latents are nonidentifiable from i.i.d. data [29, 42]; different models fitting the data equally well may yield arbitrarily different representations, thus making the recovery of a ground truth generative model impossible. While auxiliary variables, weak supervision [28, 31, 21, 43, 72, 19], or specific model constraints [29, 67, 68, 26, 23] can help identifiability, the mechanism through which the ELBO may enforce a useful inductive bias remains unclear, despite recent efforts [5, 57, 38, 15, 71].

In this work, we investigate the benefits of optimizing the ELBO for representation learning by analyzing VAEs in a *near-deterministic* limit for the conditional distribution parametrized by the nonlinear decoder. Our first result concerns the encoder's optimality in this regime. Previous works relied on the intuitive assumption that the encoder inverts the decoder in the optimum [50, 38, 71]; we formalize this *self-consistency* assumption and prove its validity for the optimal variational posterior in the near-deterministic nonlinear regime.

Using self-consistency, we show that the ELBO tends to a regularized log-likelihood—rather than to the exact one as conjectured in previous work [50]. The regularization term allows VAEs to perform what has been termed Independent Mechanism Analysis (IMA) [23]: it encourages column orthogonality of the decoder's Jacobian. This generalizes previous findings based on linearizations or approximations of the ELBO [57, 44, 38], and allows us to characterize the gap w.r.t. the log-likelihood in the deterministic limit. Our results elucidate the gap between ELBO and exact log-likelihood as a possible mechanism through which the ELBO implements a useful inductive bias. Unlike the unregularized log-likelihood, the IMA-regularized objective can help invert the data generating process under suitable assumptions [23]. We verify this by training VAEs in experiments on synthetic and image data, showing that they can recover the ground truth factors when the IMA assumptions are met.

The **contributions** of this paper can be summarized as follows:

- we characterize and prove *self-consistency* of VAEs in the near-deterministic regime (i.e., when the decoder variance tends to zero), justifying its usage in previous works (§ 3.1);
- we show that under self-consistency, the ELBO converges to a regularized log-likelihood (§ 3.2), and discuss its possible role as a useful inductive bias in representation learning;
- we test the applicability of our theoretical results in experiments on synthetic and image data, and show that VAEs recover the true latent factors when the IMA assumptions are met (§ 4).

2 Background

We will connect two unsupervised learning objectives: the ELBO in VAEs and the IMA-regularized log-likelihood. Both stem from LVMs with latent variables $\boldsymbol{z}$ distributed according to a *prior* $p_0(\boldsymbol{z})$, and a mapping from $\boldsymbol{z}$ to observations $\boldsymbol{x}$ given by a conditional generative model $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$.

Variational Autoencoders. Optimizing the data likelihood $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ in deep LVMs—i.e., finding decoder parameters $\boldsymbol{\theta}$ maximizing $\int p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})d\boldsymbol{z}$—is intractable in general, so approximate objectives are required. Variational approximations [63] replace the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ by an approximate one, called the *variational posterior* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$, which is a stochastic mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ with parameters $\boldsymbol{\phi}$. This allows to evaluate a tractable evidence lower bound (ELBO) [35, 56] of the model's log-likelihood that can be defined as

$$\text{ELBO}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})\right] - \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_0(\boldsymbol{z})\right]. \tag{1}$$

The two terms in (1) are sometimes interpreted as a reconstruction term measuring the sample quality of the decoder and a regularizer—the Kullback-Leibler Divergence (KL) between the prior and the encoder [36]. The variational approximation trades off computational efficiency with a difference w.r.t. the exact log-likelihood, which is expressed alternatively as (see [17, 36] and Appx. A)

$$\text{ELBO}(\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \text{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right], \tag{2}$$

where the KL between variational and true posteriors characterizes the *gap*: if the variational family of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ does not include $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, the ELBO will be strictly smaller than $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$.

VAEs [35] rely on the variational approximation in (1) to train deep LVMs where neural networks parametrize the *encoder* $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and the *decoder* $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$. A common modeling choice constrains the variational family of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to a factorized Gaussian with posterior means $\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$ and variances $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ for the k^{th} factor $z_k|\boldsymbol{x}$, and with a diagonal covariance $\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$; and the decoder to a factorized Gaussian, conditional on $\boldsymbol{z}$, with mean $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z})$ and an isotropic covariance in d dimensions,

$$z_k|\boldsymbol{x} \sim \mathcal{N}(\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x}), \sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2); \qquad \boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}\left(\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{z}), \gamma^{-2}\mathbf{I}_d\right). \tag{3}$$

The deterministic limit of VAEs. The stochasticity of VAEs makes it nontrivial to relate them to generative models with deterministic decoders such as Independent Component Analysis (see paragraph below), though postulating a deterministic regime (where the decoder precision γ^2 becomes infinite) is possible. Interestingly, Nielsen et al. [50] explored this deterministic limit and argued that *deterministic* VAEs optimize an exact log-likelihood, similar to normalizing flows [55, 51]. Normalizing flows model arbitrarily complex distributions using a simple base distribution $p_0(\boldsymbol{z})$ and nonlinear, *deterministic and invertible* transformations $\boldsymbol{f}^{\boldsymbol{\theta}}$. Through a change of variables,[3] the likelihood of the original variables becomes

$$\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \log p_0(\boldsymbol{z}) - \log\left|\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})\right|. \tag{4}$$

The comparison is nontrivial, since VAEs contain an encoder and a decoder, whereas normalizing flows consist of a single architecture. Nielsen et al. [50] made this analogy by resorting to what we call a *self-consistency assumption*, stating that the VAE encoder inverts the decoder. We define self-consistency in the *near-deterministic* regime: as the decoder variance goes to zero, i.e. $\gamma \to +\infty$.

Definition 1 ((Near-deterministic) self-consistency)**.** *For a fixed $\boldsymbol{\theta}$, assume that mean decoder $\boldsymbol{f}^{\boldsymbol{\theta}}$ is invertible with inverse $\boldsymbol{g}^{\boldsymbol{\theta}}$, and that a map associates each choice of decoder parameters and observation $(\boldsymbol{\theta}, \gamma, \boldsymbol{x})$ to an encoder parameter $(\boldsymbol{\theta}, \gamma, \boldsymbol{x}) \mapsto \widehat{\boldsymbol{\phi}}(\boldsymbol{\theta}, \gamma, \boldsymbol{x})$, we say the VAE is self-consistent whenever*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) \to \boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}) \quad \textit{and} \quad \boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 \to \mathbf{0}\text{ , } \textit{as } \gamma \to +\infty\,. \tag{5}$$

The encoder parameter map $\widehat{\boldsymbol{\phi}}$ reflects the choice of a particular encoder model for each $(\boldsymbol{\theta}, \gamma)$ pair:[4] in § 3.1, we study this problem by introducing and justifying a particular choice for $\widehat{\boldsymbol{\phi}}$ (see also § 5). This self-consistency assumption appears central to deterministic claims [50, 38], but has not yet been proven. In particular, Nielsen et al. [50] assume that taking the deterministic limit is well-behaved. However, VAEs' *near-deterministic* properties have not been investigated analytically.

