Learning Optimal Conditional Priors For Disentangled Representations

Graziano Mita **Maurizio Filippone** **Pietro Michiardi**
EURECOM, 06410 Biot (France)
{mita, filippone, michiardi}@eurecom.fr

Abstract

A large part of the literature on learning disentangled representations focuses on variational autoencoders (VAEs). Recent developments demonstrate that disentanglement cannot be obtained in a fully unsupervised setting without inductive biases on models and data. As such, Khemakhem et al. (2020) suggest employing a factorized prior distribution over the latent variables that is conditionally dependent on auxiliary observed variables complementing input observations. While this is a remarkable advancement toward model identifiability, the learned conditional prior only focuses on sufficiency, giving no guarantees on a minimal representation. Motivated by information theoretic principles, we propose a novel VAE-based generative model with theoretical guarantees on disentanglement. Our proposed model learns a sufficient and compact – thus optimal – conditional prior, which serves as regularization for the latent space. Experimental results indicate superior performance with respect to state-of-the-art methods, according to several established metrics proposed in the literature on disentanglement.

1 INTRODUCTION

Representation learning aims at learning data representations such that it is easier to extract useful information when building classifiers or other predictive tasks (Bengio et al., 2013). Representation learning seek to obtain the following properties: i) *expressiveness*: a reasonably-sized representation should allow to distinguish among a high number of different input configurations; ii) *abstractness*: learned representations should capture high-level features; iii) *invariance*: representation should be invariant to local changes of input configurations; iv) *interpretability*: learned representations should allow each dimension to be informative about the given task. These properties are at the core of *disentangled representations*.

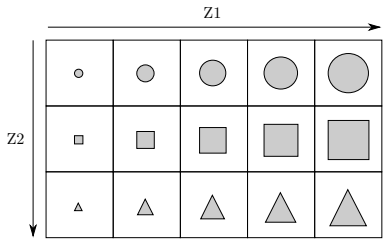

Figure 1: Toy example where each dimension of the latent space $\mathbf{z}$ controls a given generative factor: z_1 the size, z_2 the shape of the 2D objects.

In disentangled representation learning, the main assumption is that high-dimensional observations $\mathbf{x}$ are the result of a (possibly nonlinear) transformation applied to a low dimensional latent variable of independent generative factors, called *ground-truth factors*, capturing semantically meaningful concepts. Input observations can be thought of as the result of a probabilistic generative process, where a latent variable $\mathbf{z}$ is first sampled from a prior distribution $p(\mathbf{z})$, and then the observation $\mathbf{x}$ is sampled from $p(\mathbf{x}|\mathbf{z})$. The goal is to learn a representation of the data that captures the generative factors. In simple terms, illustrated in fig. 1, each dimension of the learned representation refers to a single factor of variation.

In this work, we focus on deep generative models to learn disentangled representations, in particular those based on variational autoencoders (VAE). A well known theoretical result asserts that disentanglement is essentially impossible in a fully unsupervised set-

Preliminary work. Under review by AISTATS 2021. Do not distribute.

ting, without inductive biases on models and data (Locatello et al., 2019). However, inducing a disentangled structure into the latent space $\mathbf{z}$ is feasible by incorporating auxiliary information $\mathbf{u}$ about the ground-truth factors in the model. The type and amount of supervision define different families of disentanglement methods, often classified as supervised, semi-supervised, and weakly-supervised. In most of these methods, the auxiliary variables $\mathbf{u}$ become an integral part of the latent space. However, recent work, such as IVAE (Khemakhem et al., 2020), indicate that there are alternative strategies to benefit from auxiliary information, such as using it to impose a structure on the latent space. This is done in IVAE by learning a prior distribution on the latent space $\mathbf{z}$, where the crucial aspect is that this is conditioned on the auxiliary information $\mathbf{u}$. Under mild assumptions, it is possible to show that such form of conditioning implies model identifiability (Khemakhem et al., 2020), which is equivalent to disentanglement.

While this is a remarkable advancement toward full disentanglement, theoretical guarantees fall short in practice, whereby approximate inference prevents full disentanglement. In this work, we advocate the design of conditional prior distributions that impose optimality constraints on the learned latent representation $\mathbf{z} \sim p(\mathbf{z}|\mathbf{u})$. We propose a novel generative model with theoretical guarantees on disentanglement, that learns a sufficient and compact – thus optimal – conditional prior. This improves the regularization on the function that maps input observations to latent variables, which translate in tangible improvements of disentanglement in practice.

Our main contributions are as follows:

- We present a detailed overview of VAE-based disentanglement methods using a unified notation. Our focus is on the rôle of the regularization term, which we study through the lenses of information theory. We introduce a distinction between *direct matching approaches*, in which ground-truth factors are directly matched to the latent space $\mathbf{z}$, and *indirect matching approaches*, where a prior distribution over the latents is used to structure the learned latent space.
- We improve the model presented in (Khemakhem et al., 2020), and present a new method, *Identifiable Double* VAE (IDVAE), that is identifiabile, in theory, and that learns an optimal conditional prior, which is truly desirable in practice.
- We design an experimental protocol that uses four well-known datasets, and established disentanglement metrics. We compare our method to several state-of-the-art competitors and demonstrate that IDVAE achieves superior disentanglement performance across most experiments.

2 PRELIMINARIES

Let $\mathbf{x} \in \mathbb{R}^n$ be some input observations, which are the result of a transformation of independent latent ground-truth factors $\mathbf{z} \in \mathbb{R}^d$ through a function $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^n$. Then, we have that $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a Gaussian noise term: $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, and independent of $\mathbf{z}$. We consider the following generative model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{z}), \tag{1}$$

where $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ is a vector of model parameters, $p_{\boldsymbol{\theta}}(\mathbf{z}) = \prod_{i=1}^{d} p_{\boldsymbol{\theta}}(z_i)$ represents the factorized prior probability distribution over the latents and $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ is the conditional distribution to recover $\mathbf{x}$ from $\mathbf{z}$. The decoder function $\mathbf{f}(\mathbf{z})$ determines the way $\mathbf{z}$ is transformed into $\mathbf{x}$ within $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

Assume to observe some data $\mathcal{D} = \{\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}\}$ generated by $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}|\mathbf{z})p_{\boldsymbol{\theta}^*}(\mathbf{z})$, where $\boldsymbol{\theta}^*$ are the true, but unknown parameters. Then, the goal is to learn $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z}). \tag{2}$$

When eq. (2) holds, it is possible to recover the ground-truth factors. Unfortunately, by observing $\mathbf{x}$ alone, we can estimate the marginal density $p_{\boldsymbol{\theta}}(\mathbf{x}) \approx p_{\boldsymbol{\theta}^*}(\mathbf{x})$, but there are no guarantees about learning the true generative model $p_{\boldsymbol{\theta}^*}(\mathbf{x}, \mathbf{z})$. This is only feasible for models satisfying the following implication:

$$\forall(\boldsymbol{\theta}, \boldsymbol{\theta}') : p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \implies \boldsymbol{\theta} = \boldsymbol{\theta}'. \tag{3}$$

When eq. (3) holds, the estimated and the true marginal distribution match, and their parameters match too. Then, the model is **identifiable** (Khemakhem et al., 2020) and, as a consequence, it allows one to obtain a fully disentangled representation:

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \implies p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}'}(\mathbf{x}, \mathbf{z}). \tag{4}$$

A practical goal is to aim for model identifiability up to trivial transformations, such as permutation and scaling; as long as ground-truth factors can be identified, their order and scale is irrelevant.

3 TAXONOMY OF VAE-BASED DISENTANGLEMENT

Today, a large body of work to learn disentangled representations is based on generative models, using variants of generative adversarial networks (GAN) (Goodfellow et al., 2014) or variational autoencoders (VAE)

(Kingma and Welling, 2014; Rezende et al., 2014). In this work, we focus on VAE-based approaches.

