Improving the ELBO: disentangling part 2

Part I: Challenging common assumptions of disentangling

Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Ratsch, Sylvain Gelly, Bernhard Scholkopf, Olivier Bachem

Disentangled =
if we change one factor of the true underlying representation
then only one factor of latent representation changes

Disentangled =
if we change one factor of the true underlying representation
then only one factor of latent representation changes

Imagine we generate data (faces) according to
face azimuth (a), skin illumination (s), and hair length (h)

The ideal latent representation

$$\begin{pmatrix} a \\ s \\ h \end{pmatrix}$$

Disentangled =
if we change one factor of the true underlying representation
then only one factor of latent representation changes

Disentanglement is sensitive to rotations of the latent embedding

Imagine we generate data (faces) according to
face azimuth (a), skin illumination (s), and hair length (h)

The ideal latent representation

$$\begin{pmatrix} a \\ s \\ h \end{pmatrix}$$

Rotated representation
no longer "disentangled"

$$\begin{pmatrix} 0.75a + 0.25s + 0.61h \\ 0.25a + 0.75s - 0.61h \\ -0.61a + 0.61s + 0.50h \end{pmatrix}$$

But do we have to worry about rotated latent representations?
Yes! Because the VAE objective is exactly the same under rotations of the latent space.
In other words, our optimization doesn't preferentially select for a "good" rotation

The idealized VAE objective

$$\sum_{i=1}^{N} \log p(\mathbf{x}^i)$$

Let U be a fixed rotation matrix. The rotated latent space induces a joint distribution

$$p_U(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x} \mid U^\top \mathbf{z})$$

For every $\mathbf{x}^i \in X$ *we have* $p(\mathbf{x}^i) = p_U(\mathbf{x}^i)$

Proof. We simply compute

$$\begin{aligned}
p_U(\mathbf{x}^i) &= \int p_U(\mathbf{x}^i, \mathbf{z})\, \mathrm{d}\mathbf{z} \\
&= \int p(\mathbf{z})p(\mathbf{x}^i \mid U^\top \mathbf{z})\, \mathrm{d}\mathbf{z} \\
&= \int p(U\mathbf{z})p(\mathbf{x}^i \mid \mathbf{z})\, \mathrm{d}\mathbf{z} && \text{change of variables} \\
&= \int p(\mathbf{z})p(\mathbf{x}^i \mid \mathbf{z})\, \mathrm{d}\mathbf{z} = p(\mathbf{x}^i) && \text{rotational symmetry of prior p(z)}
\end{aligned}$$

Rolinek, Zietlow et al. 2018 "Variational Autoencoders Pursue PCA Directions (by Accident)"
Locatello et al. 2019 "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

But do we have to worry about rotated latent representations?
Yes! Because the VAE objective is exactly the same under rotations of the latent space.
In other words, our optimization doesn't preferentially select for a "good" rotation

The ELBO approximation

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

$$p_U(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x} \mid U^\top \mathbf{z})$$
$$q_U(\mathbf{x}, \mathbf{z}) = q(\mathbf{x})q(U^\top \mathbf{z} \mid \mathbf{x})$$

$$\begin{aligned}
&D_{\mathrm{KL}}(q_U(\mathbf{z} \mid \mathbf{x}^i) \parallel p_U(\mathbf{z} \mid \mathbf{x}^i)) \\
&= \int q_U(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q_U(\mathbf{z} \mid \mathbf{x}^i)}{p_U(\mathbf{z} \mid \mathbf{x}^i)} \, \mathrm{d}\mathbf{z} \\
&= \int q_U(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q_U(\mathbf{z} \mid \mathbf{x}^i) \cdot p_U(\mathbf{x}^i)}{p_U(\mathbf{z}) \cdot p_U(\mathbf{x}^i \mid \mathbf{z})} \, \mathrm{d}\mathbf{z} && \text{Bayes' Rule} \\
&\overset{(3)}{=} \int q(U^\top \mathbf{z} \mid \mathbf{x}^i) \log \frac{q(U^\top \mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{z}) \cdot p(\mathbf{x}^i \mid U^\top \mathbf{z})} \, \mathrm{d}\mathbf{z} \\
&\overset{(4)}{=} \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{U}\mathbf{z}) \cdot p(\mathbf{x}^i \mid \mathbf{z})} \, \mathrm{d}\mathbf{z} && \text{change of variables} \\
&\overset{(5)}{=} \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{z}) \cdot p(\mathbf{x}^i \mid \mathbf{z})} \, \mathrm{d}\mathbf{z} && \text{rotational symmetry of prior p(z)} \\
&= \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i)}{p(\mathbf{z} \mid \mathbf{x}^i)} \, \mathrm{d}\mathbf{z} \\
&= D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}^i) \parallel p(\mathbf{z} \mid \mathbf{x}^i)),
\end{aligned}$$

