Independent Subspace Analysis for Unsupervised Learning of Disentangled Representations

Jan Stühmer
Microsoft Research

Richard E. Turner
University of Cambridge

Sebastian Nowozin
Google Brain[1]

Abstract

Recently there has been an increased interest in unsupervised learning of disentangled representations using the Variational Autoencoder (VAE) framework. Most of the existing work has focused largely on modifying the variational cost function to achieve this goal. We first show that these modifications, e.g. β-VAE, simplify the tendency of variational inference to underfit causing pathological overpruning and over-orthogonalization of learned components. Second we propose a complementary approach: to modify the probabilistic model with a structured latent prior. This prior allows to discover latent variable representations that are structured into a hierarchy of independent vector spaces. The proposed prior has three major advantages: First, in contrast to the standard VAE normal prior the proposed prior is not rotationally invariant. This resolves the problem of unidentifiability of the standard VAE normal prior. Second, we demonstrate that the proposed prior encourages a disentangled latent representation which facilitates learning of disentangled representations. Third, extensive quantitative experiments demonstrate that the prior significantly mitigates the trade-off between reconstruction loss and disentanglement over the state of the art.

1 Introduction

Recently there has been an increased interest in unsupervised learning of disentangled representations. The term *disentangled* usually describes two main objectives: First, to identify each true factor of variation with a latent variable, and second, interpretability of these latent factors (Schmidhuber, 1992; Ridgeway, 2016; Achille and Soatto, 2017). Most of this recent work is inspired by the β-VAE concept introduced in Higgins et al. (2016), which proposes to re-weight the terms in the evidence lower bound (ELBO) objective. In Higgins et al. (2016) a higher weight for the Kullback-Leibler divergence (KL) between approximate posterior and prior is proposed, and putative mechanistic explanations for the effects of this modification are studied in Burgess et al. (2017) and Chen et al. (2018). Two recent approaches, Kim and Mnih (2018) and Chen et al. (2018), propose to penalize the total correlation between the dimensions of the latent representation, therefore encouraging a factorized distribution.

These modifications of the evidence lower bound however lead to a trade-off between disentanglement and reconstruction loss and therefore the quality of the learned model. This trade-off is directly encoded in the modified objective: by increasing the β-weight of the KL-term, the relative weight of the reconstruction loss term is more and more decreased. Therefore, optimization of the modified ELBO will lead to latent encodings which have a lower KL-divergence from the prior, but at the same time lead to a higher reconstruction loss. Furthermore, we discuss in section 2.4 that using a higher weight for the KL-term amplifies existing biases of variational inference, potentially to a catastrophic extent.

There is a foundational contradiction in many approaches to disentangling deep generative models (DGMs): the standard model employed is not identifiable as it employs a standard normal prior which then undergoes a linear transformation. Any rotation of the latent space can be absorbed into the linear transform and is therefore statistically indistinguishable. If interpretability is desired, the modelling choices are setting us up to fail.

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

[1] Work done while at Microsoft Research.

We make the following contributions:

- We show that current state of the art approaches based on modified cost functions employ a trade-off between reconstruction loss and disentanglement of the latent representation.
- In section 2.3 we show that variational inference techniques are biased: the estimated components are biased towards having orthogonal effects on the data and the number of components is underestimated.
- We provide a novel description of the origin of disentanglement in β-VAE and demonstrate in section 2.4 that increasing the weight of the KL term increases the over-pruning bias of variational inference.
- To mitigate these drawbacks of existing approaches, we propose a family of rotationally asymmetric distributions for the latent prior, which removes the rotational ambiguity from the model.
- The prior allows to decompose the latent space into independent subspaces. Experiments demonstrate that this prior facilitates disentangled representations even for the unmodified ELBO objective.
- Extensive quantitative experiments demonstrate that the prior significantly mitigates the trade-off between disentanglement and reconstruction quality.

2 Background

We briefly discuss previous work on variational inference in deep generative models and two modifications of the learning objective that have been proposed to learn a disentangled representation. We discuss characteristic biases of variational inference and how the modifications of the learning objective actually accentuate these biases.

2.1 Disentangled Representation Learning

Variational Autoencoder The variational autoencoder introduced in Kingma and Welling (2014) combines a generative model, the decoder, with an inference network, the encoder. Training is performed by optimizing the *evidence lower bound* (ELBO) averaged over the empirical distribution:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - D_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\,, \quad (1)$$

where the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ is a deep learning model with parameters θ and $\boldsymbol{z}$ is sampled from the encoder $\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ with variational parameters ϕ. When choosing appropriate families of distributions, gradients through the samples $\boldsymbol{z}$ can be estimated using the *reparameterization trick*. The approximate posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is usually modelled as a multivariate Gaussian with diagonal covariance matrix and the prior $p(\boldsymbol{z})$ is typically the standard normal distribution.