A standard VAE learns the parameters of eq. (1) by introducing an inference model $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to derive a ELBO as follows:

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (5)$$

where, by abuse of notation, we write $\mathbf{x}$ in place of $\mathbf{x}^{(i)}$. This makes exposition uncluttered for the rest of the section, but, clearly, the marginal log-likelihood is composed of a sum over the marginal log-likelihoods of datapoints $\mathbf{x}^{(i)}$ (Kingma and Welling, 2014).

The distribution $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ has the rôle of a decoder, whereas $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ can be seen as an encoder, and is generally assumed to be a factorized Gaussian with a diagonal covariance matrix. Both distributions are parameterized with neural networks, with parameters $\boldsymbol{\theta}$ and variational parameters $\boldsymbol{\phi}$. The prior $p(\mathbf{z})$ is generally a factorized, isotropic unit Gaussian.

The first term of eq. (5) relates to the *reconstruction* of the input data using latent variables sampled from the variational approximation of the true posterior. The second term is a *regularization* term, which pushes the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ to match the prior on the latent space. Maximizing eq. (5) across datapoints implies learning the parameters such that the reconstruction performance is high, and the regularization term is small. Most approaches to learning disentangled representations from the literature operate on the regularization term. As such, we next focus on its rôle in determining the properties of the latent space.

3.1 Disentanglement and Regularization

The regularization term in eq. (5) is linked to disentanglement: it gauges the capacity of the inference model for learning the latent space and controls the emphasis on learning independent latent factors. Indeed, representation learning can be studied through the lenses of information theory (Achille and Soatto, 2016), where an optimal representation $\mathbf{z}$ of observed variables $\mathbf{x}$, for a generic task $\mathbf{y}$ (in our case, we aim at reconstructing $\mathbf{x}$) is defined in terms of sufficiency and minimality: $\mathbf{z}$ is *sufficient* for the task $\mathbf{y}$ if $I(\mathbf{x};\mathbf{y}) = I(\mathbf{z};\mathbf{y})$, where $I(\cdot;\cdot)$ is the mutual information; $\mathbf{z}$ is *minimal* if it compresses the input such that it discards all variability in the input that is not relevant for the task.

Since $I(\mathbf{x};\mathbf{y}) - I(\mathbf{z};\mathbf{y}) = H(\mathbf{y}|\mathbf{z}) - H(\mathbf{y}|\mathbf{x})$, finding an optimal representation is equivalent to computing the posterior $p(\mathbf{z}|\mathbf{x})$ that solves the optimization problem:

$$\begin{aligned} \min \quad & I(\mathbf{x};\mathbf{z}) \\ \text{s. t.} \quad & H(\mathbf{y}|\mathbf{z}) = H(\mathbf{y}|\mathbf{x}), \end{aligned}$$

where we denote the entropy by $H(\cdot)$. To address the difficulty of such an optimization problem, Tishby et al. (1999) introduce a generalization known as the *Information Bottleneck* (IB), which amounts to optimizing the following Lagrangian:

$$\mathcal{L}_{\text{IB}} = H(\mathbf{y}|\mathbf{x}) + \beta I(\mathbf{x};\mathbf{z}), \quad (6)$$

with the constant β controlling the trade-off between sufficiency and minimality. It is easy to show (Achille and Soatto, 2016) that eq. (5) and eq. (6) address the same optimization problem when $\beta = 1$.

Since both terms that appear in the regularization of eq. (5) are factorized Gaussians with diagonal covariance, one way to interpret the individual components z_i of the latent space is to view them as independent white noise Gaussian channels (Burgess et al., 2017). When the $\text{KL}(\cdot)$ term is zero, the latent channels z_i have zero capacity: this happens when the approximate posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ matches exactly the prior $p_{\boldsymbol{\theta}}(\mathbf{z})$. In this case, however, the reconstruction term is penalized. To increase $\text{KL}(\cdot) > 0$, it is necessary to decrease the overlap between channels, and reduce their variances.

The above understanding of the regularization term is at the basis of many variants of the original VAE model, that strive to increase the pressure on the regularization term, or elements thereof, to achieve disentanglement, without sacrificing reconstruction properties too much. For example Higgins et al. (2017) propose β-VAE, which modifies eq. (5) by introducing a hyper-parameter β to gauge the pressure on the regularization term throughout the learning process:

$$\mathcal{L}_{\beta-\text{VAE}} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \beta\text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \quad (7)$$

When $\beta > 1$, the encoder distribution $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ is pushed towards the unit Gaussian prior $p(\mathbf{z})$. This is equivalent to the IB formulation from eq. (6). In light of the discussion above, the strong penalization of the KL term in β-VAE affects the latent channel distribution, by reducing the spread of their means, and increasing their variances.

Many methods build on β-VAE (Burgess et al., 2017; Kim and Mnih, 2018; Kumar et al., 2018; Chen et al., 2018; Zhao et al., 2019), rewriting the ELBO in slightly different ways. A generalization of the KL term decomposition proposed by Hoffman and Johnson (2016); Makhzani and Frey (2017) is the following (Chen et al., 2018):

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}}[\text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))] = \\ &I(\mathbf{x};\mathbf{z}) + \text{KL}(q(\mathbf{z})||\prod_j q(z_j))) + \sum_j \text{KL}(q(z_j)||p(z_j)) \end{aligned}$$

where $q(\mathbf{z})$ is the aggregated posterior. Penalizing $I(\mathbf{x};\mathbf{z})$ can be harmful to reconstruction purposes, but enforcing a factorized aggregated posterior encourages independence across the dimensions of $\mathbf{z}$, favouring disentanglement. The dimensional independence in the latent space is forced by the second term, known as total correlation (TC). The third term is a further regularization, preventing the aggregate posterior to deviate too much from the factorized prior.

Unfortunately, unsupervised VAE-based approaches allow to approximate the data marginal distribution $p_{\boldsymbol{\theta}}(\mathbf{x})$, but there are no guarantees to recover the true joint probability distribution $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$, having access to the input observations $\mathbf{x}$ only (Khemakhem et al., 2020). Pushing the model to learn a representation with statistically independent dimensions is not a sufficient condition to obtain full disentanglement. These considerations were recently formalized in the *impossibility result* (Locatello et al., 2019), but they were already known in the nonlinear ICA literature (Comon, 1994; Hyvärinen and Pajunen, 1999).

To overcome such limitations, the idea is to incorporate an inductive bias in the model, taking the form of additional information about the ground-truth factors, which we indicate as $\mathbf{u} \in \mathbb{R}^m$. When auxiliary observed variables $\mathbf{u}$ are available, they can be used jointly with $\mathbf{z}$ to reconstruct the original input $\mathbf{x}$. These methods are usually classified under the semi/weakly supervised family. More specifically Shu et al. (2020) identify three commonly used form of weak supervision: *restricted labeling* (Kingma et al., 2014; Cheung et al., 2015; Siddharth et al., 2017; Klys et al., 2018), *match/group pairing* (Bouchacourt et al., 2018; Hosoya, 2019; Locatello et al., 2020a), and *rank pairing* (Chen and Batmanghelich, 2020a,b). In the extreme case, when all ground-truth factors are known for all the input samples, we label them as supervised disentanglement methods.

Similarly to the unsupervised counterpart, methods relying on auxiliary observed variables $\mathbf{u}$ differ in how the regularization term(s) are designed. Some approaches use a "supervised" regularization term to directly match the latent space $\mathbf{z}$ and the available ground-truth factors $\mathbf{u}$: we refer to this form of regularization as *direct matching*. An example is what we here call FullVAE method (Locatello et al., 2020b), which optimizes the following ELBO:

$$\mathcal{L}_{\text{FullVAE}} = \mathcal{L}_{\beta-\text{VAE}} - \gamma R_s(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}),\mathbf{u}), \tag{8}$$

where $R_s(\cdot)$ is a loss function between the latent and the ground-truth factors (in the original implementation it is a binary cross entropy loss). Other approaches employ a KL divergence term between the posterior and the prior over the latents: we refer to this form of regularization as *indirect matching*. In other words, direct matching methods require explicit knowledge of one or more ground-truth factors, whereas indirect matching can also use weak information about them. Shu et al. (2020) demonstrated that indirect matching methods can enforce some properties in the latent space, leading to what they define as *consistency* and *restrictiveness*. To obtain full disentanglement, a method must satisfy both properties on all the latent dimensions.