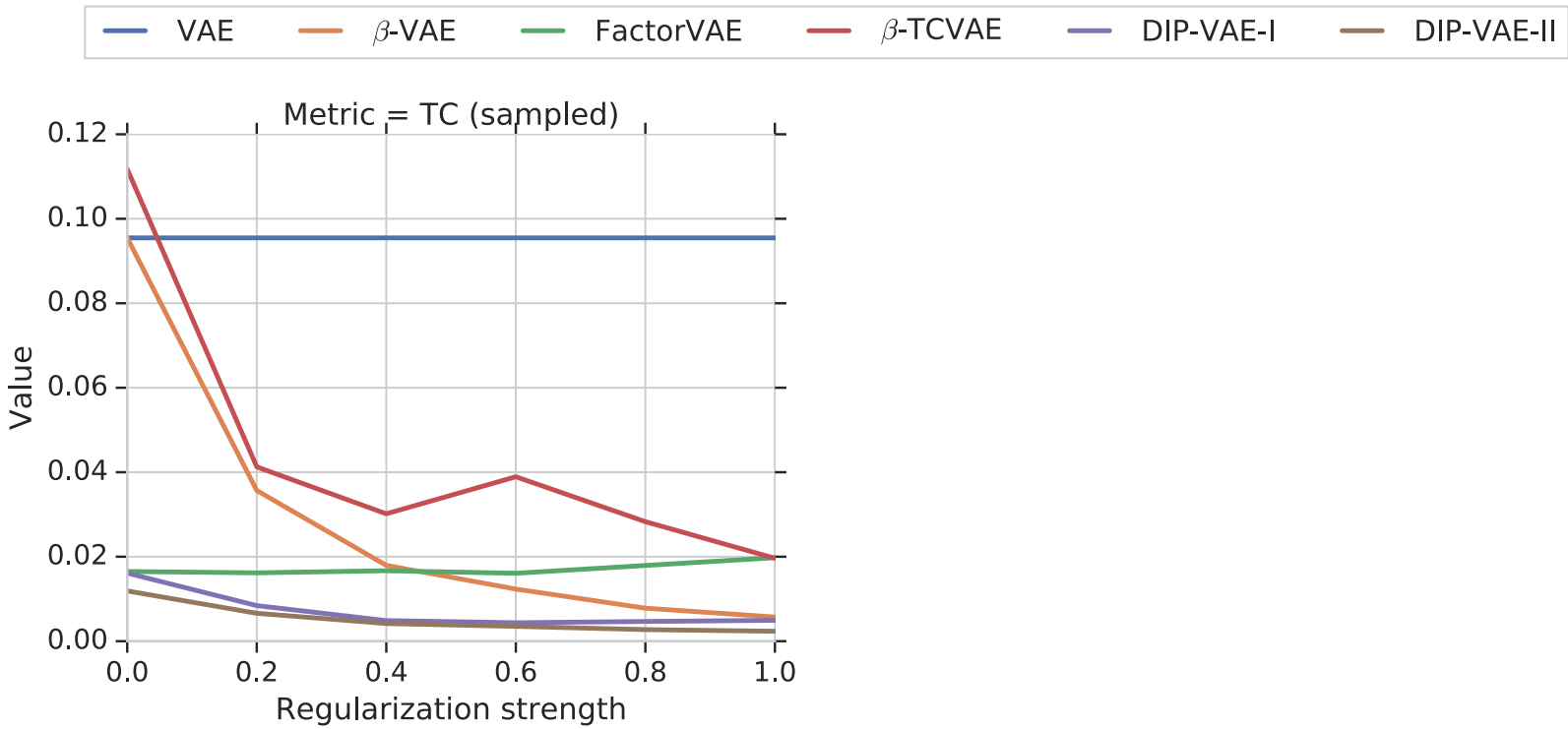
Rolinek, Zietlow et al. 2018 "Variational Autoencoders Pursue PCA Directions (by Accident)"
Locatello et al. 2019 "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Disentangled =
if we change one factor of the true underlying representation
then only one factor of latent representation changes