β-VAE Higgins et al. (2016) propose to modify the evidence lower bound objective and penalize the KL-divergence of the ELBO:

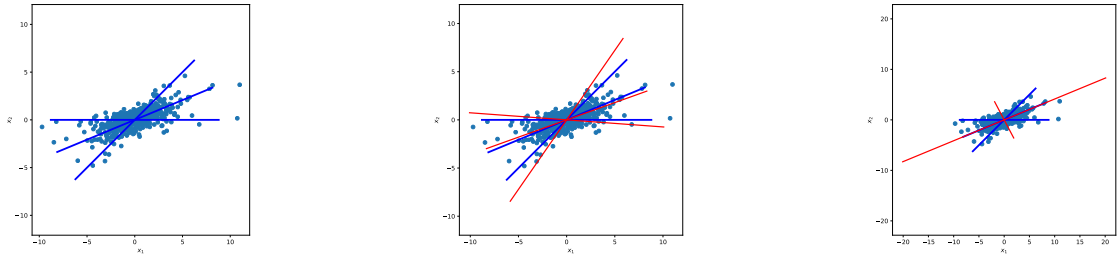
$$\mathcal{L}_{\beta\text{-ELBO}} = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right] - \beta D_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z})), \quad (2)$$

where $\beta > 1$ is a free parameter that should encourage a disentangled representation. In Burgess et al. (2017) the authors provide further thoughts on the mechanism that leads to these disentangled representations. However we will show in section 2.4 that this parameter amplifies biases of variational inference towards orthogonalization and pruning.

β-TCVAE Chen et al. (2018) propose an alternative decomposition of the ELBO, that leads to the recent variant of β-VAE called β-TCVAE. They demonstrate that β-TCVAE allows to learn representations with higher MIG score than β-VAE (Higgins et al., 2016), InfoGAN (Chen et al., 2016) and FactorVAE (Kim and Mnih, 2018). The authors propose to decompose the KL-term in the ELBO objective into three parts and to weight them independently:

$$\begin{aligned}
&\mathbb{E}_{p_\theta(\boldsymbol{x})}\left[D_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\right] = \\
&= D_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|q_\phi(\boldsymbol{z})p_\theta(\boldsymbol{x}))+ \\
&+ D_{\text{KL}}(q_\phi(\boldsymbol{z})\|\prod_j q_\phi(\boldsymbol{z}_j))+ \\
&+ \sum_j D_{\text{KL}}(q_\phi(\boldsymbol{z}_j)\|p(\boldsymbol{z}_j))\,. \quad (3)
\end{aligned}$$

The first term is the index-code mutual information, the second term is the total correlation and the third term the dimension-wise KL-divergence. Because the index-code mutual information can be viewed as an estimator for the mutual information between $p_\theta(\boldsymbol{x})$ and $q_\phi(\boldsymbol{z})$, the authors propose to exclude this term when reweighting the KL-term with the β weight. In addition to the improved objective, the authors propose a quantitative metric of disentanglement, the mutual information gap (MIG). To compute this metric, first the mutual information between every latent variable and the underlying generative factors of the dataset are evaluated. The mutual information gap is then defined as the difference of the mutual information between the latent variables with highest and second highest correlation with an underlying factor.



(a) Sparse student's t-distribution generated by three latent components.

(b) Variational inference with $\beta = 1.0$.

(c) Variational inference with $\beta = 5.0$.

Figure 1: The modified ELBO of β-VAE emphasizes orthogonalization and pruning with increasing β-weight

2.2 Related Work

In addition to the work mentioned already, we briefly review some of the influential papers: Chen et al. (2016) present a variant of a GAN that encourages an interpretable latent representation by maximizing the mutual information between the observation and a small subset of latent variables. The approach relies on optimizing a lower bound of the intractable mutual information. Kim and Mnih (2018) propose a learning objective equivalent to β-TCVAE, and train it with the density ratio trick (Sugiyama et al., 2012). Kumar et al. (2017) introduce a regulariser of the KL-divergence between the approximate posterior and the prior distribution. A parallel line of research proposes not to train a perfect generative model but instead to find a simpler representation of the data (Vedantam et al., 2017; Hinton et al., 2011a). A similar strategy is followed in semi-supervised approaches that require implicit or explicit knowledge about the true underlying factors of the data (Kulkarni et al., 2015; Kingma et al., 2014; Reed et al., 2014; Baydin et al., 2017; Hinton et al., 2011b; Zhu et al., 2017; Goroshin et al., 2015; Hsu et al., 2017; Denton et al., 2017). Existing work on structured priors for VAEs, the VAMP prior by Tomczak and Welling (2017) and the LORACs prior by Vikram et al. (2018), are modelling a clustered latent space. In our work however we introduce a prior which is a latent subspace model. The recent work of Locatello et al. (2019) challenges the whole field of unsupervised representation learning and presents a proof of unidentifiability of the latent representation. We want to emphasize however, that the proof presented in Locatello et al. (2019) only holds for priors that factorize over every latent dimension. A property which does not hold for the prior proposed in this work.