Identifiability, ICA, and IMA. Independent Component Analysis (ICA) [9, 30] models observations as the *mixing* of a latent vector $\boldsymbol{z}$ with independent components through a deterministic function $\boldsymbol{f}$, i.e., $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z}), p_0(\boldsymbol{z}) = \prod_i p_0(z_i)$.[5] In ICA the focus is on defining conditions under which the original

[3] note that in normalizing flows the change of variables is usually expressed in terms of $\boldsymbol{g}^{\boldsymbol{\theta}} = \boldsymbol{f}^{\boldsymbol{\theta}-1}$

[4] both the ELBO and $\widehat{\boldsymbol{\phi}}$ depends on the decoder precision γ: we will omit this in the following for simplicity

[5] the conditional distribution $p(\boldsymbol{x}|\boldsymbol{z})$ is therefore degenerate

latent variables can be recovered from observations—i.e., the model is “identifiable by design” [31]. The goal is to learn an unmixing $\boldsymbol{g^\theta}$ such that the recovered components $\boldsymbol{y} = \boldsymbol{g^\theta}(\boldsymbol{x})$ are estimates of the true ones up to some ambiguities (e.g., permutation and element-wise nonlinear transformations). Unfortunately, the nonlinear problem is nonidentifiable without further constraints [16, 29]: any two observationally equivalent models can yield components which are arbitrarily entangled, thus making recovery of the ground truth factors impossible. This is typically shown by suitably constructed counterexamples [29, 42], and it was argued to imply impossibility statements for unsupervised disentanglement [42, 65]. Identifiability can be recovered when *auxiliary* variables [31, 21, 33, 19] are available, or exploiting a temporal structure in the data [28, 24].

Restrictions on the mixing function class (e.g., linear [9]) are another possibility to recover identifiability [29, 67]. Recently, Gresele et al. [23] proposed restricting the function class by taking inspiration from the *principle of independent causal mechanisms* [52], in an approach termed Independent Mechanism Analysis (IMA). IMA postulates that the latent components influence the observations “independently”, where influences correspond to the partial derivatives $\partial \boldsymbol{f^\theta}/\partial z_k$, and their non-statistical independence amounts to an orthogonality condition. While full identifiability has not been proved for this model class, it was shown to rule out classical families of spurious solutions used as counterexamples to identifiability of unconstrained non-linear ICA [23, 4]. Moroever, Buchholz et al. [4] further demonstrated local identifiability of this function class. Also, IMA constraints were empirically shown [23, 62] to help recover the ground truth through regularization of the log-likelihood in (4) with an objective $\mathcal{L}_{\text{IMA}}(\boldsymbol{f^\theta}, \boldsymbol{z}) := \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - \lambda \cdot c_{\text{IMA}}(\boldsymbol{f^\theta}, \boldsymbol{z})$, where $\lambda > 0$ and the regularization term $c_{\text{IMA}}(\boldsymbol{f^\theta}, \boldsymbol{z})$ and its expectation $C_{\text{IMA}}(\boldsymbol{f^\theta}, p_0)$ are given by

$$c_{\text{IMA}}(\boldsymbol{f^\theta}, \boldsymbol{z}) = \sum_{k=1}^{d} \log \left\| \frac{\partial \boldsymbol{f^\theta}}{\partial z_k}(\boldsymbol{z}) \right\| - \log \left| \mathbf{J}_{\boldsymbol{f^\theta}}(\boldsymbol{z}) \right|; \quad C_{\text{IMA}}(\boldsymbol{f^\theta}, p_0) = \mathbb{E}_{p_0(\boldsymbol{z})}\left[c_{\text{IMA}}(\boldsymbol{f^\theta}, \boldsymbol{z}) \right], \quad (6)$$

and termed *local* (resp. *global*) IMA contrast. When $\boldsymbol{f^\theta}$ is in the IMA function class (i.e., $C_{\text{IMA}}(\boldsymbol{f^\theta}, p_0)$ vanishes), the objective is equal to the log-likelihood; otherwise, it lower bounds it.

3 Theory

Our theoretical analysis assumes that all the model’s defining densities ($p_0(\boldsymbol{z})$, $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$) are factorized. We also assume a Gaussian decoder, matching common modeling practice in VAEs.

Assumption 1 (Factorized VAE class with isotropic Gaussian decoder and log-concave prior)**.** *We are given a fixed latent prior and three parameterized classes of $\mathbb{R}^d \to \mathbb{R}^d$ mappings: the mean decoder class $\boldsymbol{\theta} \mapsto \boldsymbol{f^\theta}$, and the mean and standard deviation encoder classes, $\boldsymbol{\phi} \mapsto \boldsymbol{\mu^\phi}$ and $\boldsymbol{\phi} \mapsto \boldsymbol{\sigma^\phi}$ s.t.*

(i) $p_0(\boldsymbol{z}) \sim \prod_k m(z_k)$, with m being smooth and fully supported on $\mathbb{R}$, having bounded non-positive second-order, and bounded third-order logarithmic derivatives;
(ii) the encoder and decoder are of the form in (3)*, with isotropic decoder covariance $1/\gamma^2 \mathbf{I}_d$;*
(iii) the variational mean and variance encoder classes are universal approximators;
(iv) for all $\boldsymbol{\theta}$, $\boldsymbol{f^\theta} : \mathbb{R}^d \to \mathbb{R}^d$ is a bijection with inverse $\boldsymbol{g^\theta}$, and both are C^2 with bounded first and second order derivatives.

Crucially, *both the mean encoder and the mean decoder can be nonlinear*. Moreover, the family of log-concave priors contains the commonly-used Gaussian distribution as a special case. We study the *near-deterministic decoder* regime of such models, where $\gamma \to +\infty$. This regime is expected to model data generating processes with vanishing observation noise well—in line with the typical ICA setting—and is commonly considered in theoretical analyses of VAEs, e.g., in [50] (which additionally assumes quasi-deterministic encoders), and in [44, 38]. Unlike Nielsen et al. [50], we consider a large but finite γ, not *at* the limit $\gamma = \infty$, where the decoder is fully deterministic. In fact, for any large but finite γ, the objective is well-behaved and amenable to theoretical analysis, while the KL-divergence is undefined in the deterministic setting. The requirement in assumption *(iv)* deviates from common practice in VAEs—where observations are typically higher-dimensional—but it allows to connect VAEs and exact likelihood methods such as normalizing flows [50] (see also § 5).

Due to considering $\gamma \to +\infty$, results are stated in the following “big-O” notation for an integer p:

$$f(\boldsymbol{x}, \gamma) = g(\boldsymbol{x}, \gamma) + O_{\gamma \to +\infty}(1/\gamma^p) \iff \gamma^p \| f(\boldsymbol{x}, \gamma) - g(\boldsymbol{x}, \gamma) \| \text{ is bounded as } \gamma \to +\infty\,.$$

3.1 Self-consistency

In this section, we will prove a *self-consistency* result in the near-deterministic regime. This rests on characterizing optimal variational posteriors (i.e., those minimizing the ELBO gap w.r.t. the likelihood) for a *particular point* $\boldsymbol{x}$ and *fixed decoder parameters* $\boldsymbol{\theta}$. Based on (2), any associated optimal choice of encoder parameters satisfies

$$\widehat{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta}) \in \arg\max_{\boldsymbol{\phi}} \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi}) = \arg\min_{\boldsymbol{\phi}} \mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right] . \tag{7}$$

We call *self-consistent* ELBO the resulting achieved value, denoted as

$$\mathrm{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta}))\,. \tag{8}$$

The expression in (7) corresponds to a problem of *information projection* [10, 48] of $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ onto the set of factorized Gaussian distributions. This means that given a variational family, we search for the optimal $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ to minimize the KL to $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$. While such information projection problems are well studied for closed convex sets where they yield a unique minimizer [11], the set projected onto in our case is not convex (convex combinations of arbitrary Gaussians are not Gaussian), making this problem of independent interest. After establishing upper and lower bounds on the KL divergence (exposed in Prop. 7-8 in Appx. C.2), we obtain the following self-consistency result.

