Disentangling Disentanglement in Variational Autoencoders

Emile Mathieu [*1] **Tom Rainforth** [*1] **N. Siddharth** [*2] **Yee Whye Teh** [1]

Abstract

We develop a generalisation of disentanglement in variational autoencoders (VAEs)—*decomposition* of the latent representation—characterising it as the fulfilment of two factors: a) the latent encodings of the data having an appropriate level of overlap, and b) the aggregate encoding of the data conforming to a desired structure, represented through the prior. Decomposition permits disentanglement, i.e. explicit independence between latents, as a special case, but also allows for a much richer class of properties to be imposed on the learnt representation, such as sparsity, clustering, independent subspaces, or even intricate hierarchical dependency relationships. We show that the β-VAE varies from the standard VAE predominantly in its control of latent overlap and that for the standard choice of an isotropic Gaussian prior, its objective is invariant to rotations of the latent representation. Viewed from the decomposition perspective, breaking this invariance with simple manipulations of the prior can yield better disentanglement with little or no detriment to reconstructions. We further demonstrate how other choices of prior can assist in producing different decompositions and introduce an alternative training objective that allows the control of both decomposition factors in a principled manner.

1. Introduction

An oft-stated motivation for learning disentangled representations of data with deep generative models is a desire to achieve interpretability (Bengio et al., 2013; Chen et al., 2017)—particularly the *decomposability* (see §3.2.1 in Lipton, 2016) of latent representations to admit intuitive explanations. Most work has focused on capturing purely *independent* factors of variation (Alemi et al., 2017; Ansari and Soh, 2019; Burgess et al., 2018; Chen et al., 2018; 2017; Eastwood and Williams, 2018; Esmaeili et al., 2019; Higgins et al., 2016; Kim and Mnih, 2018; Xu and Durrett, 2018; Zhao et al., 2017), typically evaluating this using purpose-built, synthetic data (Eastwood and Williams, 2018; Higgins et al., 2016; Kim and Mnih, 2018), whose generative factors are independent by construction.

This conventional view of disentanglement, as recovering independence, has subsequently motivated the development of formal evaluation metrics for independence (Eastwood and Williams, 2018; Kim and Mnih, 2018), which in turn has driven the development of objectives that target these metrics, often by employing regularisers explicitly encouraging independence in the representations (Eastwood and Williams, 2018; Esmaeili et al., 2019; Kim and Mnih, 2018).

We argue that such an approach is not generalisable, and potentially even harmful, to learning interpretable representations for more complicated problems, where such simplistic representations cannot accurately mimic the generation of high dimensional data from low dimensional latent spaces, and more richly structured dependencies are required.

We posit a generalisation of disentanglement in VAEs—*decomposing* their latent representations—that can help avoid such pitfalls. We characterise decomposition in VAEs as the fulfilment of two factors: a) the latent encodings of data having an appropriate level of overlap, and b) the aggregate encoding of data conforming to a desired structure, represented through the prior. We emphasize that neither of these factors is sufficient in isolation: without an appropriate level of overlap, encodings can degrade to a lookup table where the latents convey little information about data, and without the aggregate encoding of data following a desired structure, the encodings do not decompose as desired.

Disentanglement *implicitly* makes a choice of decomposition: that the latent features are independent of one another. We make this *explicit* and exploit it to both provide improvement to disentanglement through judicious choices of structure in the prior, and to introduce a more general framework flexible enough to capture alternate, more complex, notions of decomposition such as sparsity, clustering, hierarchical structuring, or independent subspaces.

[*]Equal contribution [1]Department of Statistics [2]Department of Engineering, University of Oxford. Correspondence to: Emile Mathieu <emile.mathieu@stats.ox.ac.uk>, Tom Rainforth <rainforth@stats.ox.ac.uk>, N. Siddharth <nsid@robots.ox.ac.uk>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

To connect our framework with existing approaches for encouraging disentanglement, we provide a theoretical analysis of the β-VAE (Alemi et al., 2018; 2017; Higgins et al., 2016), and show that it typically only allows control of latent overlap, the first decomposition factor. We show that it can be interpreted, up to a constant offset, as the standard VAE objective with its prior annealed as $p_\theta(\boldsymbol{z})^\beta$ and an additional maximum entropy regularization of the encoder that increases the stochasticity of the encodings. Specialising this result for the typical choice of a Gaussian encoder and isotropic Gaussian prior indicates that the β-VAE, up to a scaling of the latent space, is equivalent to the VAE plus a regulariser encouraging higher encoder variance. Moreover, this objective is invariant to rotations of the learned latent representation, meaning that it does not, on its own, encourage the latent variables to take on meaningful representations any more than an arbitrary rotation of them.

We confirm these results empirically, while further using our decomposition framework to show that simple manipulations to the prior can improve disentanglement, and other decompositions, with little or no detriment to the reconstruction accuracy. Further, motivated by our analysis, we propose an alternative objective that takes into account the distinct needs of the two factors of decomposition, and use it to learn clustered and sparse representations as demonstrations of alternative forms of decomposition. An implementation of our experiments and suggested methods is provided at http://github.com/iffsid/disentangling-disentanglement.

2. Background and Related Work

2.1. Variational Autoencoders

Let $\boldsymbol{x}$ be an $\mathcal{X}$-valued random variable distributed according to an unknown generative process with density $p_{\mathcal{D}}(\boldsymbol{x})$ and from which we have observations, $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. The aim is to learn a latent-variable model $p_\theta(\boldsymbol{x}, \boldsymbol{z})$ that captures this generative process, comprising of a fixed[1] prior over latents $p(\boldsymbol{z})$ and a parametric likelihood $p_\theta(\boldsymbol{x}|\boldsymbol{z})$. Learning proceeds by minimising a divergence between the true data generating distribution and the model w.r.t θ, typically

$$\arg\min_{\boldsymbol{\theta}} \mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{x}) \parallel p_\theta(\boldsymbol{x})) = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$$

where $p_\theta(\boldsymbol{x}) = \int_{\mathcal{Z}} p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}$ is the marginal likelihood, or evidence, of datapoint $\boldsymbol{x}$ under the model, approximated by averaging over the observations.