2.3 Orthogonalization and Pruning in Variational Inference

There have been several interpretations of the behaviour of the β-VAE (Chen et al., 2018; Burgess et al., 2017; Rolinek et al., 2019). Here we provide a complementary perspective: that it enhances well known statistical biases in VI (Turner and Sahani, 2011) to produce disentangled, but not necessarily useful, representations. The form of these biases can be understood by considering the variational objective when written as an explicit lower-bound: the log-likelihood of the parameters minus the KL divergence between the approximate posterior and the true posterior

$$\mathcal{L}_{\text{ELBO}} = \log p_\theta(\boldsymbol{x}) - D_{\text{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p_\theta(\boldsymbol{z}|\boldsymbol{x}))\,. \quad (4)$$

From this form it is clear that VI's estimates of the parameters θ will be biased away from the maximum likelihood solution (the maximizer of the first term) in a direction that reduces the KL between the approximate and true posteriors. When factorized approximating distributions are used, VI will therefore be biased towards settings of the parameters that reduce the statistical dependence between the latent variables in the posterior. For example, this will bias learned components towards orthogonal directions in the output space as this reduces explaining away (e.g. in the factor analysis model, VI breaks the degeneracy of the maximum-likelihood solution finding the orthogonal PCA directions, see appendix B.8). Moreover, these biases often cause components to be pruned out (in the sense that they have no effect on the observed variables) since then their posterior sits at the prior, which is typically factorized (e.g. in an over-complete factor analysis model VI prunes out components to return a complete model, see appendix B.8). For simple linear models these effects are not pathological: indeed VI is arguably selecting from amongst the degenerate maximum likelihood solutions in a sensible way. However, for more complex models the biases are more severe: often the true posterior of the underlying model has significant dependencies (e.g. due to explaining away) and the biases can prevent the discovery of some components. For example, VAEs are known to over-prune (Burda et al., 2015; Cremer et al., 2018).

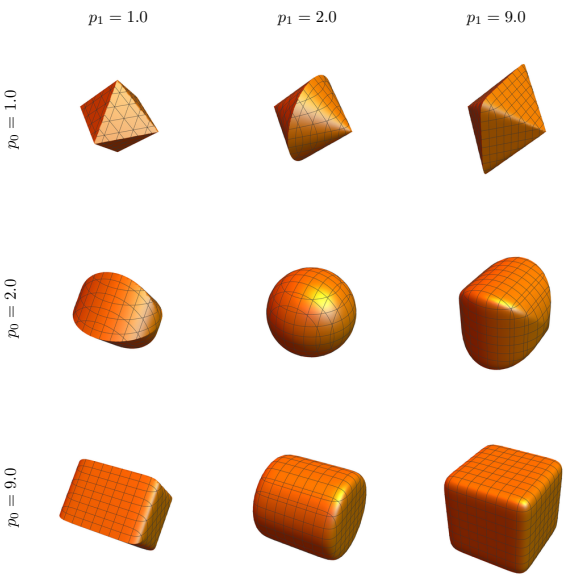


Figure 2: Iso-contours of the L^p-nested function example in equation 6 for combinations of $p_0, p_1 \in \{1, 2, 9\}$.

2.4 β-VAE Emphasizes Orthogonalization and Pruning

What happens to these biases in the β-VAE generalization when $\beta > 1$? The short answer is that they grow. This can be understood by considering coordinate ascent of the modified objective. With θ fixed, optimising q finds a solution that is closer to the prior distribution than VI due to the upweighting of the KL term in 2. With q fixed, optimization over θ returns the same solution as VI (since the prior does not depend on the parameters θ and so the value of β is irrelevant). However, since q is now closer to the prior than before, the KL bias in equation 2 will be greater. These effects are shown in the ICA example in figure 1. Also refer to appendix B.8 for further details. VI ($\beta = 1$) learns components that are more orthogonal than the underlying ones, but $\beta = 5$ prunes out one component entirely and sets the other two to be orthogonal. This is disentangled, but arguably leads to incorrect interpretation of the data. This happens even though both methods are initialised at the true model. Arguably, the β-VAE is enhancing a statistical bug in VI and leveraging this as a feature. We believe that this can be dangerous, preventing the discovery of the underlying model.

3 Latent Prior Distributions for Unsupervised Factorization

In this section we describe an approach for unsupervised learning of disentangled representations. Instead of modifying the ELBO-objective, we propose to use certain families of prior distributions $p(\boldsymbol{z})$, that lead to identifiable and interpretable models. In contrast to the standard normal distribution, the proposed priors are not rotationally invariant, and therefore allow interpretability of the latent space.

3.1 Independent Component Analysis

Independent Component Analysis (ICA) seeks to factorize a distribution into non-Gaussian factors. In order to avoid the ambiguities of latent space rotations, a non-Gaussian distribution (e.g. Laplace or Student-t distribution) is used as prior for the latent variables.