Proposition 1. *[Self-consistency of near-deterministic VAEs] Under Assumption 1, for all $\boldsymbol{x}$, $\boldsymbol{\theta}$, as $\gamma \to +\infty$, there exists at least one global minimum solution of (7). These solutions satisfy*

$$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}) = \boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x}) + O(1/\gamma) \quad \textit{and} \quad \sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = O(1/\gamma^2)\,, \textit{for all } k\,. \tag{9}$$

Prop. 1 states that minimizing the ELBO gap (equivalently, maximizing the ELBO) w.r.t. the encoder parameters $\boldsymbol{\phi}$ implies in the limit of large γ that the encoder's mean $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ tends to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$, the image of $\boldsymbol{x}$ by the *inverse* decoder. We can interpret this as the decoder "inverting" the encoder. Additionally, the variances of the encoder will converge to zero.

Let us now consider the relevance of this result for training VAEs, i.e., maximizing the expectation of the ELBO for an observed distribution $p(\boldsymbol{x})$. While maximization *only* w.r.t. $\boldsymbol{\phi}$ in (7) does not match common practice—which is learning $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ *jointly*—it models this process in the limit of large-capacity encoders. Indeed, in this case, (7) can be solved for each $\boldsymbol{x}$ as a separate learning problem, which entails that the following inequality is satisfied for any parameter choice

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})\right] &= \int p(\boldsymbol{x})\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})d\boldsymbol{x}\\ &\leq \int p(\boldsymbol{x})\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\widehat{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta}))d\boldsymbol{x} =: \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathrm{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta})\right] .\end{aligned} \tag{10}$$

The joint optimization of encoder and decoder parameters thus reduces to optimizing the subset of pairs $(\boldsymbol{\theta}, \widehat{\boldsymbol{\phi}}(\boldsymbol{x},\boldsymbol{\theta}))$, and is equivalent to optimizing the expected self-consistent ELBO, that is

$$\underset{\boldsymbol{\theta},\boldsymbol{\phi}}{\mathrm{maximize}}\, \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathrm{ELBO}(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{\phi})\right] \iff \underset{\boldsymbol{\theta}}{\mathrm{maximize}}\, \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x})}\left[\mathrm{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta})\right] \tag{11}$$

This problem reduction is aligned with the original purpose of the ELBO: building a tractable but optimal likelihood approximation. Namely, (i) ELBO* depends on the same parameters as the likelihood ($\boldsymbol{x}$, γ and $\boldsymbol{\theta}$), (ii) its gap $\mathrm{KL}\left[q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})||p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})\right]$ is minimal. The problem reduction of (11) allows us to compare the optimality of different decoders and Prop. 1 helps addressing the case of near-deterministic decoders.

3.2 Self-consistent ELBO, IMA-regularized log-likelihood and identifiability of VAEs

We want to investigate how the choice of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ implicitly regularizes the Jacobians of their means $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ and $\boldsymbol{f}^{\boldsymbol{\theta}}\,(\boldsymbol{z})$ in the near-deterministic regime. Exploiting self-consistency, we are able to precisely characterize how this happens: we formalize this in Thm. 1.

Theorem 1. *[VAEs with a near-deterministic decoder approximate the* IMA *objective] Under Assumption 1, the variational posterior satisfies*

$$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2 = \left(-\frac{d^2\log p_0}{dz_k^2}(g_k^{\boldsymbol{\theta}}(\boldsymbol{x})) + \gamma^2\left\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\,(\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x}))\right]_{:k}\right\|^2\right)^{-1} + O(1/\gamma^3)\,, \tag{12}$$

and the self-consistent ELBO (10) *approximates the* IMA*-regularized log-likelihood* (6)*:*

$$\mathrm{ELBO}^*(\boldsymbol{x};\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\boldsymbol{x}) - c_{\mathrm{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}},\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})) + O_{\gamma\to\infty}\left(1/\gamma^2\right). \tag{13}$$

https://doi.org/10.52202/068431-0875

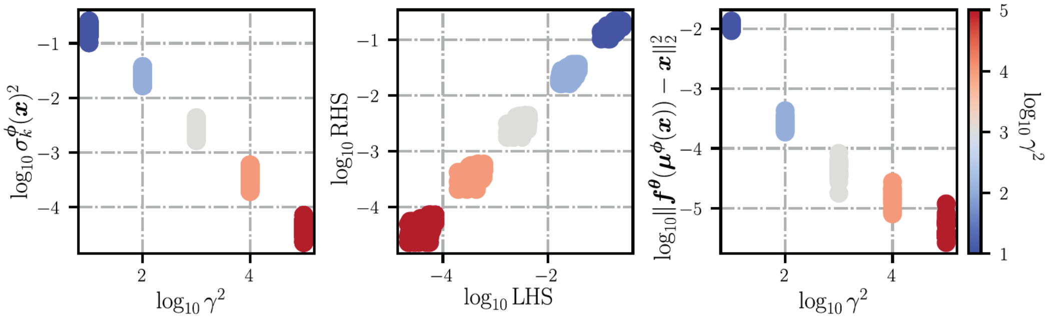


Figure 2: Self-consistency (Prop. 1) in VAE training, on a log-log plot, cf. 4.1 for details. **Left:** convergence of $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ to 0; **Center:** connecting $\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$, γ^2, and the column norms of the decoder Jacobian via LHS and RHS of (12); **Right:** convergence of $\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}\,(\boldsymbol{x})$

Proof is in Appx. B. Below, we provide a qualitative argument on the interplay between distributional assumptions in the VAE and implicit constraints on the decoder's Jacobian and its inverse.

Modeling assumptions implicitly regularize the mean decoder class $\boldsymbol{f}^{\boldsymbol{\theta}}$ under self-consistency. In the near deterministic regime, $p_{\boldsymbol{\theta}}(\boldsymbol{x})$ gets close to the pushforward distribution of the prior by the mean decoder $\boldsymbol{f}^{\boldsymbol{\theta}}{}_*\,[p_0(\boldsymbol{z})]$, which can be used to show that the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})p_0(\boldsymbol{z})/p_{\boldsymbol{\theta}}(\boldsymbol{x})$ is approximately the pushforward through the inverse mean decoder $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*\,[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$ (see Appx. A for more details). If we select a given latent $\boldsymbol{z}_0$ and denote its image by $\boldsymbol{f}^{\boldsymbol{\theta}}\,(\boldsymbol{z}_0)$, then we can locally linearize $\boldsymbol{g}^{\boldsymbol{\theta}}$ by its Jacobian $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}} = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}\,(\boldsymbol{z}_0))$, yielding a Gaussian for the pushforward distribution $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*\,[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$ with covariance ${}^1\!/\gamma^2\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}^T$. As the sufficient statistics of a Gaussian are given by its mean and covariance, the structure of the posterior covariance $\boldsymbol{\Sigma}_{\boldsymbol{z}|\boldsymbol{x}}^{\boldsymbol{\phi}}$ (which is by design diagonal, cf. (3)) is crucial for minimizing the gap in (2). Practically, this implies that in the zero gap limit, the covariances of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ should match, i.e., ${}^1\!/\gamma^2\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}^T$ will be diagonal with entries $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ and therefore $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}$ has orthogonal rows. We can express the decoder Jacobian via the inverse function theorem as $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\,(\boldsymbol{z}_0) = \mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{f}^{\boldsymbol{\theta}}\,(\boldsymbol{z}_0))^{-1}$. As the inverse of a row-orthogonal matrix has orthogonal columns, $\boldsymbol{f}^{\boldsymbol{\theta}}$ satisfies the IMA principle. Additionally, we can relate the variational posterior's variances to the column-norms of $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ as $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2 = {}^1\!/\gamma^2\|\left[\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}\,(\boldsymbol{z}_0)\right]_{:k}\|^{-2}$, as predicted by (12).