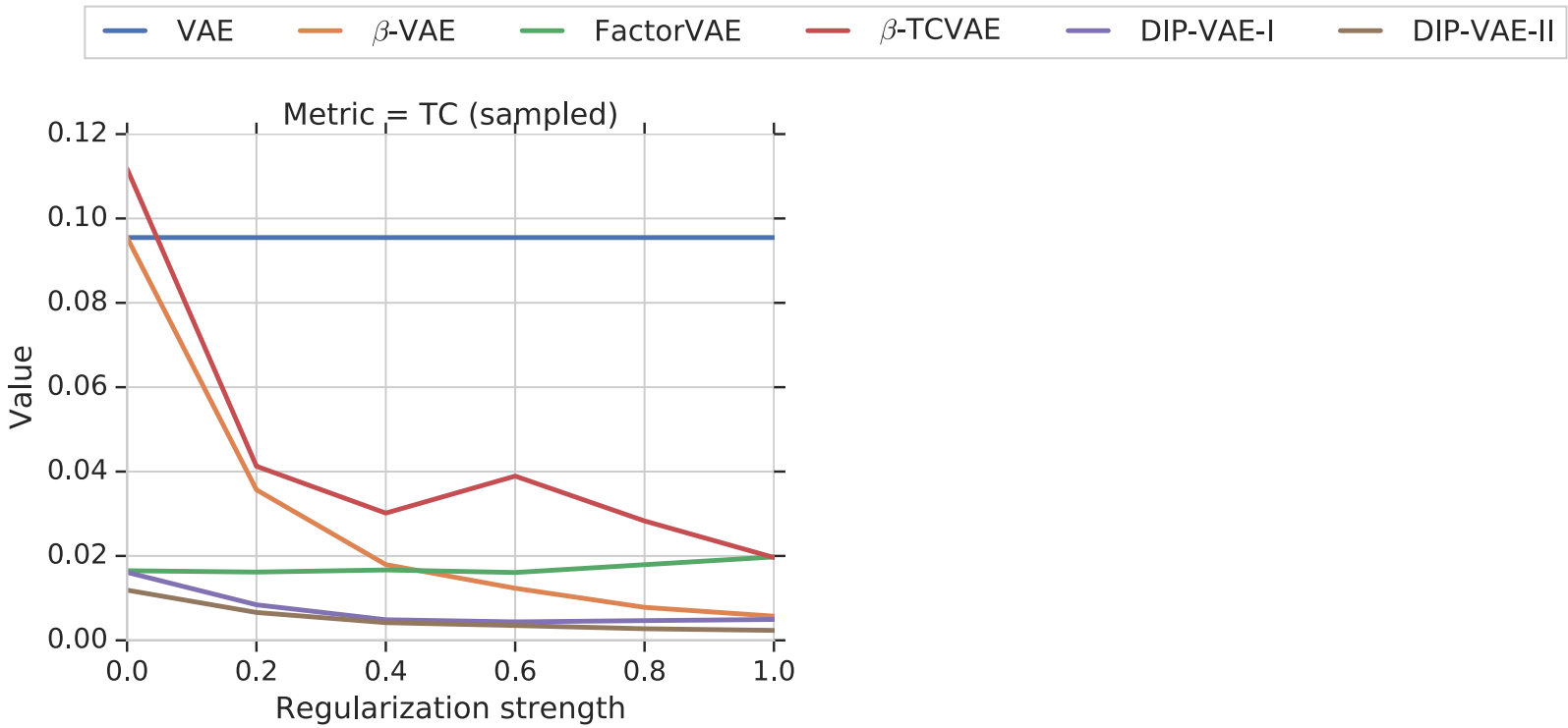
Theorem 1. *For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Then, there exists an infinite family of bijective functions $f : \mathrm{supp}(\mathbf{z}) \to \mathrm{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\boldsymbol{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \boldsymbol{u}) = P(f(\mathbf{z}) \leq \boldsymbol{u})$ for all $\boldsymbol{u} \in \mathrm{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).*

