Disentangling Disentanglement in Variational Autoencoders

E. Mathieu*, T. Rainforth*, N. Siddharth*, Y. W. Teh.
University of Oxford, Departments of Statistics & Engineering
{emile.mathieu, rainforth, y.w.teh}@stats.ox.ac.uk nsid@robots.ox.ac.uk

Overview

We develop a generalisation of disentanglement —*decomposition* of the latent representation—characterising it as the fulfilment of:

(a) the data encodings having an appropriate level of overlap
(b) the aggregate encoding of the data conforming to a desired structure, represented through the prior.

Decomposition permits disentanglement, that is, 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 introduce an alternative VAE training objective that allows one to **control the level of latent overlap**, and to **enforce alternate decompositions** than independence through the **prior distribution**.

Deconstructing the β-VAE

Theorem 1. The β-VAE [1] target

$$\mathcal{L}_\beta(\boldsymbol{x}) = \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]}_{\text{reconstruction error}} - \underbrace{\beta\,\mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))}_{\text{per data posterior regularisation}} \quad (1)$$

can be seen as the standard ELBO, $\mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi)$, 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 $f_\beta(\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{z})^\beta / F_\beta$

$$\mathcal{L}_\beta(\boldsymbol{x}) = \underbrace{\mathcal{L}(\boldsymbol{x})\,(\pi_{\theta,\beta}, q_\phi)}_{\text{ELBO with } \beta\text{-annealed prior}} + \underbrace{(\beta-1)H_{q_\phi}}_{\text{maxent}} + \underbrace{\log F_\beta}_{\text{constant}} \quad (2)$$

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

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

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}_\beta(\boldsymbol{x}; \theta^\dagger(R), \phi^\dagger(R)) \quad (3)$$

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

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

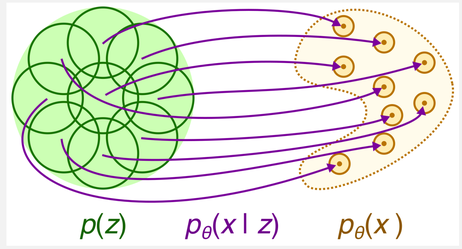
Implications:

- The β-VAE's disentanglement is largely down to control of overlap: it provides **no direct pressure** for independence
- Larger β are **not universally beneficial** for disentanglement
- Dispels the conjecture that the β-VAE objective encourages meaningful independent latent variables when using the standard choice of an isotropic Gaussian prior

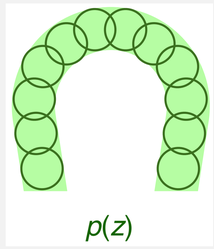
EM, TR, YWT were supported in part by the ERC under the EU's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 617071. EM was also supported by Microsoft Research through its PhD Scholarship Programme. NS was funded by EPSRC/MURI grant EP/N019474/1.

[1] 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 (ICLR)*, 2016.
[2] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *CoRR*, abs/1802.05983, 2018.
[3] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Decomposition allows richer structure than independence to be imposed on the latent encodings. It can be achieved with sufficient latent overlap and an appropriate prior.



(a) Overlap / Regularisation



(b) Desired structure

Proposal: an objective to enforce a particular *decomposition*.

$$\mathcal{L}_{\alpha,\beta}(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})] \quad \text{Reconstruct observations}$$

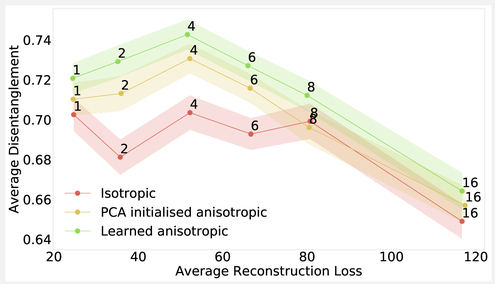
$$- \underbrace{\beta\ \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z}))}_{(a)} \quad \text{Control level of overlap} \quad (4)$$

$$- \underbrace{\alpha\ \mathbb{D}(q_\phi(\boldsymbol{z}), p(\boldsymbol{z}))}_{(b)} \quad \text{Impose desired structure}$$