Our argument indicates that minimizing the gap between the ELBO and the log-likelihood encourages column-orthogonality in $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}$ by matching the covariances of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and $\boldsymbol{g}^{\boldsymbol{\theta}}{}_*\,[p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})]$. When $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) = p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$, the gap is closed; this is only possible if the decoder is in the IMA class, for which c_{IMA} vanishes and the ELBO *tends to an exact log-likelihood*. To the best of our knowledge, we are the first to prove this for nonlinear functions, extending related work for linear VAEs [44].

Implications for identifiability of VAEs. While previous works argued that the VAE objective favors decoders with a column-orthogonal Jacobian [57, 38], they did not exactly characterize how: our result shows that the self-consistent ELBO tends to a regularized log-likelihood, where the regularization term c_{IMA} explicitly enforces this (soft) constraint. Thus, it possibly explains why VAEs are successful in learning disentangled representations: namely, the IMA function class provably rules out certain spurious solutions for nonlinear ICA [23], and the IMA-regularized log-likelihood was empirically shown to be beneficial in recovering the true latent factors. Thus, we speak about *embracing the gap*, as its functional form equips VAEs with a useful inductive bias. While the IMA function class has not yet been shown to be identifiable in the classical sense such results exist for special cases such as conformal maps ($d = 2$ [29], generalized by the very recent work in [4]), isometries [26] and for closely-related unsupervised nonlinear ICA models [69]. Moreover, Buchholz et al. [4] demonstrate a *local* form of identifiability for the IMA function class. In the following, we empirically corroborate that VAEs: 1) recover the ground truth sources when the mixing satisfies IMA, and thereby 2) achieve unsupervised disentanglement.

https://doi.org/10.52202/068431-0875

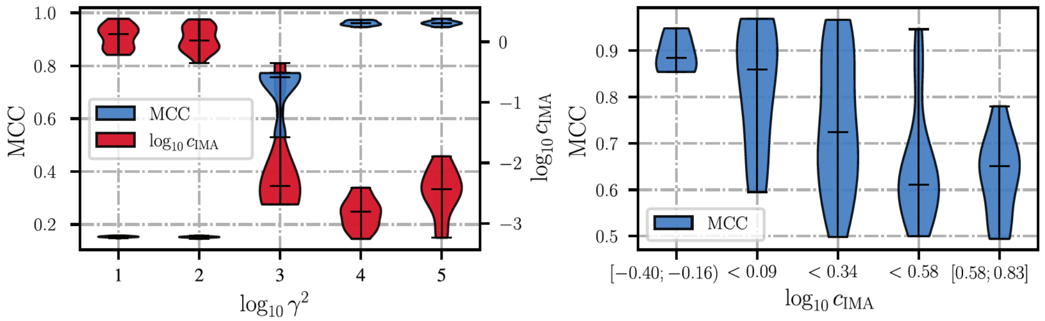


Figure 3: **Left:** c_{IMA} and Mean Correlation Coefficient (MCC) for 3-dimensional Möbius mixings **Right:** MCC depending on the *volume-preserving linear map's* c_{IMA} ($\gamma^2 = 1\text{e}5$)

4 Experiments

Our experiments serve three purposes: 1) demonstrating that self-consistency holds in practice (§ 4.1); 2) showing the relationship of the self-consistent ELBO*, the IMA-regularized and unregularized log-likelihood objectives (§ 4.2); and 3) providing empirical evidence that the connection to the IMA function class in VAEs can lead to success in learning disentangled representations (§ 4.3). More details are provided in Appx. F.

4.1 Self-consistency in practical conditions

Experimental setup. We use a 3-layer Multi-Layer Perceptron (MLP) with smooth Leaky ReLU nonlinearities [22] and orthogonal weight matrices—which intentionally does not belong to the IMA class, as our results are more general. The 60,000 source samples are drawn from a standard normal distribution and fed into a VAE composed of a 3-layer MLP encoder and decoder with a Gaussian prior. We use 20 seeds for each $\gamma^2 \in \{1\text{e}1; 1\text{e}2; 1\text{e}3; 1\text{e}4; 1\text{e}5\}$.

Results. Fig. 2 summarizes our results, featuring the *logarithms* on each axes. The **left** plot shows that the posterior variances $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ converge to zero with a $^1/_{\gamma^2}$ rate, as predicted by (9). The **center** plot shows that the expression for $\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ corresponds to (12) in the optimum of the ELBO by comparing both sides of the equation. The **right** plot shows approximate convergence of the mean encodings $\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}(\boldsymbol{x})$ with a $^1/_{\gamma}$ rate (see § 5). As $\boldsymbol{f}^{\boldsymbol{\theta}}$ is not guaranteed to be invertible, we use instead the *optimal* encoder and decoder parameters to compare $\boldsymbol{f}^{\boldsymbol{\theta}}(\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x}))$ to $\boldsymbol{x}$.

4.2 Relationship between ELBO*, IMA-regularized, and unregularized log-likelihoods

Experimental setup. We use an MLP $\boldsymbol{f}^{\boldsymbol{\theta}}$ with square upper-triangular weight matrices and invertible element-wise nonlinearities to construct a mixing not in the IMA class [23] and fix the VAE decoder to the ground truth such that (4) gives the true data log-likelihood. This way, we ensure that the unregularized and IMA-regularized log-likelihoods differ and make the claim of Nielsen et al. [50] comparable to ours. With a fixed decoder, the ELBO* depends only on $\boldsymbol{\phi}$, therefore we only train the encoder with γ^2 values from $[1\text{e}1; 1\text{e}5]$ (5 seeds each).

Results. Fig. 4 compares the difference of the estimate of ELBO* and the unregularized/IMA-regularized log-likelihoods after convergence over the whole dataset. As the decoder and the data are fixed, $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$ and C_{IMA} will not change during training, only the ELBO* does. The figure shows that as $\gamma \to +\infty$, ELBO* approaches $\mathcal{L}_{\text{IMA}}(\boldsymbol{f}^{\boldsymbol{\theta}}, \boldsymbol{z})$, as predicted by Thm. 1, and not $\log p_{\boldsymbol{\theta}}(\boldsymbol{x})$, as stated in [50]—the difference is C_{IMA}.