Generalized Gaussian Distribution A generalized version of ICA (Lee and Lewicki, 2000; Zhang et al., 2004; Lewicki, 2002; Sinz and Bethge, 2010) uses a prior from the family of *exponential power distributions* of the form

$$p_{\text{ICA}}(\boldsymbol{z}) \propto \exp\left(-\tau||\boldsymbol{z}||_p^p\right) \tag{5}$$

also called *generalized Gaussian*, *generalized Laplacian* or *p-generalized normal* distribution. Using $p = 2/(1+\kappa)$ the parameter κ is a measure of kurtosis (Box and Tiao, 1973). This family of distributions generalizes the normal ($\kappa = 0$) and the Laplacian ($\kappa = 1$) distribution. In general we get for $\kappa > 0$ *leptokurtic* and for $\kappa < 0$ *platykurtic* distributions. The choice of a leptokurtic or platykurtic distribution has a strong influence on how a generative factor of the data is represented by a latent dimension. Fig. 3 depicts two possible prior distributions over latents that represent the (x,y) spatial location of a sprite in the dSprites dataset (Matthey et al., 2017). The leptokurtic distribution expects most of the probability mass around 0 and therefore favours a projection of the x and y coordinates, which are distributed in a square, onto the diagonal. The platykurtic prior is closer to a uniform distribution and therefore encourages an axis-aligned representation. This example shows how the choice of the prior will effect the latent representation. Obviously the normal distribution is a special instance of the class of L^p-spherically symmetric distributions, and the normal distribution is the only L^2-spherically symmetric distribution with independent marginals. Equivalently (Sinz et al., 2009a) showed that this also generalizes to arbitrary values of p. The marginals of the p-generalized normal distribution are independent, and it is the only factorial model in the class of L^p-spherically symmetric distributions.

3.2 Independent Subspace Analysis

ICA can be further generalized to include independence between subspaces, but dependencies within them, by

(a) $p = 1.0$ (b) $p = 2.4$

Figure 3: Leptokurtic and platykurtic priors encourage different orientations of the encoding of the (x,y) location of a sprite in the dSprites dataset. A leptokurtic distribution (here the Laplace distribution) has, in two dimensions, contour lines along diagonal directions and expects most of the probability mass around 0. Because the (x,y) locations in dSprites are distributed in a square, the projection of the coordinates onto the diagonal fits better to the Laplace prior. A platykurtic distribution however is more similar to a uniform distribution, with axis aligned contour lines in two dimensions. This fits better to an orthogonal projection of the (x,y) location. The red and blue colour coding denotes the value of the latent variable for the respective (x,y) location of a sprite.

using a more general prior, the family of L^p-nested symmetric distributions (Hyvärinen and Hoyer, 2000; Hyvärinen and Köster, 2007; Sinz et al., 2009b; Sinz and Bethge, 2010).

L^p-nested Function To start, we take a look at a simple example of an L^p-nested function:

$$\left(|z_1|^{p_0} + \left(|z_2|^{p_1} + |z_3|^{p_1}\right)^{\frac{p_0}{p_1}}\right)^{\frac{1}{p_0}}, \tag{6}$$

with $p_0, p_1 \in \mathbb{R}$. This function is a cascade of two L^p-norms. To aid intuition we provide a visualization of this distribution in figure 4a, which depicts (6) as a tree that visualizes the nested structure of the norms. figure 2 visualizes the iso-contours of this function for different values of p_0 and p_1. We call the class of functions which employ this structure *L^p-nested*.

L^p-nested Distribution Given an L^p-nested function f and a radial density $\psi_0 : \mathbb{R} \mapsto \mathbb{R}^+$ we define the *L^p-nested symmetric distribution* following Fernandez et al. (1995) as

$$p_{\mathrm{ISA}}(\boldsymbol{z}) = \frac{\psi_0(f(\boldsymbol{z}))}{f(\boldsymbol{z})^{n-1}\mathcal{S}_f(1)}, \tag{7}$$

where $\mathcal{S}_f(1)$ is the surface area of the L^p-nested unit-sphere. This surface area can be obtained by using the gamma function:

$$\mathcal{S}_f(R) = R^{n-1}2^n \prod_{i \in I} \frac{\prod_{k=1}^{l_1} \Gamma\left[\frac{n_{i,k}}{p_i}\right]}{p_i^{l_i - 1}\Gamma\left[\frac{n_i}{p_i}\right]}, \tag{8}$$

where l_i is the number of children of a node i, n_i is the number of leaves in a subtree under the node i, and $n_{i,k}$ is the number of leaves in the subtree of the k-th children of node i. For further details we refer the reader to the excellent work of Sinz and Bethge (2010).