However, estimating $p_\theta(\boldsymbol{x})$ (or its gradients) to any sufficient degree of accuracy is typically infeasible. A common strategy to ameliorate this issue involves the introduction of a parametric inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to construct a variational evidence lower bound (ELBO) on $\log p_\theta(\boldsymbol{x})$ as follows

$$\begin{aligned}\mathcal{L}(\boldsymbol{x};\theta,\phi) &\triangleq \log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})).\end{aligned} \tag{1}$$

A variational autoencoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014) views this objective from the perspective of a deep stochastic autoencoder, taking the inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ to be an encoder and the likelihood model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ to be a decoder. Here θ and ϕ are neural network parameters, and learning happens via stochastic gradient ascent (SGA) using unbiased estimates of $\nabla_{\theta,\phi} \frac{1}{n} \sum_{i=1}^n \mathcal{L}(\boldsymbol{x}_i; \theta, \phi)$. Note that when clear from the context, we denote $\mathcal{L}(\boldsymbol{x}; \theta, \phi)$ as simply $\mathcal{L}(\boldsymbol{x})$.

2.2. Disentanglement

Disentanglement, as typically employed in literature, refers to independence among features in a representation (Bengio et al., 2013; Eastwood and Williams, 2018; Higgins et al., 2018). Conceptually, however, it has a long history, far longer than we could reasonably do justice here, and is far from specific to VAEs. The idea stems back to traditional methods such as ICA Hyvärinen and Oja (2000); Yang and Amari (1997) and conventional autoencoders Schmidhuber (1992), through to a range of modern approaches employing deep learning Achille and Soatto (2019); Chen et al. (2016); Cheung et al. (2014); Hjelm et al. (2019); Makhzani et al. (2015); Mathieu et al. (2016); Reed et al. (2014).

Of particular relevance to this work are approaches that explore disentanglement in the context of VAEs Alemi et al. (2017); Chen et al. (2018); Esmaeili et al. (2019); Higgins et al. (2016); Kim and Mnih (2018); Siddharth et al. (2017). Here one aims to achieve independence between the dimensions of the aggregate encoding, typically defined as $q_\phi(\boldsymbol{z}) \triangleq \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}[q(\boldsymbol{z}|\boldsymbol{x})] \approx \frac{1}{n}\sum_i^n q(\boldsymbol{z}|\boldsymbol{x}_i)$. The significance of $q_\phi(\boldsymbol{z})$ is that it is the marginal distribution induced on the latents by sampling a datapoint and then using the encoder to sample an encoding given that datapoint. It can thus informally be thought of as the pushforward distribution for "sampling" representations in the latent space.

Within the disentangled VAEs literature, there is also a distinction between unsupervised approaches, and semi-supervised approaches wherein one has access to the true generative factor values for some subset of data (Bouchacourt et al., 2018; Kingma et al., 2014; Siddharth et al., 2017). Our focus, however, is on the unsupervised setting.

Much of the prior work in the field has either implicitly or explicitly presumed a slightly more ambitious definition of disentanglement than considered above: that it is a measure of how well one captures *true* factors of variation (which happen to be independent by construction for synthetic data), rather than just independent factors. After all, if we wish for our learned representations to be interpretable, it is nec-

[1] Learning the prior is possible, but omitted for simplicity.

essary for the latent variables to take on clear-cut meaning.

One such definition is given by Eastwood and Williams (2018), who define it as the extent to which a latent dimension $d \in D$ in a representation predicts a true generative factor $k \in K$, with each latent capturing at most one generative factor. This implicitly assumes $D \geq K$, as otherwise the latents are unable to explain all the true generative factors. However, for real data, the association is more likely $D \ll K$, with one learning a low-dimensional abstraction of a complex process involving many factors. Consequently, such simplistic representations cannot, by definition, be found for more complex datasets that require more richly structured dependencies to be able to encode the information required to generate higher dimensional data. Moreover, for complex datasets involving a finite set of datapoints, it might not be reasonable to presume that one could capture the elements of the true generative process—the data itself might not contain sufficient information to recover these and even if it does, the computation required to achieve this through model learning is unlikely to be tractable.

The subsequent need for richly structured dependencies between latent dimensions has been reflected in the motivation for a handful of approaches (Bouchacourt et al., 2018; Esmaeili et al., 2019; Johnson et al., 2016; Siddharth et al., 2017) that explore this through graphical models, although employing mutually-inconsistent, and not generalisable, interpretations of disentanglement. This motivates our development of a decomposition framework as a means of extending beyond the limitations of disentanglement.

3. Decomposition: A Generalisation of Disentanglement

The commonly assumed notion of disentanglement is quite restrictive for complex models where the true generative factors are not independent, very large in number, or where it cannot be reasonably assumed that there is a well-defined set of "true" generative factors (as will be the case for many, if not most, real datasets). To this end, we introduce a generalization of disentanglement, *decomposition*, which at a high-level can be thought of as imposing a desired structure on the learned representations. This permits disentanglement as a special case, for which the desired structure is that $q_\phi(\boldsymbol{z})$ factors along its dimensions.

We characterise the decomposition of latent spaces in VAEs to be the fulfilment of two factors (as shown in Figure 1):

a. An "appropriate" level of overlap in the latent space—ensuring that the range of latent values capable of encoding a particular datapoint is neither too small, nor too large. This is, in general, dictated by the level of stochasticity in the encoder: the noisier the encoding process is, the higher the number of datapoints which can plausibly give rise to a particular encoding.