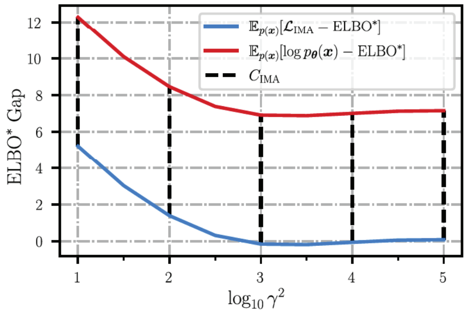


Figure 4: Comparison of the ELBO*, the IMA-regularized and unregularized log-likelihoods over different γ^2. Error bars are omitted as they are orders of magnitudes smaller

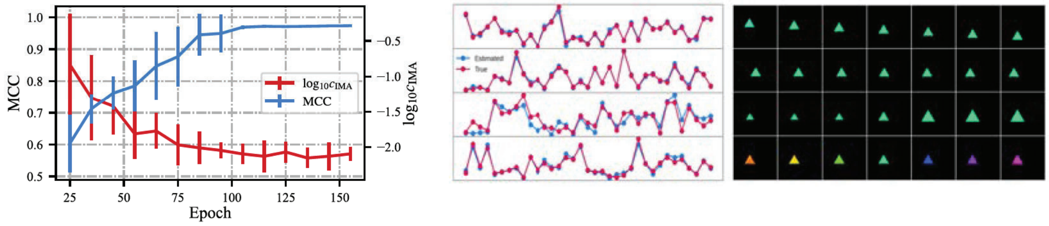


Figure 5: **Left:** c_{IMA} and MCC for Sprites [66] during training ($\gamma^2 = 1$); **Center:** true and estimated latent factors for the best trained VAE on Sprites; **Right:** the corresponding latent interpolations and MCC values (from top to bottom): y- (0.989), x-position (0.996), scale (0.933), and color (0.989)

4.3 Connecting the IMA principle, γ^2, and disentanglement

Experimental setup (synthetic). We use 3-dimensional conformal mixings (i.e., the Möbius transform [53]) from the IMA class with *uniform* ground-truth and prior distributions. Our results quantify the relationship of the decoder Jacobian's IMA-contrast and identifiability with MCC [27] and show how this translates to disentanglement—we note that MCC was already used to quantify disentanglement [72, 37]. To determine whether a mixing from the IMA class is beneficial for disentanglement, we apply a volume-preserving linear map after the Möbius transform (using 100 seeds) to make $c_{\text{IMA}} \neq 0$. Other parameters are the same as in § 4.1, with the exception of picking the best $\gamma^2 = 1\text{e}5$.
Results (synthetic). The **left** of Fig. 3 empirically demonstrates the benefits of optimizing the IMA-regularized log-likelihood. By increasing γ^2, MCC increases, while c_{IMA} decreases, suggesting that VAEs in the near-deterministic regime encourage disentanglement by enforcing the IMA principle. The **right** plot shows that when the mixing is outside the IMA class, MCC decreases, corroborating the benefits of IMA class mixings for disentanglement.
Experimental setup (image). We train a VAE (not β-VAE) with a factorized Gaussian posterior and Beta prior on a Sprites image dataset generated using the spriteworld renderer [66] with a Beta ground truth distribution. Similar to [32], we use four latent factors, namely, *x- and y-position, color and size*, and omit factors that can be problematic, such as shape (as it is discrete) and rotation (due to symmetries) [57, 37]. Our choice is motivated by [26, 18] showing that this data-generating process may approximately satisfy the IMA principle.
Results (image). The **left** of Fig. 5 indicates that VAEs can learn the true latent factors and MCC is *anticorrelated* with c_{IMA}, reinforcing the hypothesis that the data-generating process belongs to the IMA class. The **center** plot compares estimated and true latent factors from the best model (scaling and permutation indeterminacies are removed), whereas the **right** plot shows the corresponding latent interpolations—thus, connecting identifiability (measured by MCC) to disentanglement.

5 Limitations

The near-deterministic regime. Our theory relies on $\gamma \to +\infty$; this is the regime where posterior collapse may be avoided [44], and where calculating the reconstruction loss may be possible even without sampling [38]. However, in practice it may be unclear when γ^2 is large enough. This seems to be problem-dependent [57, 44], and possibly tied to the covariance of the observations [60, 59]. Moreover, large values of γ^2 may be harder to optimize due to an exploding reconstruction term in (1). This may be one explanation for the slight deviation of Fig. 2, right from our theory's predictions: while convergence of $\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ to $\boldsymbol{g}^{\boldsymbol{\theta}}$ matches the prediction in Prop. 1, its rate is not precisely the one predicted for the self-consistent ELBO (10). Another cause could be the encoder's finite capacity. Nonetheless, we have experimentally shown that for realistic hyperparameters, VAEs' behavior matches the predictions of our theory for the near-deterministic regime.

Dimensionality. The setting in § 3 requires equal dimensionality for observations $\boldsymbol{x}$ and latents $\boldsymbol{z}$, in line with work on normalizing flows [51] and nonlinear ICA [28, 31, 24] (but see, e.g., [33]). For high-dimensional images, however, it is often assumed that $\boldsymbol{x}$ lives on a lower-dimensional manifold embedded in a higher-dimensional space, where the dimensionality of $\boldsymbol{x}$ is greater than $\boldsymbol{z}$ [13]. While our theoretical results do not cover this case, we observe empirically in Fig. 5 that the predictions of our theory remain accurate when observations are high-dimensional images. Extending our theory to this setting could leverage ideas explored in, e.g., [13, 12, 7] and is left for future work.

The ELBO, the self-consistent ELBO, and amortized inference. There are in principle multiple ways to obtain self-consistency (Defn. 1). Notably, one could simply force the variational mean and variance encoder maps to behave this way; unlike [38], we model the actual behavior of VAEs trained under ELBO maximization, and obtain self-consistency as a result. For this, we assume that the optimal encoder, which minimizes the gap between ELBO and log-likelihood, can be learned. This is not guaranteed in general, since it requires universal approximation capability of the encoder. On the other hand, (10) requires *unamortized* inference to introduce ELBO*, which does not depend on $\boldsymbol{\phi}$. As in practice amortized inference may be used to efficiently estimate a single set of $\boldsymbol{\phi}$ for all $\boldsymbol{x}$ [61], it can lead to a suboptimal gap to the log-likelihood and discrepancies with our theoretical predictions.

6 Discussion

On disentanglement in unsupervised VAEs. It is widely believed that unsupervised VAEs cannot learn disentangled representations [42, 33], motivating work on models with, e.g., conditional priors [33] or sparse decoding [47]. We show that under certain assumptions, ELBO optimization can implement useful inductive biases for representation learning, yielding disentangled representations in unsupervised VAEs. However, while our results are formulated for VAEs, some of the most successful models at disentanglement are modifications thereof—e.g., β-VAEs [25, 5], with an additional parameter β multiplying the KL in (1). While they deviate from the information projection setting considered in § 3.1, their objectives are equivalent to the ELBO in a sense described in Appx. A.3, which allows us to derive convergence to the IMA-regularized likelihood objective for $\gamma/\sqrt{\beta} \to +\infty$. This encompasses the deterministic limit, and also the setting $\beta \to 0$ with constant γ described in [38]. Whether this theoretical regime matches common practice remains an open question. Overall, we stress that we uncover *one* possible mechanism through which VAEs may achieve disentanglement. By connecting to IMA [23], we discuss implications on recovering the ground truth under suitable assumptions, extending uniqueness results presented in [38]. We speculate that our success in disentanglement is probably due to selecting data sets where the mixing is in the IMA class (cf. [26, 18]), which presumably was not the case in [42].

Characterizing the ELBO gap for nonlinear models. Thm. 1 characterizes the gap between ELBO and true log-likelihood for nonlinear VAEs, and extends the linear analysis of Lucas et al. [44] and the results of Dai et al. [14] in the affine case; we also empirically characterize the gap in the deterministic limit in § 4.2. An unanticipated consequence of this result is that—consistent with [44]—VAEs optimize the IMA-regularized log-likelihood in the near-deterministic limit, and not the unregularized one, as stated in [50].