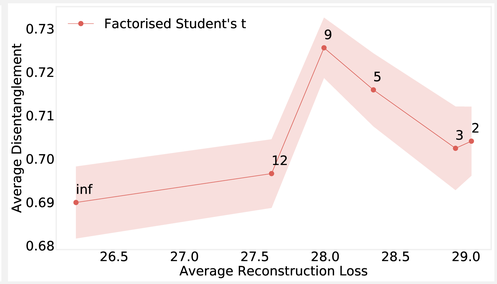
Prior for axis-aligned disentanglement

We demonstrate that substantial improvements in **disentanglement** [2] can be achieved by breaking the rotational invariance of the prior, using either

- A non-isotropic Gaussian (Fig. 2a)
- A product of Student-t's (Fig. 2b)



(a) β-VAE (i.e. (4) with $\alpha = 0$) trained on the *2D Shapes* dataset [3]. Using an anisotropic Gaussian with diagonal covariance either fixed to the principal component values or learned during training.



(b) Using $p_\theta^\nu(\boldsymbol{z}) = \prod_i \text{STUDENT-T}(z_i; \nu)$ for different degrees of freedom ν with $\beta = 1$. Note that $p_\theta^\nu(\boldsymbol{z}) \to \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \mathbf{I})$ as $\nu \to \infty$, and reducing ν only incurs a minor increase in reconstruction loss.

Sparse Prior

We demonstrate important improvements in **sparsity** by using a relaxed *Spike and Slab* prior $p(\boldsymbol{z}) = \prod_d (1-\gamma)\,\mathcal{N}(z_d; 0, 1) + \gamma\,\mathcal{N}(z_d; 0, \sigma_0^2)$, and a dimension-wise MMD regulariser.

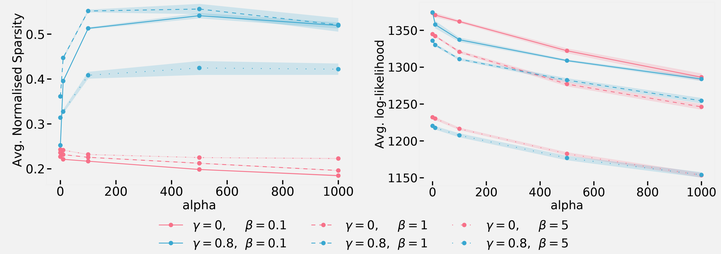


Figure 3: [Left] Sparsity vs regularisation strength α (c.f. Eq. (4), high better). [right] Avg reconstruction log-likelihood vs α (higher better).

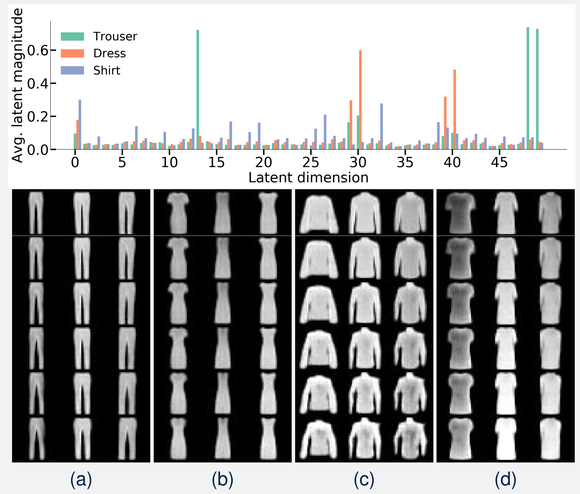


Figure 4: Qualitative evaluation of sparsity.

Clustered prior: Gaussian mixture

As shown in Figure 5, our framework allows to control

- The level of overlap
- The form of the marginal posterior

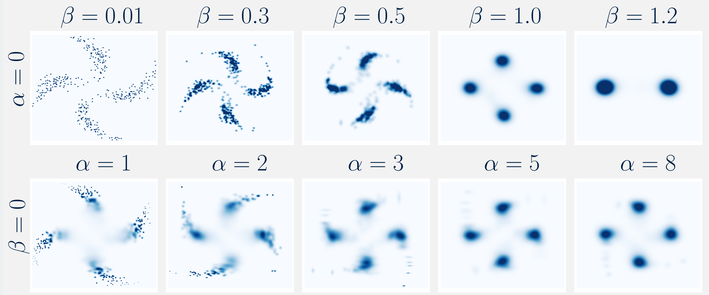


Figure 5: Density of aggregate posterior $q_\phi(\boldsymbol{z})$ for different values of α and β. Models are trained on the "pinwheels" dataset, minimising objective (4).