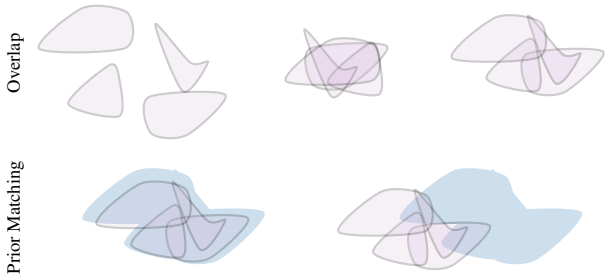

Figure 1. The two factors of decomposition. [Top] Overlap between encodings $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)$, showing cases with (l) too little overlap, (m) too much overlap, and (r) an "appropriate" level of overlap. [Bottom] Illustration of (l) good and (r) bad regularisation between the aggregate posterior $q_\phi(\boldsymbol{z})$ and the desired prior $p(\boldsymbol{z})$.

b. The aggregate encoding $q_\phi(\boldsymbol{z})$ matching the prior $p(\boldsymbol{z})$, where the latter expresses the desired dependency structure between latents.

The overlap factor (a) is perhaps best understood by considering extremes—too little, and the latents effectively become a lookup table; too much, and the data and latents do not convey information about each other. In either case, meaningfulness of the latent encodings is lost. Thus, without the *appropriate* level of overlap—dictated both by noise in the true generative process and dataset size—it is not possible to enforce meaningful structure on the latent space. Though quantitatively formalising overlap in general scenarios is surprisingly challenging (c.f. § 7 and Appendix D), we note for now that when the encoder distribution is unimodal, it is typically well-characterized by the mutual information between the data and the latents $I(\boldsymbol{x}; \boldsymbol{z})$.

The regularisation factor (b) enforces a congruence between the (aggregate) latent embeddings of data and the dependency structures expressed in the prior. We posit that such structure is best expressed in the prior, as opposed to explicit independence regularisation of the marginal posterior (Chen et al., 2018; Kim and Mnih, 2018), to enable the *generative* model to express the desired decomposition, and to avoid potentially violating self-consistency between the encoder, decoder, and true data generating distributions. The prior also provides a rich and flexible means of expressing desired structure by defining a generative process that encapsulates dependencies between variables, as with a graphical model.

Critically, *neither factor is sufficient in isolation*. An inappropriate level of overlap in the latent space will impede interpretability, irrespective of quality of regularisation, as the latent space need not be meaningful. Conversely, without the pressure to regularise to the prior, the latent space is under no constraint to exhibit the desired structure.

Decomposition is inherently subjective as we must choose the structure of the prior we regularise to depending on how we intend to use our learned model or what kind of features we would like to uncover from the data. This may at first

seem unsatisfactory compared to the seemingly objective adjustments often made to the ELBO by disentanglement methods. However, disentanglement *is itself* a subjective choice for the decomposition. We can embrace this subjective nature through judicious choices of the prior distribution; ignoring this imposes unintended assumptions which can have unwanted effects. For example, as we will later show, the rotational invariance of the standard prior $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, I)$ can actually hinder disentanglement.

4. Deconstructing the β-VAE

To connect existing approaches to our proposed framework, we now consider, as a case study, the β-VAE (Higgins et al., 2016)—an adaptation of the VAE objective (ELBO) to learn better-disentangled representations. Specifically, it scales the KL term in the standard ELBO by a factor $\beta > 0$ as

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta \,\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| \, p(\boldsymbol{z})). \quad (2)$$

Hoffman et al. (2017) showed that the β-VAE target can be viewed as a standard ELBO with the alternative prior $r(\boldsymbol{z}) \propto q_\phi(\boldsymbol{z})^{(1-\beta)} p(\boldsymbol{z})^\beta$, along with terms involving the mutual information and the prior's normalising constant.

We now introduce an alternate deconstruction as follows

Theorem 1. *The β-VAE target $\mathcal{L}_\beta(\boldsymbol{x})$ can be interpreted in terms of the standard ELBO, $\mathcal{L}\left(\boldsymbol{x}; \pi_{\theta,\beta}, q_\phi\right)$, for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x}, \boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p(\boldsymbol{z})^\beta / F_\beta$ as*

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathcal{L}\left(\boldsymbol{x}; \pi_{\theta,\beta}, q_\phi\right) + (\beta - 1) H_{q_\phi} + \log F_\beta \quad (3)$$

where $F_\beta \triangleq \int_{\boldsymbol{z}} p(\boldsymbol{z})^\beta d\boldsymbol{z}$ is constant given β, and H_{q_ϕ} is the entropy of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$.

Proof. All proofs are given in Appendix A. □

Clearly, the second term in (3), enforcing a maximum entropy regulariser on the posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$, allows the value of β to affect the overlap of encodings in the latent space. We thus see that it provides a means of controlling decomposition factor (a). However, it is itself not sufficient to enforce disentanglement. For example, the entropy of $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ is independent of its mean $\mu_\theta(\boldsymbol{x})$ and is independent to rotations of $\boldsymbol{z}$, so it is clearly incapable of discouraging certain representations with poor disentanglement. All the same, having the wrong level of regularization can, in turn, lead to an inappropriate level of overlap and undermine the ability to disentangle. Consequently, this term is still important.

Although the precise impact of prior annealing depends on the original form of the prior, the high-level effect is the same—larger values of β cause the effective latent space to collapse towards the modes of the prior. For uni-modal priors, the main effect of annealing is to reduce the scaling of $\boldsymbol{z}$; indeed this is the only effect for generalized Gaussian distributions. While this would appear not to have any tangible effects, closer inspection suggests otherwise—it ensures that the scaling of the encodings matches that of the prior. Only incorporating the maximum-entropy regularisation will simply cause the scaling of the latent space to increase. The rescaling of the prior now cancels this effect, ensuring the scaling of $q_\phi(\boldsymbol{z})$ matches that of $p(\boldsymbol{z})$.

Taken together, this implies that the β-VAE's ability to encourage disentanglement is predominantly through *direct* control over the level of overlap. It places no other direct constraint on the latents to disentangle (although in some cases, the annealed prior may inadvertently encourage better disentanglement), but instead helps avoid the pitfalls of inappropriate overlap. Amongst other things, this explains why large β is not universally beneficial for disentanglement, as the level of overlap can be increased too far.