Extensions to related work. Several papers discuss the (near-)deterministic regime [50, 57, 38, 13]. For example, Nielsen et al. [50] postulate a deterministic VAE with the encoder inverting the decoder. Also Kumar and Poole [38] work in that regime, but without justifying the relationship between the encoder and decoder. Although they show that the choice of $p_0(\boldsymbol{z})$ and $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ influences uniqueness (by, e.g., ruling out rotations), this does not imply recovering the true latents. Our approach formalizes (Defn. 1), proves (Prop. 1), and demonstrates the practical feasibility of (§ 4) the near-deterministic regime. To the best of our knowledge, all previous work relied on the linear case [44] or a (linear) approximation and the evaluation of the ELBO *around a point* to show the inductive bias on the decoder Jacobian. However, our main result (Thm. 1) yields a nonlinear equation where the decoder Jacobian can be evaluated at *any point* and is equipped with a convergence bound. Moreover, the consistency of VAE estimation for identifiable models [33] requires guarantees on $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$; our result helps proving these. Dai and Wipf [13] use a non-factorized Gaussian variational posterior and prove in their Thm. 2 (including the $\dim \boldsymbol{x} = \dim \boldsymbol{z}$ case) that in the deterministic limit their κ-simple VAE can fit perfectly arbitrary observed data (barring few assumptions), while the ELBO gap tends to zero. In contrast, we use a factorized variational posterior; this prevents the ELBO gap to vanish in the deterministic limit, except in the special case of a decoder mean in the IMA class fitting the data perfectly. Dai and Wipf [13] take the limit of $\gamma \to +\infty$ (here using γ as the square root of the decoder precision and not the decoder variance as used in [13]) to relate encoder and decoder properties in this limit in their Thm. 5, similarly to Prop. 1. In contrast to our nonlinear analysis, this is derived when optimizing w.r.t. both encoder and decoder parameters, and with a non-factorized encoder assumption, leading to fundamentally different behavior of the solutions in the deterministic limit. The work done by Sliwa et al. [62], simultaneously to ours, showcases an extensive empirical study highlighting that the IMA contrast allows distinguishing true and spurious solutions for a broad range of cases and outperforms standard regularizers such as weight decay. We discuss extended connections to the literature in Appx. D and Appx. E.

https://doi.org/10.52202/068431-0875

Covariance structure and IMA. We have shown that specific choices for encoder and decoder covariances regularize the decoder Jacobian, such that closing the ELBO gap constrains the decoder to belong to the IMA class. Following our intuition (Fig. 1), assuming factorized $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ and isotropic $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$, IMA holds only for the *decoder*; since in the other direction the pushforward of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ through $\boldsymbol{f}^{\boldsymbol{\theta}}$ has covariance $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})\,\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})^T$, which cannot be used to make row orthogonality statements on $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$ in the general case. Additionally, we conjecture that assuming an isotropic encoder would constrain IMA to hold in both encoding and decoding directions (as both $\mathbf{J}_{\boldsymbol{f}^{\boldsymbol{\theta}}}(\boldsymbol{z})$ and $\mathbf{J}_{\boldsymbol{g}^{\boldsymbol{\theta}}}(\boldsymbol{x})$ need to be column-orthogonal), such that the resulting decoder mean is constrained to have orthogonal columns of equal norms, which is a defining property of conformal maps [4]. On the other hand, we conjecture that if the observation model is not isotropic, but the encoder model is, IMA would only tend to be enforced for the mean *encoder* Jacobian, converging to the inverse decoder mean in the deterministic limit.

Implications for recovering the true latent factors using unsupervised VAEs. Convergence of the ELBO to the IMA-regularized log-likelihood suggests that unsupervised VAEs may recover the true factors of variation according to current identifiability results of the IMA class [4]. This is based on the following reasoning: *If the ground truth generative model belongs to the IMA class, unsupervised learning of the model with an infinite capacity VAE will, in the deterministic limit, ensure a solution that perfectly fits the data and whose decoder mean is also in the IMA class (by joint optimization of both the likelihood and the regularization term). Identifiability of the IMA class implies that the VAE will learn the true decoder (up to acceptable ambiguities); then, since self-consistency guarantees that the encoder inverts the decoder, the encoder infers the ground truth generative factors associated to observations.* Although strict identifiability for all functions in the IMA class remains to be proven, three concurrent papers provide guarantees that go towards identifiability: Leemann et al. [41] proves identifiability for a subset of the IMA class in the context of concept discovery; Zheng et al. [70] shows identifiability of nonlinear ICA by assuming a specific sparsity structure of the decoder Jacobian (called *structural sparsity*); whereas Buchholz et al. [4] introduce the concept of *local identifiability* and proves that IMA is locally identifiable.

Moreover, as mentioned in the above paragraph, we suspect that closing the ELBO gap with an isotropic encoder (while the encoder in Thm. 1 is only constrained to have diagonal covariance) constrains the decoder to be a conformal map. This is an interesting constraint, as nonlinear ICA with conformal mixings are identifiable: the two-dimensional case was first addressed with some additional constraints in [29], while the general case (in arbitrary dimension) was shown to rule out certain spurious solutions for conformal mixings [23], and finally proven to be identifiable by Buchholz et al. [4] in concurrent work. Hence, we conjecture that given a ground truth generative model with a conformal map from latent to observation space, and an unsupervised VAEs with isotropic Gaussian encoders and decoders, the true latent factors can be recovered.

Conclusion. We provide a theoretical justification for VAEs' widely-used self-consistency assumption in the near-deterministic regime of small decoder variance. Using this result, we show that the self-consistent ELBO converges to the IMA-regularized log-likelihood, and not to the unregularized one. Thus, we can characterize the gap between ELBO and true log-likelihood and reason about its role as an inductive bias for representation learning in nonlinear VAEs. We characterize a set of assumptions under which unsupervised VAEs can be expected to disentangle and we demonstrate this behavior in experiments on synthetic and image data.

Acknowledgments and Disclosure of Funding

The authors thank the anonymous reviewers for their suggestions. This work was supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A & 01IS18039B, and by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Wieland Brendel acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Dominik Zietlow and Patrik Reizinger. Patrik Reizinger acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program.

References

[1] Alexander A. Alemi, Ben Poole, Ian Fischer, Joshua V. Dillon, Rif A. Saurous, and Kevin Murphy. Fixing a broken ELBO. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 159–168. PMLR, 2018. 2

[2] Khaled Alyani, Marco Congedo, and Maher Moakher. Diagonality measures of hermitian positive-definite matrices with application to the approximate joint diagonalization problem. *Linear Algebra Appl.*, 528:290–320, September 2017. ISSN 0024-3795. doi: 10.1016/j.laa.2016.08.031. URL `https://doi.org/10.1016/j.laa.2016.08.031`. 23, 24, 42

[3] Christopher M Bishop. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0. URL `https://doi.org/10.1007/978-0-387-45528-0`. 1

[4] Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function Classes for Identifiable Nonlinear Independent Component Analysis. *Advances in Neural Information Processing Systems*, 36, 2022. 4, 6, 10

[5] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018. 2, 9

[6] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3), jun 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL `https://doi.org/10.1145/1970392.1970395`. 41

[7] Anthony L Caterini, Gabriel Loaiza-Ganem, Geoff Pleiss, and John P Cunningham. Rectangular flows for manifold learning. *Advances in Neural Information Processing Systems*, 34, 2021. 8

[8] Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 2615–2625, 2018. 2, 21

[9] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314, 1994. 3, 4

[10] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley series in telecommunications. John Wiley & Sons, Inc., New York, 1991. ISBN 0471062596, 0471200611. doi: 10.1002/0471200611. URL `https://doi.org/10.1002/0471200611`. 5