A recent work by Khemakhem et al. (2020) establishes a theoretical framework to obtain model identifiability, which is related to disentanglement. They propose a new generative model called IVAE, that learns a disentangled representation using a factorized prior from the exponential family, crucially conditioned on $\mathbf{u}$:

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp\left[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})\right], \tag{9}$$

where Q_i is the base measure, $Z_i(\mathbf{u})$ is the normalizing constant, $\mathbf{T}_i = [T_{i,1}, \cdots, T_{i,k}]$ are the sufficient statistics, and $\boldsymbol{\lambda}_i(\mathbf{u})$ are the corresponding parameters. The dimension of each sufficient statistic k is fixed. In practical applications, the conditional prior is chosen to be a Gaussian location-scale family, where $(\lambda_{i,0}, \lambda_{i,1})$ represent the mean and variance of each latent dimension z_i as a function of $\mathbf{u}$. Then, it is possible to derive the following ELBO for the IVAE model:

$$\begin{aligned}\mathcal{L}_{\text{IVAE}} = &\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \\ &- \beta\text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})||p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})).\end{aligned} \tag{10}$$

In eq. (21), we recognize the usual structure of a reconstruction and a regularization term. A remarkable advancement of the IVAE model relates to identifiability properties: next, we build on IVAE and propose a new approach that inherits identifiability, but that learns an optimal conditional prior.

4 LEARNING OPTIMAL CONDITIONAL PRIORS

The main differentiating factor between semi/weakly supervised methods and the IVAE approach is the way auxiliary variables $\mathbf{u}$ affect the latent space $\mathbf{z}$. In the first family of methods, auxiliary variables are integral part of the latent space, e.g. SS-VAE (Kingma et al., 2014). According to Shu et al. (2020), such auxiliary variables must be chosen carefully, as they impact the disentanglement of the corresponding models, favoring consistency or restrictiveness. Instead, the key idea of IVAE is to use auxiliary variables to impose a structure on the latent space, through the definition of a

conditional prior on them. However, while the IVAE model achieves identifiability, there are no guarantees for the conditional prior to be optimal.

Recall that the goal of disentangled representation learning is to find a representation $\mathbf{z}$ that is aligned with a particular optimal representation, among the many that might exist: one that reflects the generative ground-truth factors. Khemakhem et al. (2020) develop Theorem 1 (which we report in appendix A) showing that in theory, and under some conditions, the IVAE model learns a disentangled representation (up to a class of transformation). In particular, condition (iv) of Theorem 1 focuses on the connection between $\mathbf{u}$ and $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$. It asserts that, in order for a model to be identifiable, there must exist $d\,k+1$ points $\mathbf{u}$ such that the $d\,k \times d\,k$ matrix L, defined as $L = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), \cdots, \boldsymbol{\lambda}(\mathbf{u}_{dk}) - \boldsymbol{\lambda}(\mathbf{u}_0))$, is invertible, that is the columns of the matrix L defined in terms of the conditional prior parameters $\boldsymbol{\lambda}(\mathbf{u}_i)$ in eq. (9) are linearly independent.

Theorem 1 requires auxiliary variables $\mathbf{u}$ to be expressive enough to recover all the independent factors. In information theoretic terms, $\mathbf{u}$ must be sufficient to recover the ground-truth factors. Then, in principle, it is possible to learn an optimal, disentangled representation of $\mathbf{z}$, by sampling from an *appropriate* conditional prior $p(\mathbf{z}|\mathbf{u})$. The conditional prior $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$ in IVAE is one of such appropriate distributions, even if it is not optimal, since it does not guarantee minimality.

In practice, however, all the VAE-based methods we discuss in this work are approximate: as a consequence, theoretical guarantees fall short. Even the IVAE method does not achieve full disentanglement.

Our key idea is to define a new method to learn a different conditional prior $p(\mathbf{z}|\mathbf{u})$, one that is constrained to be optimal, thus both sufficient and minimal. While Theorem 1 still applies, we demonstrate that in practice, whereby approximations spoil theoretical results, learning an optimal conditional prior is truly desirable, since it offers superior regularization quality.

4.1 Identifiable Double VAE – IDVAE

Let $\mathbf{x} \in \mathbb{R}^n$, and $\mathbf{u} \in \mathbb{R}^m$ be two observed random variables, and $\mathbf{z} \in \mathbb{R}^d$ a low-dimensional latent variable, with $d \leq n$. Let also $\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\epsilon}$, with $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$. We approximate the injective function $\mathbf{f}(\cdot)$ with a neural network with parameters $\boldsymbol{\theta}$. Then, consider the following generative models, illustrated in fig. 2:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{u}), \tag{11}$$

$$p_{\boldsymbol{\vartheta}}(\mathbf{z}, \mathbf{u}) = p_{\boldsymbol{\vartheta}}(\mathbf{u}|\mathbf{z})p(\mathbf{z}), \tag{12}$$

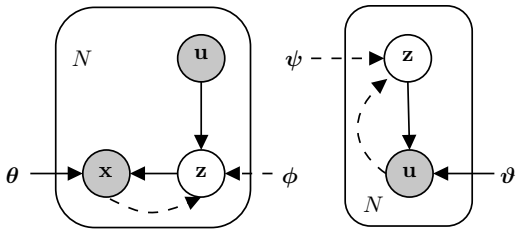

Figure 2: Directed graphical model of IDVAE. Solid lines: generative process. Dashed lines: variational approximation.

where $(\boldsymbol{\theta}, \boldsymbol{\vartheta})$ are the model parameters. Equation (11) corresponds to the process of generating $\mathbf{x}$ given the latents $\mathbf{z}$, where the prior $p(\mathbf{z}|\mathbf{u})$ is conditionally factorial, i.e., each element of $z_i \in \mathbf{z}$ has a univariate Gaussian distribution given the conditioning auxiliary variable $\mathbf{u}$. Similarly, eq. (12) formalizes the process to obtain $\mathbf{u}$ given $\mathbf{z}$, where $p_{\boldsymbol{\vartheta}}(\mathbf{u}|\mathbf{z})$ is a Gaussian distribution with diagonal covariance, and $p(\mathbf{z})$ is a factorized, isotropic unit Gaussian distribution.

Given a dataset $\mathcal{D} = \{(\mathbf{x}^{(1)}, \mathbf{u}^{(1)}), \cdots, (\mathbf{x}^{(N)}, \mathbf{u}^{(N)})\}$ of observations generated according to eq. (11) and eq. (12), we are interested in finding a variational bound for the marginal data log-likelihood $p(\mathbf{x}, \mathbf{u})$, which we derive as follows:

$$\log p(\mathbf{x}, \mathbf{u}) = \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}, \mathbf{u})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}), \tag{13}$$

where, by abuse of notation, we write $\mathbf{x}$ and $\mathbf{u}$ in place of $\mathbf{x}^{(i)}$ and $\mathbf{u}^{(i)}$, which we do hereafter as well.