4.1. Special Case – Gaussians

We can gain further insights into the β-VAE in the common use case—assuming a Gaussian prior, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma)$, and Gaussian encoder, $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x})\right)$. Here it is straightforward to see that annealing simply scales the latent space by $1/\sqrt{\beta}$, i.e. $f_\beta(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma/\beta)$. Given this, it is easy to see that a VAE trained with the adjusted target $\mathcal{L}\left(\boldsymbol{x}; \pi_{\theta,\beta}, q_\phi\right)$, but appropriately scaling the latent space, will behave identically to one trained with the original target $\mathcal{L}(\boldsymbol{x})$. It will also have an identical ELBO as the expected reconstruction is trivially the same, while the KL between Gaussians is invariant to scaling both equally. More precisely, we have the following result.

Corollary 1. *If $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \Sigma)$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x})\right)$, then,*

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}\left(\boldsymbol{x}; \theta', \phi'\right) + \frac{(\beta - 1)}{2} \log |S_{\phi'}(\boldsymbol{x})| + c \quad (4)$$

where θ' and ϕ' represent rescaled networks such that

$$p_{\theta'}(\boldsymbol{x} \mid \boldsymbol{z}) = p_\theta\left(\boldsymbol{x} \mid \boldsymbol{z}/\sqrt{\beta}\right),$$
$$q_{\phi'}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_{\phi'}(\boldsymbol{x}), S_{\phi'}(\boldsymbol{x})\right),$$
$$\mu_{\phi'}(\boldsymbol{x}) = \sqrt{\beta}\mu_\phi(\boldsymbol{x}), \quad S_{\phi'}(\boldsymbol{x}) = \beta S_\phi(\boldsymbol{x}),$$

and $c \triangleq \frac{D(\beta-1)}{2}\left(1 + \log \frac{2\pi}{\beta}\right) + \log F_\beta$ is a constant, with D denoting the dimensionality of $\boldsymbol{z}$.

Noting that as c is irrelevant to the training process, this indicates an equivalence, up to scaling of the latent space, between training with the β-VAE objective and a maximum-entropy regularised version of the standard ELBO

$$\mathcal{L}_{H,\beta}(\boldsymbol{x}) \triangleq \mathcal{L}(\boldsymbol{x}) + \frac{(\beta - 1)}{2} \log |S_\phi(\boldsymbol{x})|, \quad (5)$$

whenever $p(\boldsymbol{z})$ and $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ are Gaussian. Note that we implicitly presume suitable adjustment of neural-network hyper-parameters and the stochastic gradient scheme to account for the change of scaling in the optimal networks.

Moreover, the stationary points for the two objectives $\mathcal{L}_\beta(\boldsymbol{x};\theta,\phi)$ and $\mathcal{L}_{H,\beta}(\boldsymbol{x};\theta',\phi')$ are equivalent (c.f. Corollary 2 in Appendix A), indicating that optimising for (5) leads to networks equivalent to those from optimising the β-VAE objective (2), up to scaling the encodings by a factor of $\sqrt{\beta}$. Under the isotropic Gaussian prior setting, we further have the following result showing that the β-VAE objective is invariant to rotations of the latent space.

Theorem 2. *If* $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z};0,\sigma I)$ *and* $q_\phi(\boldsymbol{z}\mid\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z};\mu_\phi(\boldsymbol{x}),S_\phi(\boldsymbol{x}))$*, then for all rotation matrices* R*,*

$$\mathcal{L}_\beta(\boldsymbol{x};\theta,\phi) = \mathcal{L}_\beta(\boldsymbol{x};\theta^\dagger(R),\phi^\dagger(R)) \tag{6}$$

where $\theta^\dagger(R)$ *and* $\phi^\dagger(R)$ *are transformed networks such that*

$$p_{\theta^\dagger}(\boldsymbol{x}\mid\boldsymbol{z}) = p_\theta\big(\boldsymbol{x}\mid R^T\boldsymbol{z}\big),$$
$$q_{\phi^\dagger}(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z};R\mu_\phi(\boldsymbol{x}),RS_\phi(\boldsymbol{x})R^T\right).$$

This shows that the β-VAE objective does not directly encourage latent variables to take on meaningful representations when using the standard choice of an isotropic Gaussian prior. In fact, on its own, it encourages latent representations which match the true generative factors no more than it encourages *any arbitrary rotation* of these factors, with such rotations capable of exhibiting strong correlations between latents. This view is further supported by our empirical results (see Figure 2), where we did not observe any gains in disentanglement (using the metric from Kim and Mnih (2018)) from increasing $\beta > 0$ with an isotropic Gaussian prior trained on the *2D Shapes* dataset (Matthey et al., 2017). It may also go some way to explaining the extremely high levels of variation we found in the disentanglement-metric scores between different random seeds at train time.

It should be noted, however, that the value of β can indirectly influence the level of disentanglement when using a mean-field assumption for the encoder distribution (i.e. restricting $S_\phi(x)$ to be diagonal). As noted by Rolinek et al. (2018); Stühmer et al. (2019), increasing β can reinforce existing inductive biases, wherein mean-field assumptions encourage representations which reduce dependence between the latent dimensions (Turner and Sahani, 2011).

5. An Objective for Enforcing Decomposition

Given the characterisation set out above, we now develop an objective that incorporates the effect of both factors (a) and (b). Our analysis of the β-VAE tells us that its objective allows direct control over the level of overlap, i.e. factor (a). To incorporate direct control over the regularisation (b) between the marginal posterior and the prior, we add a divergence term $\mathbb{D}(q_\phi(z),p(\boldsymbol{z}))$, yielding

$$\begin{aligned}\mathcal{L}_{\alpha,\beta}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{z})] \\ -\beta\ \mathrm{KL}(q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\,\|\,p(\boldsymbol{z})) - \alpha\,\mathbb{D}(q_\phi(\boldsymbol{z}),p(\boldsymbol{z}))\end{aligned} \tag{7}$$

allowing control over how much factors (a) and (b) are enforced, through appropriate setting of β and α respectively.