Independent Subspace Analysis The family of L^p-nested distributions allows a generalization of ICA called independent subspace analysis (ISA). ISA uses a subclass of L^p-nested distributions, which are defined by functions of the form

$$f(\boldsymbol{z}) = \left(\left(\sum_{j=1}^{n_1} |z_j|^{p_1}\right)^{\frac{p_0}{p_1}} + \ldots \right.$$
$$\left. \ldots + \left(\sum_{j=n_1+\cdots+n_{l-1}+1}^{n} |z_j|^{p_l}\right)^{\frac{p_0}{p_l}}\right)^{\frac{1}{p_0}}, \tag{9}$$

and correspond to trees of depth two. The tree structure of this subclass of functions is visualized in figure 4b where each v_i, $i = 1, \ldots, l_0$ denotes the function value of the L^p-norm evaluated over a node's children. The components z_j of z that contribute to each v_i form a subspace

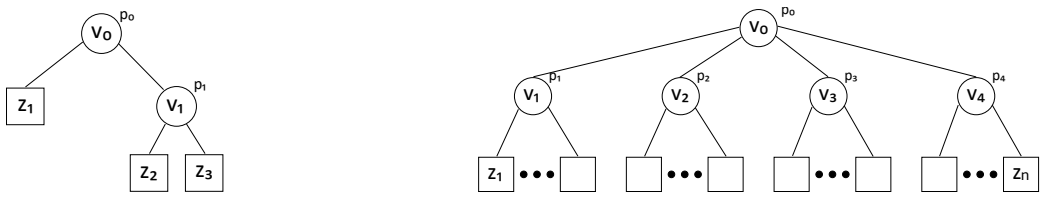
$$\mathcal{V}_i = \left\{ z_j \,\middle|\, j = a \ldots b \text{ with } a = \sum_{k=1}^{i-1} n_k + 1,\ b = a + n_i \right\}. \tag{10}$$

The subspaces $\mathcal{V}_1, \ldots, \mathcal{V}_{l_0}$ become independent when using the radial distribution (Sinz and Bethge, 2010)

$$\psi_0(v_0) = \frac{p_0 v_0^{n-1}}{\Gamma\left[\frac{n}{p_0}\right] s^{\frac{n}{p_0}}} \exp\left(-\frac{v_0^{p_0}}{s}\right). \tag{11}$$

We can interpret this as a generalization of the Chi-distribution: it is the radial distribution of an L^p-nested distribution that becomes equivalent to the Chi-distribution in the case of an L^2-spherically symmetric (Gaussian) distribution.

ISA-VAE We propose to choose the latent prior $p_{\mathrm{ISA}}(\boldsymbol{z})$ (Eq. 7) with $f(\boldsymbol{z})$ from the family of ISA



(a) Tree corresponding to Eq. 6

(b) Tree visualization of Eq. 9, an L^p-nested ISA model.

Figure 4: Tree representation of L^p-nested distributions. a) Tree of the example provided in Eq. 6. b) Tree corresponding to an L^p-nested ISA model.

models of the form of Eq. 9, which allows us to define independent subspaces in the latent space.[2] The Kulback-Leibler divergence of the ELBO-objective can be estimated by Monte-Carlo sampling. This leads to an ELBO-objective of the form

$$\begin{aligned}\mathcal{L}_{\text{ISA-VAE}} =& \mathbb{E}_{z\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}) - \right.\\ & \left. -\beta\left(\log q_\phi\left(\boldsymbol{z}|\boldsymbol{x}\right) - \log p_{\text{ISA}}\left(\boldsymbol{z}\right)\right)\right], \quad (12)\end{aligned}$$

which only requires to compute the log-density of the prior that is readily accessible from the density defined in Eq. 7. As discussed in Roeder et al. (2017) this form of the ELBO even has potential advantages (variance reduction) in comparison to a closed form KL-divergence.

Sampling and the Reparameterization Trick If we want to sample from the generative model we have to be able to sample from the prior distribution. Sinz and Bethge (2010) describe an exact sampling approach to sample from an L^p-nested distribution, which we reproduce as Algorithm 1 in the appendix. Note that during training we only have to sample from the approximate posterior q_ϕ, which we do not have to modify and which can remain a multivariate Gaussian distribution following the original VAE approach. As a consequence, the reparameterization trick can be applied (Kingma and Welling, 2014).

Experiments in the following section demonstrate that the proposed prior supports unsupervised learning of disentangled representation even for the unmodified ELBO objective ($\beta = 1$).

4 Experiments

In our experiments, we evaluate the influence of the proposed prior distribution on disentanglement and on the quality of the reconstruction on the dSprites dataset (Matthey et al., 2017), which contains images of three different shapes undergoing transformations of their position (32 different positions in x and 32 different positions in y), scale (6) and rotation (40), and on the dataset 3D Faces (Paysan et al., 2009) that was also used for evaluation in Chen et al. (2016), which consists of synthetic images of faces with the latent factors azimut (21), elevation (11) and lighting (11). Further we present results on the cars3d dataset (Reed et al., 2015). We follow the same procedure as in Locatello et al. (2019) and perform an extensive evaluation with 50 experiments for each parameter setting, resulting in a total number of 1280 experiments.[3]