Since the KL term is non-negative, we have the following variational lower bound:

$$\log p(\mathbf{x}, \mathbf{u}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}). \tag{14}$$

Now, we can write the ELBO, which resembles that of eq. (21), but includes an additional term:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] \\ - \beta \mathrm{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})) + \log p(\mathbf{u}),\end{aligned} \tag{15}$$

where we include the parameter β to gauge the pressure on the the KL term. Note that we use a prior on $\mathbf{z}$ such that it is conditioned $\mathbf{u}$, and this is the crucial aspect of IVAE and IDVAE. Next, we derive a bound for $\log p(\mathbf{u})$ by focusing on the generative model in eq. (12). Our objective is to minimize $\mathrm{KL}(q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})||p(\mathbf{z}|\mathbf{u}))$, which yields:

$$\mathcal{L}_{\text{prior}}(\boldsymbol{\vartheta}, \boldsymbol{\psi}) = \mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})}[\log p_{\boldsymbol{\vartheta}}(\mathbf{u}|\mathbf{z})] - \mathrm{KL}(q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})||p(\mathbf{z})). \tag{16}$$

Note that the variational distribution $q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})$ approximates the conditional prior $p(\mathbf{z}|\mathbf{u})$ in our model.

Combining eq. (15) and eq. (16), we obtain:

$$\begin{aligned}&\mathcal{L}_{\text{IDVAE}}(\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\vartheta}, \boldsymbol{\psi})\\&\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \beta\text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})||p(\mathbf{z}|\mathbf{u}))\\&+ \mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})}[\log p_{\boldsymbol{\vartheta}}(\mathbf{u}|\mathbf{z})] - \text{KL}(q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})||p(\mathbf{z})). \qquad (17)\end{aligned}$$

The ELBO for our model, which we call IDVAE since it includes two variational autoencoders, consists of two contributions with independent parameters. Thus, when we optimize the ELBO by summing across all datapoints, e.g. using a doubly stochastic approach (Titsias and Lázaro-Gredilla, 2014) and automatic differentiation, the two parts can be treated separately. We further make the modeling assumption of using, for the conditional prior $p(\mathbf{z}|\mathbf{u})$, its variational approximation $q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})$: when we optimize $\mathcal{L}_{\text{IVAE}}$, we use the approximate conditional prior, and keep its parameters $\boldsymbol{\psi}$ fixed.

4.2 Properties of IDVAE

IDVAE is a versatile method to learn disentangled representations from observations $\mathbf{x}$ and additionally observed variables $\mathbf{u}$. Unlike existing methods, that accept only a particular type of information – such as content, grouping, ranking, either about single ground-truth factors or all of them – IDVAE works with any variable $\mathbf{u}$. IDVAE uses an *optimal representation of the conditional prior* to structure *the whole* latent space $\mathbf{z}$, whereas the majority of previous work use such information to affect only a portion of the latent variable. Furthermore, unlike IVAE, IDVAE is specifically designed to learn an optimal conditional prior: when $\mathbf{u}$ is informative enough about the ground-truth factors, our conditional prior is a better representation of such ground-truth factors. The experimental evidence we present in section 5 indicates that the effects of optimality are tangible.

IDVAE inherits from IVAE the theoretical guarantees to recover a disentangled representation, which is linked to model identifiability. It is easy to observe that Theorem 1 from (Khemakhem et al., 2020) also applies to IDVAE. Indeed, all conditions in the Theorem, including condition (iv), hold also for IDVAE, since the optimality property of the learned representation for the conditional prior is more restrictive than the sufficiency property that characterizes the conditional prior learned in IVAE.

Note that in section 4.1 we focus on the case where for each input observation $\mathbf{x}^{(i)}$ there exists a corresponding auxiliary variable $\mathbf{u}^{(i)}$. In appendix C we derive the ELBO of the IDVAE method for the semi-supervised case, which is beneficial in practical applications where $\mathbf{u}$ is available for a subset of input observations only.

Dataset	Size	Ground-truth factors (distinct values)
DSprites	737'280	shape(3), scale(6), orientation(40), posX(32), posY(32)
Cars3D	17'568	elevation(4), azimuth(24), object type (183)
Shapes3D	480'000	floor color(10), wall color(10), object color(8), object size(8), object type(4), azimuth(15)
SmallNORB	24'300	elevation(5), azimuth(9), lighting(18)

Table 1: Main characteristics of the datasets.

5 EXPERIMENTS

5.1 Experimental settings

Methods. We compare IDVAE with three disentanglement methods: β-VAE, FullVAE, IVAE. β-VAE (Higgins et al., 2017) is a baseline for indirect matching methods where no ground-truth factor is known at training time and the only way to enforce a disentangled representation is by increasing the strength of the regularization term through the hyper-parameter β. FullVAE (Locatello et al., 2020b) is the representative of direct matching methods: it can be considered as a standard β-VAE with an additional regularization term, weighted by an hyper-parameter γ, to match the latent space to the target ground-truth factors. As done in the original implementation, we use a binary cross entropy loss, where the targets are normalized in $[0, 1]$. We set $\beta = 1$, to measure the impact of the supervised loss term only. IVAE is another indirect matching method where the regularization term, weighted again by β, involves a conditional prior. All methods have been implemented in PyTorch (Paszke et al., 2019).

Datasets. We consider four common datasets in the disentanglement literature, where observations are images built as a deterministic function of known generative factors: DSprites (Higgins et al., 2017), Shapes3D (Kim and Mnih, 2018), Cars3D (Reed et al., 2015) and SmallNORB (LeCun et al., 2004). We have full control on the generative process and explicit access to the ground-truth factors. All ground-truth factors are normalized in the range $[0, 1]$; for discrete factors, we implicitly assume an ordering before applying normalization. All images are reshaped to a 64×64 size. A short description of the four datasets is reported in table 1. Implementations of the generative process for each dataset are based on the code provided by Locatello et al. (2019).

Disentanglement metrics. In the literature, several metrics have been proposed to measure disentanglement, with known advantages and disadvantages, and ability to capture different aspects of disentan-

glement. We report the results for some of the most popular metrics: beta score (Higgins et al., 2017), MIG (Chen et al., 2018), SAP (Kumar et al., 2018), modularity and explicitness (Ridgeway and Mozer, 2018), all with values between 0 and 1. The implementation of the metrics is based on Locatello et al. (2019). For further details, refer to appendix E.

Experimental protocol. In order to fairly evaluate the impact of the regularization terms, all the four selected methods have the same convolutional architecture (widely used in most of the recent works), optimizer, hyper-parameters of the optimizer and batch size. The latent dimension $\mathbf{z}$ is fixed to the true number of ground-truth factors. In practical applications, when this information is not known, it has to be selected during the validation procedure. The conditional prior in IVAE is a MLP network; in IDVAE we use a simple MLP VAE.

We tried six different values of regularization strength associated to the target regularization term of each method – β for β-VAE, IVAE and IDVAE, and γ for FullVAE: $[1, 2, 4, 6, 8, 16]$. These are recurring values in the disentanglement literature. For each model configuration and dataset, we run the training procedure with 10 random seeds. Indeed, all the methods are susceptible to initialization values. After 300'000 training iterations, every model is evaluated according to the disentanglement metrics described above.

For FullVAE, IVAE and IDVAE, all ground-truth factors are provided as input, although IVAE and IDVAE work as well with a subset of them (or with any other additionally observed variable). Although this is an artificially simple protocol, our results show that obtaining full disentanglement is hard, even when all ground-truth factors are known.

5.2 Results

Disentanglement Evaluation. In Figure 3 we report, for each method and for each dataset, the ranges of the beta score and explicitness values with a box-plot. The variance of the box-plots is due to the random seeds and regularization strengths, which are the only parameters we vary. The remaining evaluation metrics can be found in appendix F, but they are essentially all correlated (Locatello et al., 2019).

Overall, we observe, as expected, that β-VAE is often the worst method. Indeed, it has no access to any additional information at training time except the data itself. Despite this, β-VAE disentanglement performance is surprisingly not that far from FullVAE that directly matches the latent space with the ground-truth factors. In some cases, β-VAE obtains very high beta scores (see outliers), such as for DSprites and Cars3D datasets, confirming the sensitivity to random initialization of unsupervised methods (Locatello et al., 2019). Note also that FullVAE exhibits inconsistent performance across the four datasets.