Note that such an additional term has been previously considered by Kumar et al. (2017), with $\mathbb{D}(q_\phi(\boldsymbol{z}),p(\boldsymbol{z})) = \mathrm{KL}(q_\phi(\boldsymbol{z})\,\|\,p(\boldsymbol{z}))$, although for the sake of tractability they rely instead on moment matching using covariances. There have also been a number of approaches that decompose the standard VAE objective in different ways (e.g. Dilokthanakul et al., 2019; Esmaeili et al., 2019; Hoffman and Johnson, 2016) to expose $\mathrm{KL}(q_\phi(\boldsymbol{z})\,\|\,p(\boldsymbol{z}))$ as a component, but, as we discuss in Appendix C, this can be difficult to compute correctly in practice, with common approaches leading to highly biased estimates whose practical behaviour is very different than the divergence they are estimating, unless very large batch sizes are used.

Wasserstein Auto-Encoders (Tolstikhin et al., 2018) formulate an objective that includes a general divergence term between the prior and marginal posterior, computed using either maximum mean discrepancy (MMD) or a variational formulation of the Jensen-Shannon divergence (a.k.a GAN loss). However, we find that empirically, choosing the MMD's kernel and numerically stabilising its U-statistics estimator to be tricky, and designing and learning a GAN to be cumbersome and unstable. Consequently, the problems of choosing an appropriate $\mathbb{D}(q_\phi(\boldsymbol{z}),p(\boldsymbol{z}))$ and generating reliable estimates for this choice are tightly coupled, with a general purpose solution remaining an important open problem; see further discussion in Appendix C.

6. Experiments

6.1. Prior for Axis-Aligned Disentanglement

We first show how subtle changes to the prior distribution can yield improvements in disentanglement. The standard choice of an isotropic Gaussian has previously been justified by the correct assertion that the latents are independent under the prior (Higgins et al., 2016). However, as explained in § 4.1, the rotational invariance of this prior means that it does not directly encourage axis-aligned representations. Priors that break this rotational invariance should be better suited for learning disentangled representations. We assess this hypothesis by training a β-VAE (i.e. (7) with $\alpha = 0$) on the *2D Shapes* dataset (Matthey et al., 2017) and evaluating disentanglement using the metric of Kim and Mnih (2018).

Figure 2 demonstrates that notable improvements in disentanglement can be achieved by using non-isotropic priors: for a given reconstruction loss, implicitly fixed by β, non-isotropic Gaussian priors got better disentanglement scores, with further improvement achieved when the prior variance is learnt. With a product of Student-t priors $p_\nu(\boldsymbol{z})$ (noting $p_\nu(\boldsymbol{z}) \to \mathcal{N}(\boldsymbol{z};\mathbf{0},\mathbf{I})$ as $\nu\to\infty$), reducing ν only incurred a minor reconstruction penalty, for improved disentanglement.

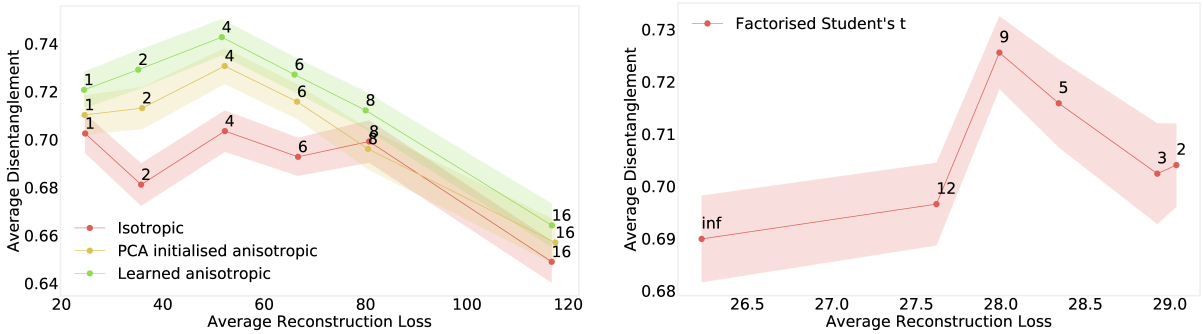

Figure 2. Reconstruction loss vs disentanglement metric of Kim and Mnih (2018). [Left] Using an anisotropic Gaussian with diagonal covariance either learned, or fixed to principal-component values of the dataset. Point labels represent different values of β. [Right] Using $p_\nu(\boldsymbol{z}) = \prod_d \text{STUDENT-T}(z_d; \nu)$ for different ν with $\beta = 1$. Note the different x-axis scaling. Shaded areas represent ± 2 standard errors for estimated mean disentanglement calculated using 100 separately trained networks. We thus see that the variability on the disentanglement metric is very large, presumably because of stochasticity in whether learned dimensions correspond to true generative factors. The variability in the reconstruction was only negligible and so is not shown. See Appendix B for full experimental details.

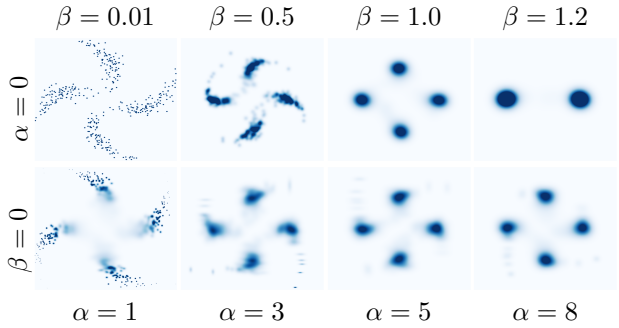

Figure 3. Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for spirals dataset with a mixture of Gaussian prior.

Interestingly, very low values of ν caused the disentanglement score to drop again (though still giving higher values than the Gaussian). We speculate that this may be related to the effect of heavy tails on the disentanglement metric itself, rather than being an objectively worse disentanglement. Another interesting result was that for an isotropic Gaussian prior, as per the original β-VAE setup, no gains at all were achieved in disentanglement by increasing β.