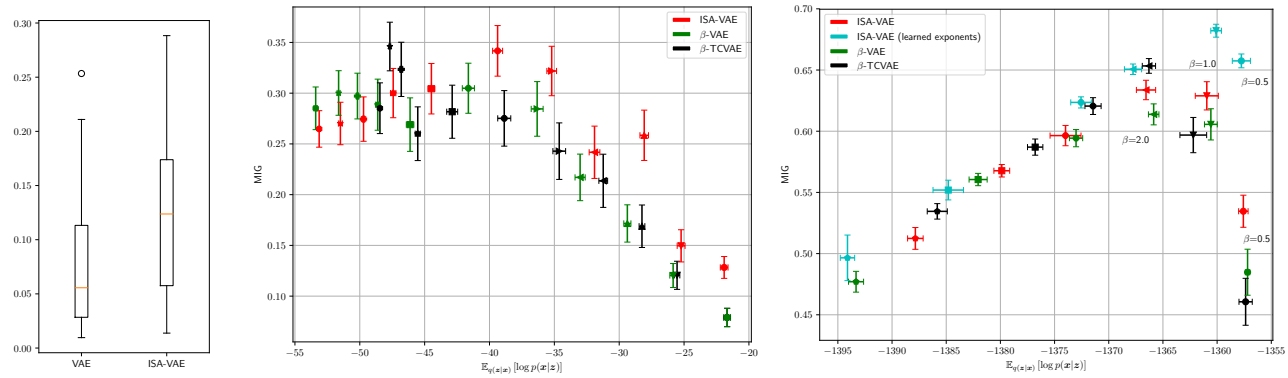
Disentanglement Metrics To provide a quantitative evaluation of disentanglement we compute the disentanglement metric *Mutual Information Gap* (MIG) that was proposed in Chen et al. (2018). The MIG score measures how much mutual information a latent dimension shares with the underlying factor, and how well this latent dimension is separated from the other latent factors. Therefore the MIG measures the two desired properties usually referred to with the term *disentanglement*: a factorized latent representation, and interpretability of the latent factors. Chen et al. (2018) compare the MIG metric to existing disentanglement metrics (Higgins et al., 2016; Kim and Mnih, 2018) and demonstrate that the MIG is more effective and that other metrics do not allow to capture both properties in a desirable way.

Reconstruction Quality To quantify the reconstruction quality, we report the expected log-likelihood of the reconstructed data $\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]$. In our opinion this measure is more informative than the ELBO, frequently reported in existing work, e.g. Chen et al. (2018), especially when varying the β parameter, the weighting of the KL term, which is part of the ELBO and therefore affects its value.

Comparison Baselines Chen et al. (2018) demonstrate that β-TCVAE, a modification of the β-VAE, enables learning of representations with higher disen-

[2] Independend of this work, Higgins et al. (2018) recently proposed to use independent vector subspaces as latent representations to define a new notion of disentanglement.

[3] Source code to repoduce our experiments is available at `https://github.com/microsoft/isa-vae`



(a) dSprites, $\beta = 1.0$ (b) dSprites, $\beta = 1.0, 1.5, \ldots, 6.0$ (c) 3D faces, $\beta = 0.5, 1.0, 2.0, 3.0, 4.0, 5.0$

Figure 5: Comparison of the different approaches for different values of β. (a) The proposed prior facilitates disentanglement, as demonstrated when comparing it to the standard normal prior when using the unmodified ELBO ($\beta = 1.0$). (b), (c) Scatter plots of MIG-score (higher score is better) and reconstruction quality (larger values to the right are better)/ Results for lower β and better reconstruction quality are on the right. Error bars denote the standard error. The proposed approach ISA-VAE allows a better trade-off between disentanglement and reconstruction quality, often outperforming β-VAE and β-TCVAE with respect to MIG score and reconstruction loss on both datasets. (b) On dSprites the baselines reach better MIG scores only for models with poor reconstruction quality. (c) Learning the exponents allows to improve the trade-off between disentanglement and reconstruction loss even further, clearly outperforming the baselines for the unmodefied objective $\beta = 1.0$ and for $\beta = 0.5$. Layout of the ISA model: $l_0 = 5$, $l_{1,\ldots,5} = 4$.

tanglement score than β-VAE (Higgins et al., 2016), InfoGAN (Chen et al., 2016), and FactorVAE (Kim and Mnih, 2018). Therefore we choose β-TCVAE as a baseline for comparison and also compare against β-VAE (Higgins et al., 2016) which for $\beta = 1$ includes the standard VAE with normal prior. To allow a quantitiative comparison with exisiting work we evaluate on the datasets dSprites Matthey et al. (2017) and 3D Faces (Paysan et al., 2009) that were already used in Chen et al. (2016).