[11] I. Csiszar and F. Matus. Information projections revisited. *IEEE Trans. Inf. Theory*, 49 (6):1474–1490, June 2003. ISSN 0018-9448. doi: 10.1109/tit.2003.810633. URL `https://doi.org/10.1109/tit.2003.810633`. 5

[12] Edmond Cunningham and Madalina Fiterau. A change of variables method for rectangular matrix-vector products. In *International Conference on Artificial Intelligence and Statistics*, pages 2755–2763. PMLR, 2021. 8

[13] Bin Dai and David Wipf. Diagnosing and enhancing VAE models. In *International Conference on Learning Representations*, 2018. 8, 9, 21, 41

[14] Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with robust PCA and the role of emergent sparsity in variational autoencoder models. *The Journal of Machine Learning Research*, 19(1):1573–1614, 2018. 9

[15] Bin Dai, Ziyu Wang, and David Wipf. The usual suspects? Reassessing blame for VAE posterior collapse. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 2313–2322. PMLR, 13–18 Jul 2020. URL `https://proceedings.mlr.press/v119/dai20c.html`. 2

https://doi.org/10.52202/068431-0875

[16] George Darmois. Analyse des liaisons de probabilité. In *Proc. Int. Stat. Conferences 1947*, page 231, 1951. 4

[17] Carl Doersch. Tutorial on Variational Autoencoders. *ArXiv preprint*, abs/1606.05908, 2016. 1, 3, 22

[18] David L. Donoho and Carrie Grimes. Image Manifolds which are Isometric to Euclidean Space. *J. Math. Imaging Vis.*, 23(1):5–24, July 2005. ISSN 0924-9907, 1573-7683. doi: 10.1007/s10851-005-4965-4. URL https://doi.org/10.1007/s10851-005-4965-4. 8, 9, 46

[19] Élisabeth Gassiat, Sylvain Le Corff, and Luc Lehéricy. Deconvolution with unknown noise distribution is possible for multivariate signals. *Ann. Statist.*, 50(1), February 2022. ISSN 0090-5364. doi: 10.1214/21-aos2106. URL https://doi.org/10.1214/21-aos2106. 2, 4

[20] Partha Ghosh, Mehdi S. M. Sajjadi, Antonio Vergari, Michael J. Black, and Bernhard Schölkopf. From variational to deterministic autoencoders. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 21, 41

[21] Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The Incomplete Rosetta Stone problem: Identifiability results for Multi-view Nonlinear ICA. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 115 of *Proceedings of Machine Learning Research*, pages 217–227. PMLR, July 2019. URL https://proceedings.mlr.press/v115/gresele20a.html. 2, 4

[22] Luigi Gresele, Giancarlo Fissore, Adrián Javaloy, Bernhard Schölkopf, and Aapo Hyvärinen. Relative gradient optimization of the Jacobian term in unsupervised deep learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 7, 43, 44, 45

[23] Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanisms analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pages 28233–28248. Curran Associates, Inc., December 2021. URL https://proceedings.neurips.cc/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf. 2, 4, 6, 7, 9, 10, 23, 43

[24] Hermanni Hälvä and Aapo Hyvärinen. Hidden Markov nonlinear ICA: Unsupervised learning from nonstationary time series. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of the Thirty-Sixth Conference on Uncertainty in Artificial Intelligence, UAI 2020, virtual online, August 3-6, 2020*, volume 124 of *Proceedings of Machine Learning Research*, pages 939–948. AUAI Press, 2020. 4, 8

[25] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 2, 9, 21

[26] Daniella Horan, Eitan Richardson, and Yair Weiss. When is unsupervised disentanglement possible? *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 8, 9, 46

[27] Aapo Hyvärinen and Hiroshi Morioka. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3765–3773, 2016. 8

[28] Aapo Hyvärinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort*

https://doi.org/10.52202/068431-0875

Lauderdale, FL, USA, volume 54 of *Proceedings of Machine Learning Research*, pages 460–469. PMLR, 2017. 2, 4, 8

[29] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/s0893-6080(98)00140-3. URL `https://doi.org/10.1016/s0893-6080(98)00140-3`. 2, 4, 6, 10, 24, 41

[30] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, Inc., New York, May 2001. ISBN 047140540X, 0471221317. doi: 10.1002/0471221317. URL `https://doi.org/10.1002/0471221317`. 3

[31] Aapo Hyvärinen, Hiroaki Sasaki, and Richard E. Turner. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019. 2, 4, 8

[32] Jack Brady and Geoffrey Roeder. iSprites: A Dataset for Identifiable Multi-Object representation Learning. In *ICML2020: Workshop on Object-Oriented Learning*, 2020. URL `https://github.com/oolworkshop/oolworkshop.github.io/blob/master/pdf/OOL_25.pdf`. 8, 46

[33] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020. 4, 8, 9

[34] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2654–2663. PMLR, 2018. 2, 21

[35] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 1, 3, 21

[36] Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000056. URL `https://doi.org/10.1561/2200000056`. arXiv: 1906.02691. 3

[37] David A. Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 8, 46

[38] Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs. In *International Conference on Machine Learning*, pages 5480–5490. PMLR, 2020. 2, 3, 4, 6, 8, 9, 21, 23, 41, 46

[39] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 2

[40] Abhishek Kumar, Ben Poole, and Kevin Murphy. Regularized Autoencoders via Relaxed Injective Probability Flow. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 4292–4301. PMLR, 2020. 41

[41] Tobias Leemann, Michael Kirchhof, Yao Rong, Enkelejda Kasneci, and Gjergji Kasneci. Disentangling Embedding Spaces with Minimal Distributional Assumptions, June 2022. URL `http://arxiv.org/abs/2206.13872`. arXiv:2206.13872 [cs, stat]. 10

[42] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019. 2, 4, 9

[43] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6348–6359. PMLR, 2020. 2

[44] James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't blame the ELBO! a linear VAE perspective on posterior collapse. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9403–9413, 2019. 2, 4, 6, 8, 9, 19, 41, 42, 46

[45] Jerrold E. Marsden and Anthony Tromba. *Vector calculus*. W.H. Freeman and Company, New York, sixth edition edition, 2012. ISBN 978-1-4292-1508-4. 39

[46] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 4402–4412. PMLR, 2019. 41

[47] Gemma E Moran, Dhanya Sridhar, Yixin Wang, and David M Blei. Identifiable variational autoencoders via sparse decoding. *arXiv preprint arXiv:2110.10804*, 2021. 9

[48] Kevin P Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012. 1, 5

[49] Akira Nakagawa, Keizo Kato, and Taiji Suzuki. Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding. In *Proceedings of the 38th International Conference on Machine Learning*, pages 7916–7926. PMLR, July 2021. ISSN: 2640-3498. 40

[50] Didrik Nielsen, Priyank Jaini, Emiel Hoogeboom, Ole Winther, and Max Welling. SurVAE flows: Surjections to bridge the gap between VAEs and flows. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2, 3, 4, 7, 9, 41

[51] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *ArXiv preprint*, abs/1912.02762, 2019. 3, 8

[52] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference – Foundations and Learning Algorithms*. MIT Press, Cambridge, MA, USA, 2017. 4

[53] Robert Phillips. Liouville's theorem. *Pac. J. Math.*, 28(2):397–405, February 1969. ISSN 0030-8730, 0030-8730. doi: 10.2140/pjm.1969.28.397. URL `https://doi.org/10.2140/pjm.1969.28.397`. 8, 43, 45