IDVAE emerges as the best method, except for SmallNORB, where FullVAE performs slightly better. For this dataset, we note that there are no considerable differences among methods, since most of the box-plots overlap. It is interesting to observe that IDVAE always outperforms IVAE: learning an optimal conditional prior as we propose in this work, offers substantial benefits in terms of disentanglement. Finally, although both IVAE and IDVAE have theoretical guarantees on disentanglement and use the full set of ground-truth factors as input, they do not always obtain the maximum evaluation score in practice.

Impact of the regularization strength. The disentanglement performance of each method might change drastically as a function of the regularization strength: some approaches might work significantly better in some ranges and very badly in others. In fig. 4, we plot, for each method and for each dataset, the median of the beta score and explicitness evaluation values as a function of the regularization strength. This is also useful to see if there are methods that consistently dominate others. Additional disentanglement score results can be found in appendix F.

Across all the datasets, IDVAE achieves the best median scores for a wide range of regularization strengths. In Shapes3D, IVAE dominates all the other methods (IDVAE is largely dominant also considering the remaining evaluations metrics). The performance of IVAE and FullVAE can match that of IDVAE in some datasets, but the behavior is not consistent: if we focus on beta score, IVAE is the worst method in DSprites– even β-VAE performs better – and its performance on SmallNORB is not sufficient; FullVAE behaves well for DSprites and SmallNORB, but it is on pair only with β-VAE in Cars3D and Shapes3D.

By observing the evolution of the disentanglement scores, it appears that there is no clear strategy to choose the regularization strength. For IDVAE, in datasets such as DSprites and Cars3D, the regularization strength does not significantly affect the beta score; in Shapes3D and SmallNORB, we note instead an increasing and decreasing monotonic trend, respectively. The situation is still different if we look at the explicitness. It is plausible to deduce that the regularization strength is both model and data specific, and it is also affected by the choice of the disentanglement metric.

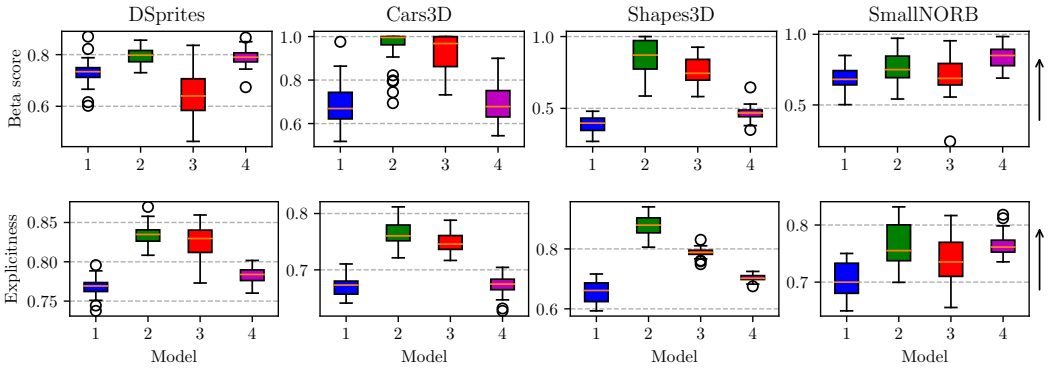

Figure 3: Beta score and explicitness (1=β-VAE, 2=IDVAE, 3=IVAE, 4=FULLVAE).

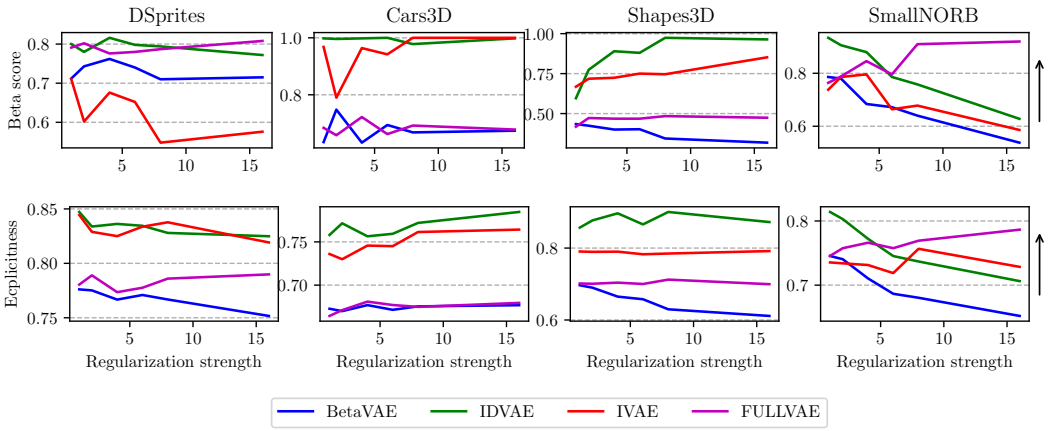

Figure 4: Beta score and explicitness median as a function of the regularization strength.

5.3 Limitations

Our experimental campaign has some clear limitations. First of all, we choose to use the same convolutional architecture for all the tested methods. We do not vary the optimization hyper-parameters and the dimension of the latent variable $\mathbf{z}$. As a consequence, we cannot ensure that every method runs in its best conditions, but this choice was dictated by practical reasons, such as to contain training times.

We use the whole set of ground-truth factors as auxiliary variables, which is the best possible setup. When auxiliary information is not available for all input samples, it is possible to use a semi supervised variant of IDVAE. Both IDVAE and IVAE can use any kind of auxiliary variable: they are not restricted to using, e.g., labels corresponding to input data.

In this work we do not study the effectiveness of the disentanglement methods to solve complex down-stream tasks. However, most of the disentanglement metrics from the literature can be seen as a proxy for the performance of a simple down-stream task, i.e., matching each latent dimension with the corresponding ground-truth factor.

6 CONCLUSION

In this work we made a step further in the design of identifiable generative models, where the goal is to learn disentangled representations of input observations. We proposed a VAE-based model, which uses a prior that encodes ground-truth factor information captured by auxiliary observed variables. The key idea was to learn an optimal representation of the latent space, defined by an inference network on the posterior of the latent variables, given the auxiliary variables. Such posterior is then used as a prior on the latent variables of a generative model, whose inference network learns a mapping between input observations and latents. The results obtained by executing our experimental protocol offer evidence that our approach outperforms existing alternatives to learn disentangled representations, according to several established metrics. In addition, our method exhibit a more stable performance across datasets and hyper-parameters.

References

A. Achille and S. Soatto. Information dropout: learning optimal representations through noise. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013.

D. Bouchacourt, R. Tomioka, and S. Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proc. of the 32nd AAAI Conf. on Artif. Intel.*, AAAI, 2018.

C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner. Understanding disentangling in β-vae. In *Proc. of the 30th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2017.

J. Chen and K. Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proc. of the 34th AAAI Conf. on Artif. Intel.*, AAAI, 2020a.

J. Chen and K. Batmanghelich. Robust ordinal vae: Employing noisy pairwise comparisons for disentanglement. *ArXiv*, 2020b.

T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

B. Cheung, J. A. Livezey, A. K. Bansal, and B. A. Olshausen. Discovering hidden factors of variation in deep networks. In *CoRR*, 2015.

P. Comon. Independent component analysis, a new concept? *Signal Process.*, 36:287–314, 1994.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. of the 27th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2014.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proc. of the 5th Int. Conf. on Learn. Repr.*, ICLR, 2017.

M. D. Hoffman and M. J. Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Adv. in Approx. Bayes. Infer.*, NeurIPS, 2016.

H. Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *Proc. of the 28th Int. Joint Conf. on Artif. Intel.*, IJCAI, 2019.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12:429–439, 1999.

I. Khemakhem, D. P. Kingma., R. P. Mont, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proc. of the 23rd Int. Conf. on Artif. Intel. and Stat.*, AISTATS, 2020.