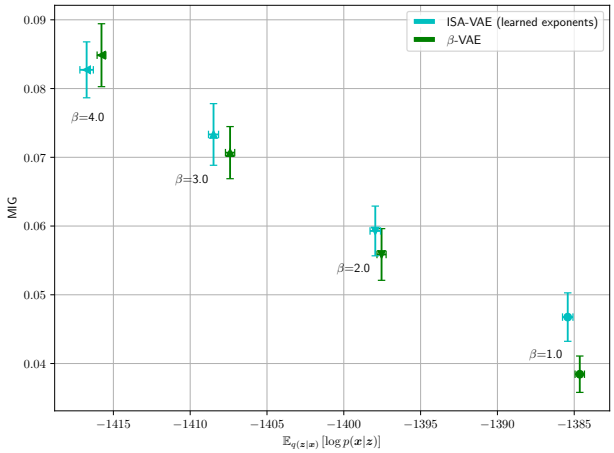
Architecture of the Encoder and Decoder To allow a quantitative comparison with existing work and reproducible results we use the same architecture for the decoder and encoder as presented in Chen et al. (2018). We reproduce the description of the encoder and decoder in appendix A.4

Choosing the ISA-layout In our experience the layout only needs to provide sufficiently many independent vector spaces for learning the representations. If more than the required latent dimensions are provided, unused latent dimensions are usually pruned away.

Choosing the Exponents As we previously discussed in section 3.1 it is important if the prior is leptokurtic or platykurtic. For p_0 we chose the value $p_0 = 2.1$, which results in a platykurtic prior for the distribution over the subspaces, which leads to a rotationally invariant prior. For simplicity, we choose the same exponent p_1 for all the subspaces. For dSprites, a platykurtic distribution fits best to the desired orientation of the x- and y-coordinate (Compare to Fig. 3). To allow a factorized distribution the exponent has to be different to p_0, thus we chose $p_1 = 2.2$. On the 3D faces dataset a leptokyrtic distribution with $p_1 = 1.9$ provided better results than a platykurtic distribution.

Learning the Exponents Instead of choosing a fixed set of exponents, the exponents can also be learned during training. We use the modified ELBO as objective function and optimize encoder, decoder, and the exponents of the prior at the same time. We keep $p_0 = 2.1$ fixed and, beginning from $p_{1,\ldots,k} = 2.0$ optimize the exponent of each subspace individually. Interestingly, optimizing the exponents during training allows to improve the trade-off between disentanglement and reconstruction loss even further. We report results on learning the exponents on the 3D faces dataset in Fig. 5c (15 experiments per beta value) and on the cars3d dataset (Reed et al., 2015) in Fig. 6 (35 experiments per beta value). Histograms of the learnt exponents are shown in appendix A.2.

Hyperparameters To allow reproducibility and a comparison of our results we chose the same hyperparameters as in Chen et al. (2018) and Locatello et al. (2019). We present a table with the evaluated hyperparameters in appendix A.3.



(a) cars3d, $\beta = 1.0, 2.0, 3.0, 4.0$

Figure 6: Comparison on the cars3d dataset. Scatter plot of MIG-score (higher score is better) and reconstruction quality (larger values to the right are better). Error bars denote the standard error. For the unmodified ELBO (beta=1, rightmost data points), ISA-VAE with learned exponents outperforms β-VAE with respect to the MIG score. These results confirm that the prior facilitates disentangled representations. This advantage however decreases for larger beta values.

4.1 Support of the Prior to Learn Disentangled Representations

First, we investigate the ability of the prior to support unsupervised learning of disentangled representations for the unmodified ELBO-objective ($\beta = 1$) and compare the distribution of MIG scores that can be reached with ISA-VAE and the standard VAE in Fig. 5a. We perform $n = 50$ experiments each. We observe a higher mean, median and maximum quantile of disentanglement scores for ISA-VAE which indicates that the prior facilitates to learn interpretable representations even when using the unmodified ELBO objective with $\beta = 1$.

4.2 Trade-off between Disentanglement and Reconstruction Loss

Since the proposed prior facilitates learning of disentangled representations, not only a higher disentanglement score can be reached, but also higher scores are reached for smaller values of β, when compared to the original approaches. This leads to a clear improvement of the trade-off between disentanglement and reconstruction loss. The improvement of this trade-off is demonstrated in Fig. 5, where we plot both the disentanglement score and the reconstruction loss for varying values of β. ISA-β-VAE reaches high values of the disentanglement score for smaller values of β which at the same time preserves a higher quality of the reconstruction than the respective original approaches.

The results on both datasets show that the increase of the MIG score for the baseline method β-TCVAE comes at the cost of a lower reconstruction quality. This difference in the reconstruction quality becomes visible in the quality of the reconstructed images, especially for the more complex heart shape. Please refer to the appendix where we present latent traversals in appendix A.5. With the proposed approach ISA-VAE the reconstruction quality can be increased while at the same time providing a higher disentanglement. This trade-off can be even further improved when learning the exponents $p_{1,\dots,k}$ of the prior during training. Results for ISA-VAE with learned exponents on the 3d faces dataset are depicted in Fig. 5c and on the cars 3d dataset in Fig. 6.

When learning the exponents, we observe the highest difference of disentanglement scores on the cars 3d dataset for low values of β. On the 3d faces cataset, the highest MIG scores among all approaches are reached with ISA-VAE with learned exponents for the unmodified ELBO objective ($\beta = 1.0$), outperforming the existing approaches by a large margin. Both results strongly support our hypothesis, that the proposed prior facilitates learning of disentangled representations even for the unmodified ELBO objective.