6.2. Clustered Prior

We next consider an alternative decomposition one might wish to impose—*clustering* of the latent space. For this, we use the "pinwheels" dataset from (Johnson et al., 2016) and a mixture of four equally-weighted Gaussians as our prior. We then conduct an ablation study to observe the effect of varying α and β in $\mathcal{L}_{\alpha,\beta}(\mathbf{x})$ (as per (7)) on the learned representations, taking the divergence to be $\text{KL}\left(p(\boldsymbol{z})||q_\phi(\boldsymbol{z})\right)$ (see Appendix B for details).

We see in Figure 3 that increasing β increases the level of overlap in $q_\phi(\boldsymbol{z})$, as a consequence of increasing the encoder variance for individual datapoints. When β is too large, the encoding of a datapoint loses meaning. Also, as a single datapoint encodes to a Gaussian distribution, $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is unable to match $p(\boldsymbol{z})$ exactly. Because $q_\phi(\boldsymbol{z}|\boldsymbol{x}) \to q_\phi(\boldsymbol{z})$ when $\beta \to \infty$, this in turn means that overly large values of β actually cause a mismatch between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$ (see top right of Figure 3). Increasing α, instead always improved the match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$. Here, the finiteness of the dataset and the choice of divergence results in an increase in overlap with increasing α, but only up to the level required for a non-negligible overlap between the nearby datapoints: large values of α did not cause the encodings to collapse to a mode.

6.3. Prior for Sparsity

Finally, we consider a commonly desired decomposition—sparsity, which stipulates that only a small fraction of available factors are employed. That is, a *sparse representation* (Olshausen and Field, 1996) can be thought of as one where each embedding has a significant proportion of its dimensions *off*, i.e. close to 0. Sparsity has often been considered for feature-learning (Coates and Ng, 2011; Larochelle and Bengio, 2008) and employed in the probabilistic modelling literature (Lee et al., 2007; Ranzato et al., 2007).

Common ways to achieve sparsity are through a specific penalty (e.g. l_1) or a careful choice of prior (peaked at 0). Concomitant with our overarching desire to encode requisite structure in the prior, we adopt the latter, constructing a sparse prior as $p(\boldsymbol{z}) = \prod_d (1-\gamma)\,\mathcal{N}(z_d; 0, 1) + \gamma\,\mathcal{N}(z_d; 0, \sigma_0^2)$ with $\sigma_0^2 = 0.05$. This mixture distribution can be interpreted as a mixture of samples being either *off* or *on*, whose proportion is set by the weight parameter γ. We use this prior to learn a VAE for the *Fashion-MNIST* dataset (Xiao et al., 2017) using the objective $\mathcal{L}_{\alpha,\beta}(\mathbf{x})$ (as per (7)), taking the divergence to be an MMD with a kernel that only considers difference between the marginal distributions (see Appendix B for details).

We measure a representation's sparsity using the *Hoyer*

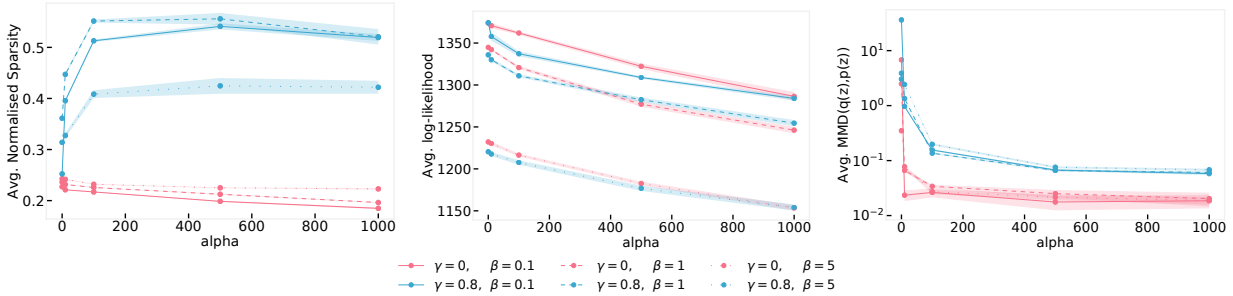

Figure 4. [Left] Sparsity vs regularisation strength α (c.f. (7), high better). [Center] Average reconstruction log-likelihood $\mathbb{E}_{p_D(\boldsymbol{x})}[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]]$ vs α (higher better). [Right] Divergence (MMD) vs α (lower better). Note here that the different values of γ represent regularizations to different distributions, with regularization to a Gaussian (i.e. $\gamma = 0$) much easier to achieve than the sparse prior, hence the lower divergence. Shaded areas represent ± 2 standard errors in the mean estimate calculated using 8 separately trained networks. See Appendix B for full experimental details.

extrinsic metric (Hurley and Rickard, 2008). For $\boldsymbol{y} \in \mathbb{R}^d$,

$$\text{Hoyer}\,(\boldsymbol{y}) = \frac{\sqrt{d} - \|\boldsymbol{y}\|_1/\|\boldsymbol{y}\|_2}{\sqrt{d}-1} \in [0,1],$$

yielding 0 for a fully dense vector and 1 for a fully sparse vector. Rather than employing this metric directly to the mean encoding of each datapoint, we first normalise each dimension to have a standard deviation of 1 under its aggregate distribution, i.e. we use $\bar{z}_d = z_d/\sigma(z_d)$ where $\sigma(z_d)$ is the standard deviation of dimension d of the latent encoding taken over the dataset. This normalisation is important as one could achieve a "sparse" representation simply by having different dimensions vary along different length scales (something the β-VAE encourages through its pruning of dimensions (Stühmer et al., 2019)), whereas we desire a representation where different datapoints "activate" different features. We then compute overall sparsity by averaging over the dataset as Sparsity $= \frac{1}{n}\sum_i^n$ Hoyer $(\bar{\boldsymbol{z}}_i)$. Figure 4 (left) shows that substantial sparsity can be gained by replacing a Gaussian prior ($\gamma = 0$) by a sparse prior ($\gamma = 0.8$). It further shows substantial gains from the inclusion of the aggregate posterior regularization, with $\alpha = 0$ giving far low sparsity than $\alpha > 0$, when using our sparse prior. The use of our sparse prior did not generally harm the reconstruction compared. Large values of α did slightly worsen the reconstruction, but this drop-off was much slower than increases in β (note that α is increased to much higher levels than β). Interestingly, we see that β being either too low or too high also harmed the sparsity.