H. Kim and A. Mnih. Disentangling by factorising. In *Proc. of the 35th Int. Conf. on Mach. Learn.*, ICML, 2018.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *Proc. of the 2nd Int. Conf. on Learn. Repr.*, ICLR, 2014.

D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Proc. of the 27th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2014.

J. Klys, J. Snell, and R. Zemel. Learning latent subspaces in variational autoencoders. In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

A. Kumar, P. Sattigeri, and A. Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *Proc. of the 6th Int. Conf. on Learn. Repr.*, ICLR, 2018.

Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proc. of the 2004 IEEE Comput. Society Conf. on Comput. Vision and Pat. Recogn.*, CVPR, 2004.

F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proc. of the 36th Int. Conf. on Mach. Learn.*, ICML, 2019.

F. Locatello, B. Poole, G. Rätsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *Proc. of the 37th Int. Conf. on Mach. Learn.*, ICML, 2020a.

F. Locatello, M. Tschannen, S. Bauer, G. Rätsch, B. Schölkopf, and O. Bachem. Disentangling factors of variation using few labels. In *Proc. of the 8th Int. Conf. on Learn. Repr.*, ICLR, 2020b.

A. Makhzani and B. J. Frey. Pixelgan autoencoders. In *Proc. of the 30th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2017.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. of the 32nd Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS. Curran Associates, Inc., 2019.

S. E. Reed, Y. Zhang, Y. Zhang, and H. Lee. Deep visual analogy-making. In *Proc. of the 28th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS. Curran Associates, Inc., 2015.

D. J. Rezende, S. Mohamed, and D. Wiestra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of the 31st Int. Conf. on Mach. Learn.*, ICML, 2014.

K. Ridgeway and M. C. Mozer. Learning deep disentangled embeddings with the f-statistic loss. In *Proc. of the 31st Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2018.

R. Shu, Y. Chen, A. Kumar, S. Ermon, and B. Poole. Weakly supervised disentanglement with guarantees. In *Proc. of the 8th Int. Conf. on Learn. Repr.*, ICLR, 2020.

N. Siddharth, B. Paige, J. van de Meent, A. Desmaison, N. Goodman, P. Kohli, F. Wood, and P. Torr. Learning disentangled representations with semi-supervised deep generative models. In *Proc. of the 30th Int. Conf. on Neural Inf. Proc. Sys.*, NeurIPS, 2017.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proc. of the 34th Annual Allert. Conf. on Comm. Contr and Comput.*, 1999.

M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In *Proc. of the 31st Int. Conf. on Mach. Learn.*, ICML, 2014.

S. Zhao, J. Song, and S. Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proc. of the 33rd AAAI Conf. on Artif. Intel.*, AAAI, 2019.

A THE MAIN IDENTIFIABILITY THEOREM

For completeness, we report here the main thereom by (Khemakhem et al., 2020) and a short analysis of its constraints to obtain identifiable models.

Let $\mathbf{x} \in \mathcal{R}^n$, $\mathbf{u} \in \mathcal{R}^m$ be two observed random variables, and $\mathbf{z} \in \mathcal{R}^d$ a latent variable. Let $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ be the parameters of the following conditional generative model:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}|\mathbf{u}) = p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}), \tag{18}$$

$$p_{\mathbf{f}}(\mathbf{x}|\mathbf{z}) = p_{\epsilon}(\mathbf{x} - f(\mathbf{z})). \tag{19}$$

The prior on the latent variables $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$ is conditionally factorial, where each element $z_i \in \mathbf{z}$ has a univariate exponential family distribution given conditioning variable $\mathbf{u}$:

$$p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u}) = \prod_i \frac{Q_i(z_i)}{Z_i(\mathbf{u})} \exp[\sum_{j=1}^{k} T_{i,j}(z_i)\lambda_{i,j}(\mathbf{u})]. \tag{20}$$

IVAE is a practical method to estimate the model defined by eqs. (18) to (20). It optimizes a lower bound on the data log-likelihood defined by:

$$\mathcal{L}_{\text{IVAE}} = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\mathbf{f}}(\mathbf{x}|\mathbf{z})] - \beta \text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})||p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})).$$

Let know define formally the concept of identifiability.

Definition 1. *Let $\sim$ be an equivalence relation on Θ. We say that a model is identifiable up to $\sim$ if $p_{\boldsymbol{\theta}}(\mathbf{x}) = p_{\boldsymbol{\theta}'}(\mathbf{x}) \Longrightarrow \boldsymbol{\theta} \sim \boldsymbol{\theta}'$.*

Definition 2. *Let $\sim$ be the equivalence relation on Θ defined as follows: $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\mathbf{f}', \mathbf{T}', \boldsymbol{\lambda}') \Leftrightarrow \exists \mathbf{A}, \mathbf{c}$: $\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{A}\mathbf{T}'(\mathbf{f}'^{-1}(\mathbf{x})) + \mathbf{c}, \forall \mathbf{x} \in \mathcal{X}$, where $\mathbf{A}$ is a $dk \times dk$ matrix and $\mathbf{c}$ is a dk vector. If A is invertible, we denote this relation by $\sim_A$. If A is a block permutation matrix, we denote it by $\sim_P$.*

Definition 2 is defining a specific equivalence relation that allows to recover the sufficient statistics of our model up to a linear matrix multiplication. The question now is: under which condition it is possible for a model to identifiable according to the equivalence relation in definition 2? Khemakhem et al. (2020) have defined the following theorem:

Theorem 1. *Assume we observe data sampled from eqs. (18) to (20), with parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$. Assume the following holds:*

i The set $\{\mathbf{x} \in \mathcal{X} : \phi_{\epsilon}(\mathbf{x}) = 0\}$ has measure zero, where ϕ_{ϵ} is the characteristic function of the density p_{ϵ} defined in eq. (19).

ii The mixing function f in eq. (19) is injective.

iii The sufficient statistics $T_{i,j}$ in eq. (20) are differentiable almost everywhere, and linearly independent on any subset of $\mathcal{X}$ of measure greater than zero.

iv Being k the dimensionality of the sufficient statistics $\boldsymbol{T}_i$ in eq. (20) and d the dimensionality of $\mathbf{z}$, there exist $dk + 1$ distinct point $\mathbf{u}^0, ..., \mathbf{u}^{dk}$ such that the matrix L defined as follows is invertible:

$$L = (\boldsymbol{\lambda}(\mathbf{u}_1) - \boldsymbol{\lambda}(\mathbf{u}_0), ..., \boldsymbol{\lambda}(\mathbf{u}_{dk}) - \boldsymbol{\lambda}(\mathbf{u}_0))$$

Then the parameters $(\mathbf{f}, \mathbf{T}, \boldsymbol{\lambda})$ are $\sim_A$-identifiable.

Theorem 1 is a theoretical guarantee for identifiability of generative models as defined in eqs. (18) to (20). Here, we will simply try to have a general understanding of its assumptions and implications. Let's focus on the assumptions first: assumption (i) is simply saying that there is a countable (non-continuous) set of samples $\mathbf{x} \in \mathcal{X}$ such that $\mathbf{x} = f(\mathbf{z}) + \epsilon$; assumption (ii) ensures that every observation $\mathbf{x}$ has a unique image in the latent space; assumption (iii) is used to guarantee that the linear transformation A is invertible (this is true for a Gaussian conditional prior). Assumption (iv) focuses on the connection between $\mathbf{u}$ and $p_{\mathbf{T},\boldsymbol{\lambda}}(\mathbf{z}|\mathbf{u})$. It ensures

that L is invertible (this is true as long as $\lambda_{i,j}(\mathbf{u})$ are generated randomly and independently). In other words, the columns of the matrix L defined in terms of the conditional prior parameters $\boldsymbol{\lambda}(\mathbf{u}_i)$ in eq. (20) are linearly independent. This requires for auxiliary variables $\mathbf{u}$ to be expressive enough to recover all the independent sources.