5 Conclusion

We presented a structured prior for unsupervised learning of disentangled representations in deep generative models. We choose the prior from the family of L^p-*nested symmetric distributions* which enables the definition of a hierarchy of independent subspaces in the latent space. In contrast to the standard normal prior that is often used in training of deep generative models the proposed prior is not rotationally invariant and therefore enhances the interpretability of the latent space. We demonstrate in our experiments, that a combination of the proposed prior with existing approaches for unsupervised learning of disentangled representations allows a significant improvement of the trade-off between disentanglement and reconstruction loss. This trade-off can be improved further by learning the parameters of the prior during training.

References

Achille, A. and Soatto, S. (2017). Emergence of invariance and disentangling in deep representations. *arXiv preprint arXiv:1706.01350*.

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2017). Automatic differentiation in

machine learning: a survey. *Journal of machine learning research*, 18(153):1–153.

Box, G. E. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading,Mass.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2017). Understanding disentangling in β-vae. In *Learning Disentangled Representations: From Perception to Control Workshop*.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. (2018). Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., and Abbeel, P. (2016). Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180.

Cremer, C., Li, X., and Duvenaud, D. (2018). Inference suboptimality in variational autoencoders. *arXiv preprint arXiv:1801.03558*.

Denton, E. L. et al. (2017). Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pages 4414–4423.

Fernandez, C., Osiewalski, J., and Steel, M. F. (1995). Modeling and inference with v-spherical distributions. *Journal of the American Statistical Association*, 90(432):1331–1340.

Goroshin, R., Bruna, J., Tompson, J., Eigen, D., and LeCun, Y. (2015). Unsupervised learning of spatiotemporally coherent metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 4086–4093.

Higgins, I., Amos, D., Pfau, D., Racanière, S., Matthey, L., Rezende, D. J., and Lerchner, A. (2018). Towards a definition of disentangled representations. *CoRR*, abs/1812.02230.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. (2016). β-vae: Learning basic visual concepts with a constrained variational framework.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011a). Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer.

Hinton, G. E., Krizhevsky, A., and Wang, S. D. (2011b). Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer.

Hsu, W.-N., Zhang, Y., and Glass, J. (2017). Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pages 1878–1889.

Hyvärinen, A. and Hoyer, P. (2000). Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural computation*, 12(7):1705–1720.

Hyvärinen, A. and Köster, U. (2007). Complex cell pooling and the statistics of natural images. *Network: Computation in Neural Systems*, 18(2):81–100.

Kim, H. and Mnih, A. (2018). Disentangling by factorising. *arXiv preprint arXiv:1802.05983*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. (2014). Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *International Conference on Learning Representations*.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. (2015). Deep convolutional inverse graphics network. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2539–2547. Curran Associates, Inc.

Kumar, A., Sattigeri, P., and Balakrishnan, A. (2017). Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

Lee, T.-w. and Lewicki, M. (2000). The generalized gaussian mixture model using ica.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature neuroscience*, 5(4):356.

Locatello, F., Bauer, S., Lučić, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. F. (2019). Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. (2017). dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/.

Paysan, P., Knothe, R., Amberg, B., Romdhani, S., and Vetter, T. (2009). A 3d face model for pose and illumination invariant face recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301. Ieee.

Reed, S., Sohn, K., Zhang, Y., and Lee, H. (2014). Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pages 1431–1439.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. (2015). Deep visual analogy-making. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 1252–1260. Curran Associates, Inc.

Ridgeway, K. (2016). A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*.

Roeder, G., Wu, Y., and Duvenaud, D. K. (2017). Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6925–6934. Curran Associates, Inc.

Rolinek, M., Zietlow, D., and Martius, G. (2019). Variational autoencoders pursue pca directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12406–12415.

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

Sinz, F. and Bethge, M. (2010). Lp-nested symmetric distributions. *Journal of Machine Learning Research*, 11(Dec):3409–3451.

Sinz, F., Gerwinn, S., and Bethge, M. (2009a). Characterization of the p-generalized normal distribution. *Journal of Multivariate Analysis*, 100(5):817–820.

Sinz, F. H., Simoncelli, E. P., and Bethge, M. (2009b). Hierarchical modeling of local image features through L^p-nested symmetric distributions. In *Advances in neural information processing systems*, pages 1696–1704.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press.

Tomczak, J. M. and Welling, M. (2017). Vae with a vampprior. *arXiv preprint arXiv:1705.07120*.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, T., and Chiappa, S., editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press.

Vedantam, R., Fischer, I., Huang, J., and Murphy, K. (2017). Generative models of visually grounded imagination. *arXiv preprint arXiv:1705.10762*.

Vikram, S., Hoffman, M. D., and Johnson, M. J. (2018). The loracs prior for vaes: Letting the trees speak for the data. *arXiv preprint arXiv:1810.06891*.

Zhang, L., Cichocki, A., and Amari, S.-i. (2004). Self-adaptive blind source separation based on activation functions adaptation. *IEEE Transactions on Neural Networks*, 15(2):233–244.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint*.