[54] Lutz Prechelt. Early stopping - but when? In *Neural Networks: Tricks of the Trade, volume 1524 of LNCS, chapter 2*, pages 55–69. Springer-Verlag, 1997. 43, 45, 46

https://doi.org/10.52202/068431-0875

[55] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis R. Bach and David M. Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1530–1538. JMLR.org, 2015. 3

[56] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014. 1, 3

[57] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12406–12415. IEEE, June 2019. doi: 10.1109/cvpr.2019.01269. URL https://doi.org/10.1109/cvpr.2019.01269. 2, 6, 8, 9, 21, 40, 41, 46

[58] Paul K. Rubenstein. Variational Autoencoders are not autoencoders, Jan 2019. URL http://paulrubenstein.co.uk/variational-autoencoders-are-not-autoencoders/. 1

[59] Oleh Rybkin, Kostas Daniilidis, and Sergey Levine. Simple and effective VAE training with calibrated decoders. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 9179–9189. PMLR, 2021. 8, 22, 46

[60] Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. 2021. 8, 46

[61] Rui Shu, Hung H Bui, Shengjia Zhao, Mykel J Kochenderfer, and Stefano Ermon. Amortized inference regularization. *Advances in Neural Information Processing Systems*, 31, 2018. 9

[62] Joanna Sliwa, Shubhangi Ghosh, Vincent Stimper, Luigi Gresele, and Bernhard Schölkopf. Probing the Robustness of Independent Mechanism Analysis for Representation Learning, July 2022. URL http://arxiv.org/abs/2207.06137. arXiv:2207.06137 [cs, stat]. 4, 9

[63] Michael Struwe. *Variational Methods*, volume 991. Springer Berlin Heidelberg, 2000. ISBN 9783662041963, 9783662041949. doi: 10.1007/978-3-662-04194-9. URL https://doi.org/10.1007/978-3-662-04194-9. 3

[64] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. B*, 61(3):611–622, August 1999. ISSN 1369-7412, 1467-9868. doi: 10.1111/1467-9868.00196. URL https://doi.org/10.1111/1467-9868.00196. 41

[65] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 4

[66] Nicholas Watters, Loic Matthey, Sebastian Borgeaud, Rishabh Kabra, and Alexander Lerchner. Spriteworld: A flexible, configurable reinforcement learning environment. https://github.com/deepmind/spriteworld/, 2019. URL https://github.com/deepmind/spriteworld/. 8, 46

[67] Kun Zhang and Laiwan Chan. Minimal nonlinear distortion principle for nonlinear Independent Component Analysis. *Journal of Machine Learning Research*, 9(Nov):2455–2487, 2008. 2, 4

[68] Kun Zhang and Aapo Hyvarinen. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv:1205.2599 [cs, stat]*, 2012. arXiv: 1205.2599. 2

[69] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA with unconditional priors. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022. 6

[70] Yujia Zheng, Ignavier Ng, and Kun Zhang. On the Identifiability of Nonlinear ICA: Sparsity and Beyond, June 2022. URL `http://arxiv.org/abs/2206.07751`. arXiv:2206.07751 [cs, stat]. 10

[71] Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying Inductive Biases for (Beta-) VAE Based Architectures. In *International Conference on Machine Learning*, pages 12945–12954. PMLR, 2021. 2, 40

[72] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR, 2021. 2, 8

Acronyms

ELBO evidence lower bound
IMA Independent Mechanism Analysis

i.i.d. independent and identically distributed
ICA Independent Component Analysis

KL Kullback-Leibler Divergence

LVM Latent Variable Model

MCC Mean Correlation Coefficient

MLE Maximum Likelihood Estimation
MLP Multi-Layer Perceptron
MSE Mean Squared Error

PCA Principal Component Analysis
PPCA Probabilistic Principal Component Analysis

SVD Singular Value Decomposition

VAE Variational Autoencoder

Nomenclature

Independent Mechanism Analysis

C_{IMA} global IMA contrast
α scalar field
$\mathbf{D}$ general diagonal matrix
$\mathbf{O}$ orthogonal matrix
$\boldsymbol{y}$ reconstructed sources
$\mathcal{L}_{\text{IMA}}$ IMA loss function
c_{IMA} local IMA contrast

Variational Autoencoder

$\mathbf{V}$ weight matrix of a linear encoder
$\mathbf{W}$ weight matrix of a linear decoder
$\boldsymbol{\mu}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ optimal mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{\mu}^{\boldsymbol{\phi}}(\boldsymbol{x})$ mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{\phi}$ parameters of the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{\sigma}^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ optimal variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$\boldsymbol{\theta}$ parameters of the decoder $p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$
γ square root of the precision of the VAE decoder
$\boldsymbol{\Sigma}^{\boldsymbol{\phi}}_{\boldsymbol{z}|\boldsymbol{x}}$ covariance matrix of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$\mathcal{L}_{\beta}$ β-VAE loss function
$\boldsymbol{f}^{\boldsymbol{\theta}}$ decoder
$\boldsymbol{g}^{\boldsymbol{\theta}}$ inverse decoder
$\widehat{\boldsymbol{\phi}}$ optimal parameters of the variational posterior $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$
$p(\boldsymbol{x})$ data distribution
$p_0(\boldsymbol{z})$ latent prior distribution
$p_{\boldsymbol{\theta}}(\boldsymbol{z}|\boldsymbol{x})$ true posterior distribution of the decoded samples of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$, parametrized by $\boldsymbol{\theta}$
$p_{\boldsymbol{\theta}}(\boldsymbol{x})$ marginal likelihood
$p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z})$ conditional distribution of the decoded samples of the VAE, mapping $\boldsymbol{z} \mapsto \boldsymbol{x}$, parametrized by $\boldsymbol{\theta}$

$q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ variational posterior of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ parametrized by $\boldsymbol{\phi}$

$q_{\widehat{\boldsymbol{\phi}}}(\boldsymbol{z}|\boldsymbol{x})$ optimal variational posterior of the VAE, mapping $\boldsymbol{x} \mapsto \boldsymbol{z}$ parametrized by $\boldsymbol{\phi}$

$\mu_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})$ optimal mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\mu_k^{\boldsymbol{\phi}}(\boldsymbol{x})$ mean of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\sigma_k^{\widehat{\boldsymbol{\phi}}}(\boldsymbol{x})^2$ optimal variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$\sigma_k^{\boldsymbol{\phi}}(\boldsymbol{x})^2$ variance of $q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})$ in dimension k

$g^{\boldsymbol{\theta}}$ inverse decoder component

$\mathbf{H}$ Hessian matrix

$\mathbf{I}_d$ d-dimensional identity matrix

$\mathbf{J}$ Jacobian matrix

$\boldsymbol{\Sigma}$ covariance matrix

$\boldsymbol{x}$ observation vector

$\boldsymbol{z}$ latent vector

$\mathcal{X}$ observation space

d dimensionality of the observation space $\mathcal{X}$

z latent single component

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] See § 3, § 4, § 6
   (b) Did you describe the limitations of your work? [Yes] See § 6
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Appx. H
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We use no data about human subjects.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See § 3
   (b) Did you include complete proofs of all theoretical results? [Yes] See § 3, Appx. B
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] See the supplementary material
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See § 4, Appx. F
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See § 4
   (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 Appx. F
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] See § 4 and the supplementary material
   (b) Did you mention the license of the assets? [Yes] See the supplementary material
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We included the code and the experiment logs.
   (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]

https://doi.org/10.52202/068431-0875