When the four conditions are satisfied, we achieve $\sim_A$-identifiability: we can recover the sufficient statistics of the latents up to a linear matrix multiplication. The remaining theorems can be found in Khemakhem et al. (2020).

B ELBO DERIVATION FOR IDVAE

$$
\begin{aligned}
\log p(\mathbf{x}, \mathbf{u}) &= log \int p(\mathbf{x}, \mathbf{u}, \mathbf{z}) d\mathbf{z} = \\
&= log \int p(\mathbf{x}|\mathbf{u}, \mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u}) d\mathbf{z} = \\
&= log \int \frac{p(\mathbf{x}|\mathbf{u}, \mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{z}|\mathbf{x}, \mathbf{u})} q(\mathbf{z}|\mathbf{x}, \mathbf{u}) d\mathbf{z} \geq \\
&\geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log \frac{p(\mathbf{x}|\mathbf{u}, \mathbf{z}) p(\mathbf{z}|\mathbf{u}) p(\mathbf{u})}{q(\mathbf{z}|\mathbf{x}, \mathbf{u})}] = \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p(\mathbf{x}|\mathbf{u}, \mathbf{z})] - KL(q(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})) + \log p(\mathbf{u}), \\
\log p(\mathbf{u}) &= log \int p(\mathbf{u}, \mathbf{z}) d\mathbf{z} \geq \\
&\geq \mathbb{E}_{q(\mathbf{z}|\mathbf{u})}[\log p(\mathbf{u}|\mathbf{z})] - KL(q(\mathbf{z}|\mathbf{u})||p(\mathbf{z})).
\end{aligned}
$$

C A SEMISUPERVISED VERSION OF IDVAE

Both IVAE and IDVAE assumes to have access to an additional information together with the input observations. This assumption might be unsatisfiable in practice. That's why we designed a variation of IDVAE (it applies also to IVAE) that works in semi-supervised settings. Following the same reasoning of Kingma et al. (2014), we end up with a new objective function that consists of two terms:

$$
\begin{aligned}
\mathcal{L}_{SS-\text{IDVAE}} &= \sum_{(\mathbf{x},\mathbf{u})\sim p_l} \mathcal{L}_{labeled}(\mathbf{x}, \mathbf{u}) + \sum_{\mathbf{x}\sim p_u} \mathcal{L}_{unlabeled}(\mathbf{x}), && (21) \\
\mathcal{L}_{labeled}(\mathbf{x}, \mathbf{u}) &= \mathcal{L}_{\text{IDVAE}}(\mathbf{x}, \mathbf{u}), && (22) \\
\mathcal{L}_{unlabeled}(\mathbf{x}) &= \mathbb{E}_{q_{\boldsymbol{\zeta}}(\mathbf{u}|\mathbf{x})}[\mathcal{L}_{labelled}(\mathbf{x}, \mathbf{u})] - \mathcal{H}(q_{\boldsymbol{\zeta}}(\mathbf{u}|\mathbf{x})), && (23) \\
\mathcal{L}_{\text{IDVAE}}(\mathbf{x}, \mathbf{u}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \text{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u})||p(\mathbf{z}|\mathbf{u})) + \\
&\quad + \mathbb{E}_{q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})}[\log p_{\boldsymbol{\vartheta}}(\mathbf{u}|\mathbf{z})] - \text{KL}(q_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u})||p(\mathbf{z})).
\end{aligned}
$$

As you can notice, there is a new term: $q_{\boldsymbol{\zeta}}(\mathbf{u}|\mathbf{x})$ is used to derive $\mathbf{u}$ from $\mathbf{x}$ when $\mathbf{u}$ is not provided as input. To be precise, we should add to eq. (21) a third term – $\mathbb{E}_{(\mathbf{x},\mathbf{u})\sim p_l}[-\log q_{\boldsymbol{\zeta}}(\mathbf{u}|\mathbf{x})]$ – such that it can learn also from labeled data. We can obtain the semisupervised version of IVAE by simply replacing $\mathcal{L}_{\text{IDVAE}}$ with $\mathcal{L}_{\text{IVAE}}$ in the equations above.

D MODEL ARCHITECTURES, PARAMETERS AND HYPERPARAMETERS

Everything is summarized in tables 2 to 4.

E IMPLEMENTATION OF DISENTANGLEMENT METRICS

Beta score The idea behind the beta score (Higgins et al., 2017) is to fix a random ground-truth factor and sample two mini batches of observations from the corresponding generative model. The encoder is then used to

obtain a learned representation from the observations (with a ground-truth factor in common). The dimension-wise absolute difference between the two representation is computed and a simple linear classifier C is used to predict the corresponding ground-truth factor. This is repeated $batch\_size$ times and the accuracy of the predictor is the disentanglement metric score.

MIG - Mutual Information Gap The mutual information gap (MIG) (Chen et al., 2018) is computed as the average, normalized difference between the highest and second highest mutual information of each ground-truth factor with the dimensions of the learned representation. As done in Locatello et al. (2019), we consider the mean representation. and compute the discrete mutual information by binning each dimension of the mean learned representation into n_bins bins.

Separated Attribute Predictability - (SAP) According to Kumar et al. (2018), the Separated Attribute Predictability (SAP) score is computed from a score matrix where each entry is the linear regression or classification score (in case of discrete factors) of predicting a given ground-truth factors with a given dimension of the learned representation. The (SAP) score is the average difference of the prediction error of the two most predictive learned dimensions for each factor. As done in (Locatello et al., 2019), we use a linear SVM as classifier.

Modularity and Explicitness A representation is modular if each dimension depends on at most one ground-truth factor. Ridgeway and Mozer (2018) propose to measure the Modularity as the average normalized squared difference of the mutual information of the factor of variations with the highest and second-highest mutual information with a dimension of the learned representation. A representation is explicit if it is easy to predict a factor of variation. To compute the explicitness, they train a one-versus-rest logistic regression classifier to predict the ground-truth factor of variation and measeure its ROC-AUC. In the current implementation, observations are discretized into n_bins bins.

As explained in the main paper, the implementation of the selected disentanglement evaluation metrics is based on Locatello et al. (2019). We report the main parameters in table 5.

F FULL EXPERIMENTS

All experiments are summarized in figs. 5 and 6.