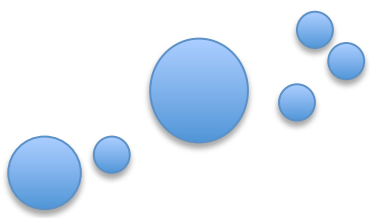
Locatello et al. 2019 “Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Can current methods enforce an uncorrelated representation?

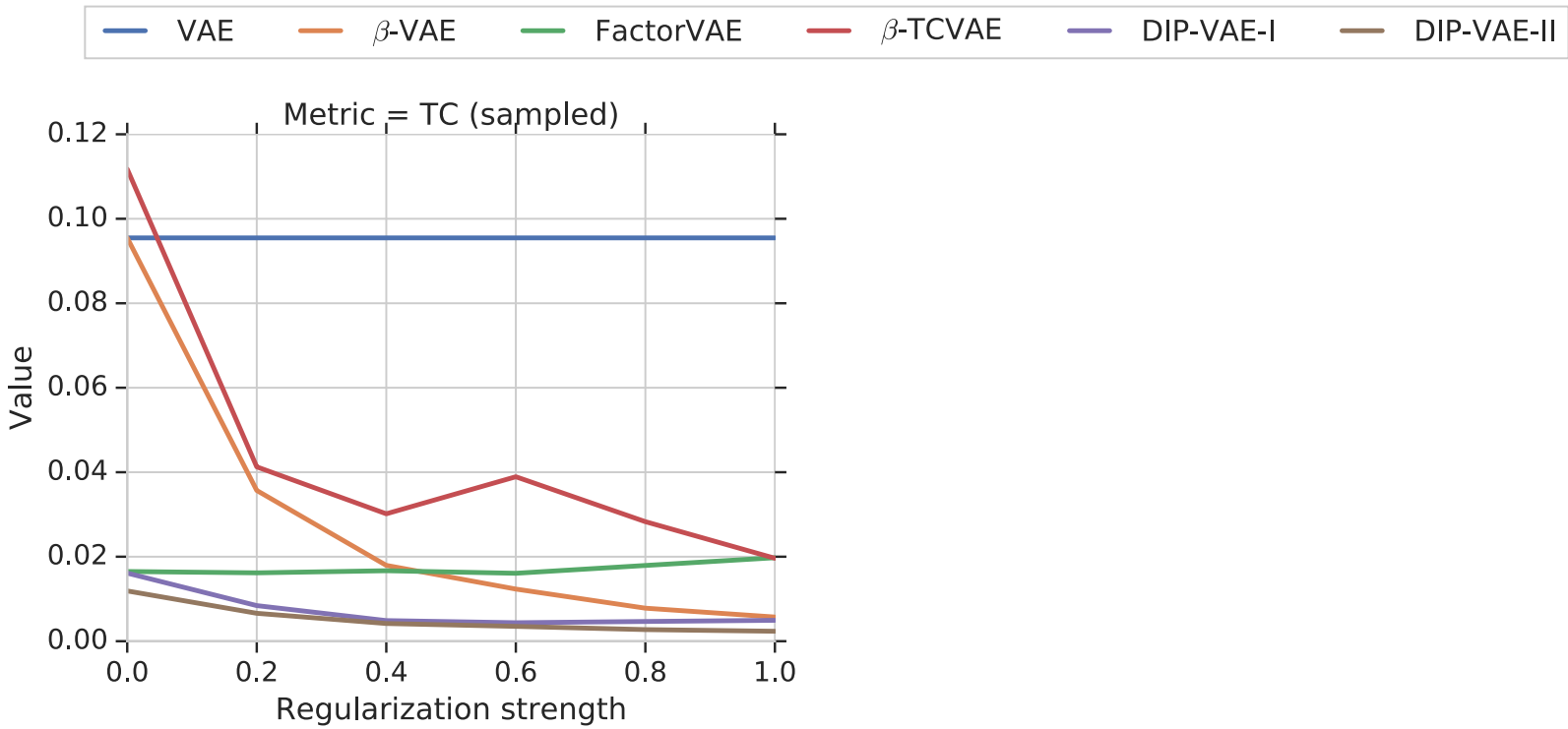


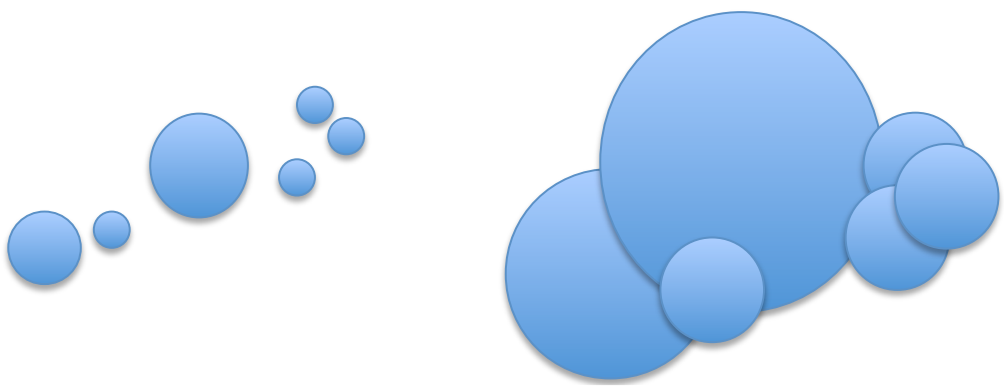
Can current methods enforce an uncorrelated representation?