We explore the qualitative effects of sparsity in Figure 5, using a network trained with $\alpha = 1000, \beta = 1$, and $\gamma = 0.8$, corresponding to one of the models in Figure 4 (left). The top plot shows the average encoding magnitude for data corresponding to 3 of the 10 classes in the *Fashion-MNIST* dataset. It clearly shows that the different classes (trousers, dress, and shirt) predominantly encode information along different sets of dimensions, as expected for sparse representations (c.f. Appendix B for plots for all classes). For each of these classes, we explore the latent space along a particular 'active' dimension—one with high average encoding magnitude—to observe if they capture meaningful features in the image. We first identify a suitable 'active' dimension for a given instance (top row) from the dimension-wise magnitudes of its encoding, by choosing one, say d, where the magnitude far exceeds σ_0^2. Given encoding value $\boldsymbol{z}_d$, we then interpolate along this dimension (keeping all others fixed) in the range $(\boldsymbol{z}_d, \boldsymbol{z}_d + \text{sign}(\boldsymbol{z}_d))$; the sign of $\boldsymbol{z}_d$ indicating the direction of interpolation. Exploring the latent space in such a manner demonstrates a variety of consistent feature transformations in the image, both within class (a, b, c), and across classes (d), indicating that these sparse dimensions do capture meaningful features in the image.

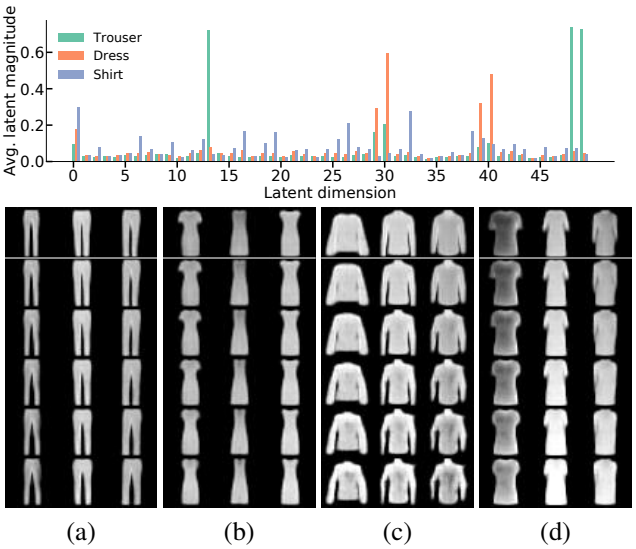

Figure 5. Qualitative evaluation of sparsity. [Top] Average encoding magnitude over data for three example classes in *Fashion-MNIST*. [Bottom] Latent interpolation (↓) for different datapoints (top layer) along particular 'active' dimensions. (a) Separation between the legs of trousers (dim 49). (b) Top/Collar width of dresses (dim 30). (c) Shirt shape (loose/fitted, dim 19). (d) Style of sleeves across different classes—t-shirt, dress, and coat (dim 40).

Concurrent to our work, Tonolini et al. (2019) also considered imposing sparsity in VAEs with a spike-slab prior (such

that $\sigma_0 \to 0$). In contrast to our work, they do not impose a constraint on the aggregate encoder, nor do they evaluate their results with a quantitative sparsity metric that accounts for the varying length scales of different latent dimensions

7. Discussion

Characterising Overlap Precisely formalising what constitutes the level of overlap in the latent space is surprisingly subtle. Prior work has typically instead considered controlling the level of compression through the mutual information between data and latents $I(\boldsymbol{x}; \boldsymbol{z})$ (Alemi et al., 2018; 2017; Hoffman and Johnson, 2016; Phuong et al., 2018), with, for example, (Phuong et al., 2018) going on to discuss how controlling the compression can "explicitly encourage useful representations." Although $I(\boldsymbol{x}; \boldsymbol{z})$ provides a perfectly serviceable characterisation of overlap in a number of cases, the two are not universally equivalent and we argue that it is the latter which is important in achieving useful representations. In particular, if the form of the encoding distribution is not fixed—as when employing normalising flows, for example—$I(\boldsymbol{x}; \boldsymbol{z})$ does not necessarily characterise overlap well. We discuss this in greater detail in Appendix D.

However, when the encoder is unimodal with fixed form (in particularly the tail behaviour is fixed) and the prior is well-characterised by Euclidean distances, then these factors have a substantially reduced ability to vary for a given $I(\boldsymbol{x}; \boldsymbol{z})$, which subsequently becomes a good characterisation of the level of overlap. When $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is Gaussian, controlling the variance of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ (with a fixed $q_\phi(\boldsymbol{z})$) should similarly provide an effective means of achieving the desired overlap behaviour. As this is the most common use case, we leave the development of more a general definition of overlap to future work, simply noting that this is an important consideration when using flexible encoder distributions.

Can VAEs Uncover True Generative Factors? In concurrently published work, Locatello et al. (2019) question the plausibility of learning unsupervised disentangled representations with meaningful features, based on theoretical analyses showing an equivalence class of generative models, many members of which could be entangled. Though their analysis is sound, we posit a counterargument to their conclusions, based on the *stochastic* nature of the encodings used during training. Namely, that this stochasticity means that they need not give rise to the same ELBO scores (an important exception is the rotational invariance for isotropic Gaussian priors). Essentially, the encoding noise forces nearby encodings to relate to similar datapoints, while standard choices for the likelihood distribution (e.g. assuming conditional independence) ensure that information is stored in the encodings, not just in the generative network. These restrictions mean that the ELBO prefers smooth representations and, provided the prior is not rotationally invariant, means that there no longer need be a class of different representations with the same ELBO; simpler representations are preferred to more complex ones.