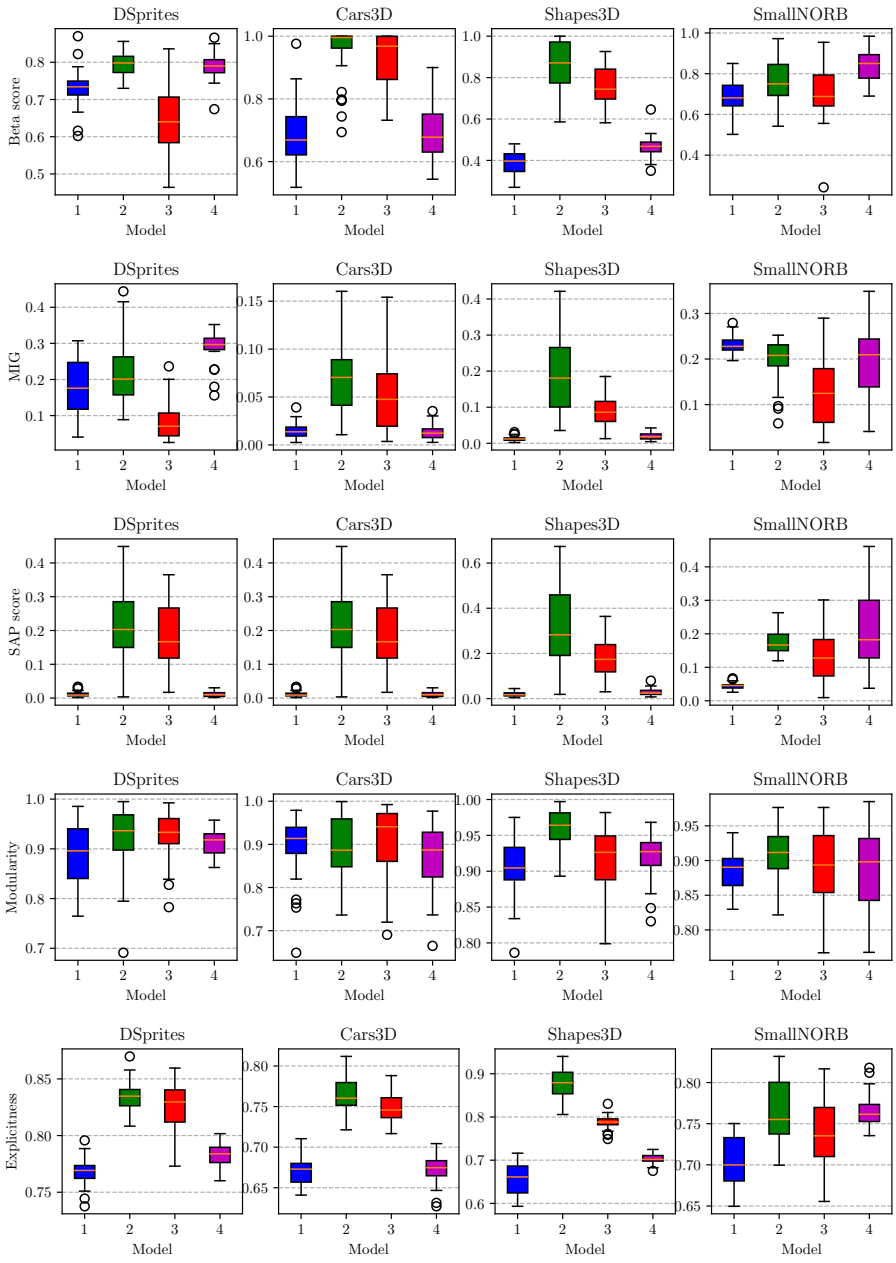

Figure 5: Beta score, MIG, SAP score, Modularity and Explicitness for each method on DSprites, Cars3D, Shapes3D, SmallNORB. (1=β-VAE, 2=IDVAE, 3=IVAE, 4=FullVAE).

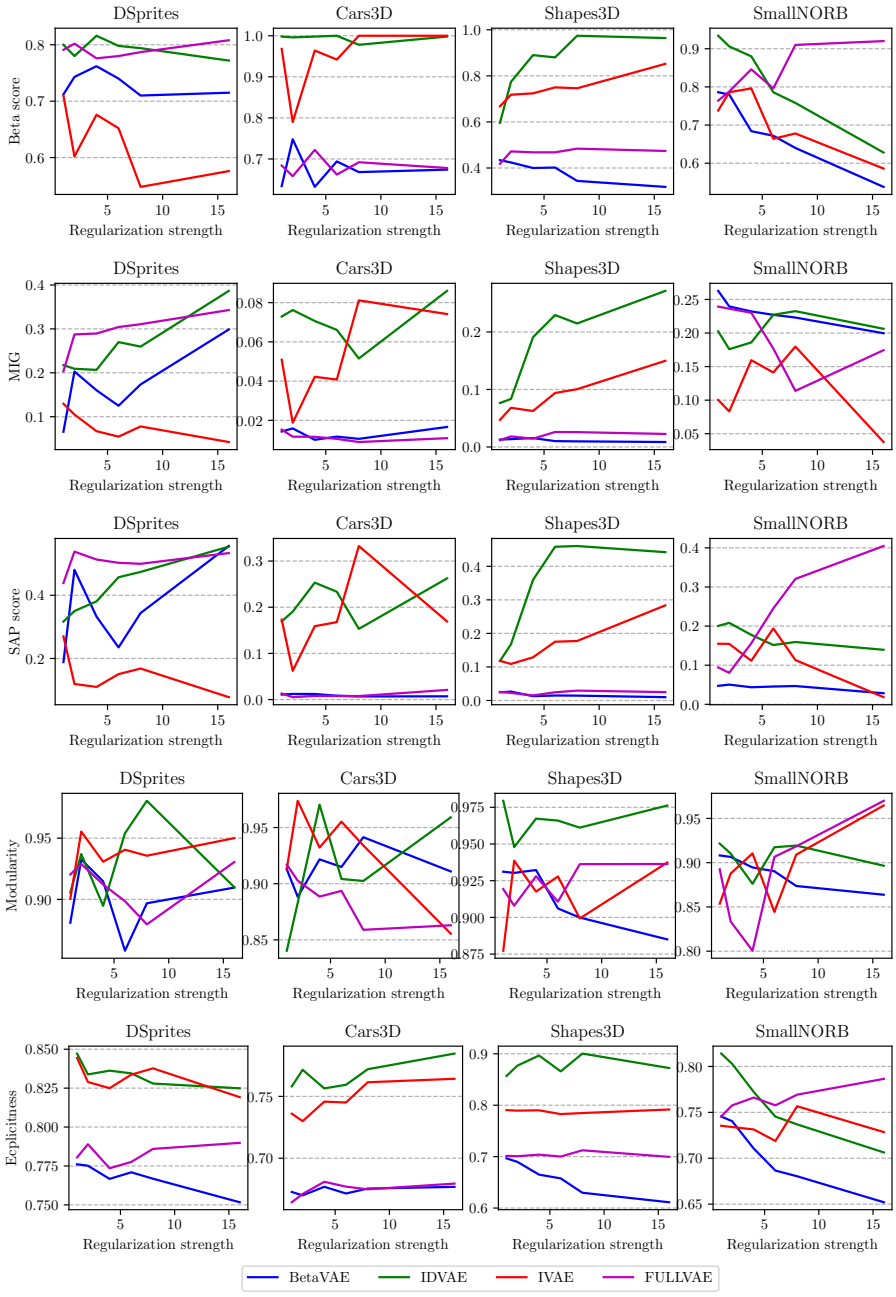

Figure 6: Median of Beta score, MIG, SAP score, Modularity and Explicitness as a function of the regularization strength, for each method on DSprites, Cars3D, Shapes3D, SmallNORB.

Encoder	**Decoder**
Input: $64 \times 64 \times$ number of channels	Input: $\mathbb{R}^d$, where d is the number of ground-truth factors
4×4 conv, 32 ReLU, stride 2	FC, 256 ReLU
4×4 conv, 32 ReLU, stride 2	FC, $4 \times 4 \times 64$ ReLU
4×4 conv, 64 ReLU, stride 2	4×4 upconv, 64 ReLU, stride 2
4×4 conv, 64 ReLU, stride 2	4×4 upconv, 32 ReLU, stride 2
FC 256*, FC $2 \times d$	4×4 upconv, 32 ReLU, stride 2
	4×4 upconv, number of channels, stride 2

Table 2: Main Encoder-Decoder architecture. In IVAE and IDVAE, we give $\mathbf{u}$ as input to the fully connected layer of the Encoder which size becomes $256 + d$.

Conditional Prior Encoder	**Conditional Prior Decoder**
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC, 1000 leaky ReLU	FC, 1000 leaky ReLU
FC $2 \times d$	FC d

Table 3: IDVAE Conditional Prior Encoder-Decoder architecture. IVAE uses the encoder only.

Parameter	**Values**
batch_size	64
optimizer	Adam
Adam: beta1	0.9
Adam: beta2	0.999
Adam: epsilon	1e-8
Adam: learning_rate	1e-4
training_steps	300'000

Table 4: Common hyperparameters to each of the considered methods.

Disentanglement metrics	**Parameters**
Beta score	train_size=10'000, test_size=5'000, batch_size=64, predictor=logistic_regression
MIG	train_size=10'000, n_bins=20
SAP score	train_size=10'000, test_size=5'000, batch_size=16, predictor=linearSVM, C=0.01
Modularity and Explicitness	train_size=10'000, test_size=5'000, batch_size=16, n_bins=20

Table 5: Disentanglement metrics and their parameters.