The exact form of the encoding distribution is also important here. For example, imagine we restrict the encoder variance to be isotropic and then use a two dimensional prior where one latent dimension has a much larger variance than the other. It will be possible to store more information in the prior dimension with higher variance (as we can spread points out more relative to the encoder variance). Consequently, that dimension is more likely to correspond to an important factor of the generative process than the other. Of course, this does not imply that this is a true factor of variation in the generative process, but neither is the meaning that can be attributed to each dimension completely arbitrary.

All the same, we agree that an important area for future work is to assess when, and to what extent, one might expect learned representations to mimic the true generative process, and, critically, when it should not. For this reason, we actively avoid including any notion of a true generative process in our definition of decomposition, but note that, analogously to disentanglement, it permits such extension in scenarios where doing so can be shown to be appropriate.

8. Conclusions

In this work, we explored and analysed the fundamental characteristics of learning disentangled representations, and showed how these can be generalised to a more general framework of *decomposition* (Lipton, 2016). We characterised the learning of decomposed latent representation with VAEs in terms of the control of two factors: i) overlap in the latent space between encodings of different datapoints, and ii) regularisation of the aggregate encoding distribution to the given prior, which encodes the structure one would wish for the latent space to have.

Connecting prior work on disentanglement to this framework, we analysed the β-VAE objective to show that its contribution to disentangling is primarily through direct control of the level of overlap between encodings of the data, expressed by maximising the entropy of the encoding distribution. In the commonly encountered case of assuming an isotropic Gaussian prior and an independent Gaussian posterior, we showed that control of overlap is the *only* effect of the β-VAE. Motivated by this observation, we developed an alternate objective for the ELBO that allows control of the two factors of decomposability through an additional regularisation term. We then conducted empirical evaluations using this objective, targeting alternate forms of decompositions such as clustering and sparsity, and observed the effect of varying the extent of regularisation to the prior on the quality of the resulting clustering and sparseness of the learnt embeddings. The results indicate that we were successful in attaining those decompositions.

Acknowledgements

EM, TR, YWT were supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 617071. TR research leading to these results also received funding from EPSRC under grant EP/P026753/1. EM was also supported by Microsoft Research through its PhD Scholarship Programme. NS was funded by EPSRC/MURI grant EP/N019474/1.

References

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50), 2019.

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken ELBO. In *International Conference on Machine Learning*, pages 159–168, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Abdul Fatir Ansari and Harold Soh. Hyperprior induced unsupervised disentanglement of latent representations. In *AAAI Conference on Artificial Intelligence*, 2019.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence*, 2018.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loïc Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *CoRR*, abs/1804.03599, 2018.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. Variational Lossy Autoencoder. 2017.

Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 921–928. Omnipress, 2011.

Nat Dilokthanakul, Nick Pawlowski, and Murray Shanahan. Explicit information placement on latent variables using auxiliary generative modelling task, 2019. URL https://openreview.net/forum?id=H1l-SjA5t7.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Babak Esmaeili, Hao Wu, Sarthak Jain, N Siddharth, Brooks Paige, and Jan-Willem van de Meent. Hierarchical Disentangled Representations. *Artificial Intelligence and Statistics*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the International Conference on Learning Representations*, 2016.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*, 2019.

Matthew D Hoffman and Matthew J Johnson. ELBO surgery: yet another way to carve up the variational evidence lower bound. In *Workshop on Advances in Approximate Bayesian Inference, NIPS*, pages 1–4, 2016.

Matthew D Hoffman, Carlos Riquelme, and Matthew J Johnson. The β-VAE's Implicit Prior. In *Workshop on Bayesian Deep Learning, NIPS*, pages 1–5, 2017.

Niall P. Hurley and Scott T. Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55:4723–4741, 2008.

Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000.

Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in Neural Information Processing Systems*, pages 2946–2954. 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, 2018.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. *arXiv.org*, November 2017.

Hugo Larochelle and Yoshua Bengio. Classification using discriminative restricted boltzmann machines. In *International Conference on Machine Learning*, pages 536–543, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, pages 801–808. MIT Press, 2007.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *International Conference on Machine Learning*, 2019.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pages 5040–5048, 2016.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

Mary Phuong, Max Welling, Nate Kushman, Ryota Tomioka, and Sebastian Nowozin. The mutual autoencoder: Controlling information in latent code representations, 2018. URL https://openreview.net/forum?id=HkbmWqxCZ.

Marc Ranzato, Christopher Poultney, Sumit Chopra, and Yann L. Cun. Efficient learning of sparse representations with an energy-based model. In *Advances in Neural Information Processing Systems*, pages 1137–1144. MIT Press, 2007.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. 2014.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). *arXiv preprint arXiv:1812.06775*, 2018.

Jürgen Schmidhuber. Learning factorial codes by predictability minimization. *Neural Computation*, 4(6):863–879, 1992.

N. Siddharth, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pages 5925–5935, 2017.

Jan Stühmer, Richard Turner, and Sebastian Nowozin. ISA-VAE: Independent subspace analysis with variational autoencoders, 2019. URL https://openreview.net/forum?id=rJl_NhR9K7.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations*, 2018.

Francesco Tonolini, Bjorn Sand Jensen, and Roderick Murray-Smith. Variational sparse coding, 2019. URL https://openreview.net/forum?id=SkeJ6iR9Km.

Richard E. Turner and Maneesh Sahani. Two problems with variational expectation maximisation for time-series models. *D. Barber, T. Cemgil, and S. Chiappa (eds.), Bayesian Time series models, chapter 5*, page 109–130, 2011.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Jiacheng Xu and Greg Durrett. Spherical Latent Spaces for Stable Variational Autoencoders. In *Conference on Empirical Methods in Natural Language Processing*, 2018.

Howard Hua Yang and Shun-ichi Amari. Adaptive online learning algorithms for blind separation: maximum entropy and minimum mutual information. *Neural computation*, 9(7):1457–1482, 1997.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *CoRR*, abs/1706.02262, 2017. URL http://arxiv.org/abs/1706.02262.