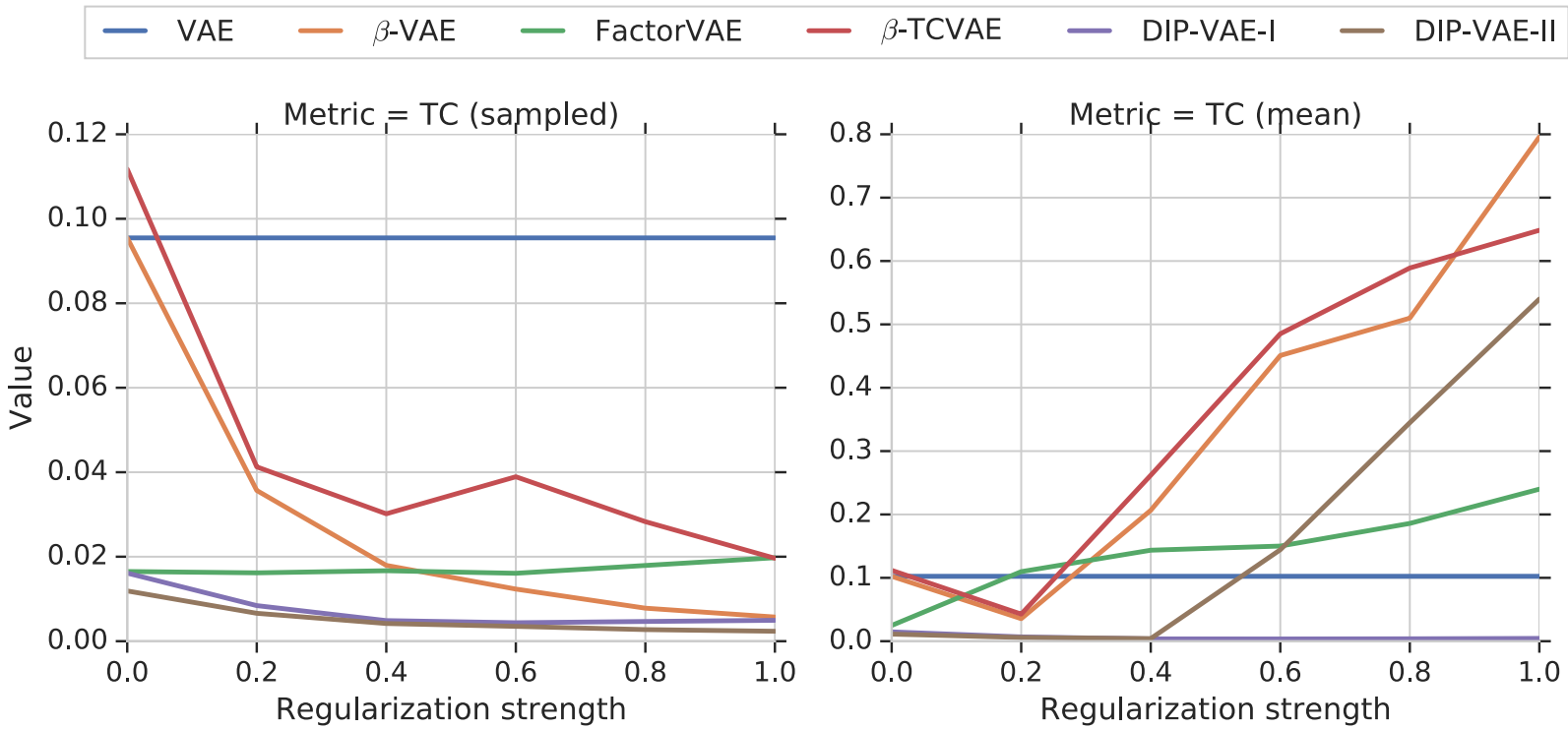


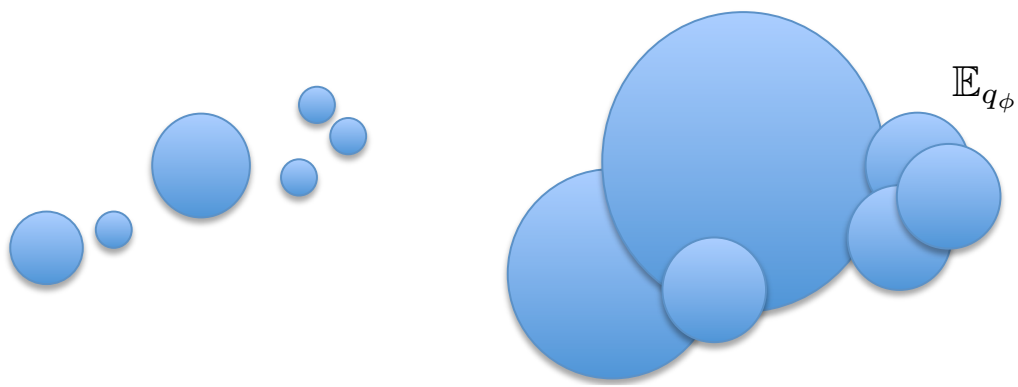
Can current methods enforce an uncorrelated representation?





Can current methods enforce an uncorrelated representation?





$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$$

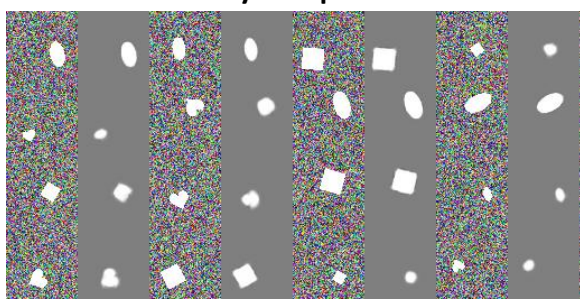
$$p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, I)$$

Do different methods for quantifying disentanglement agree?

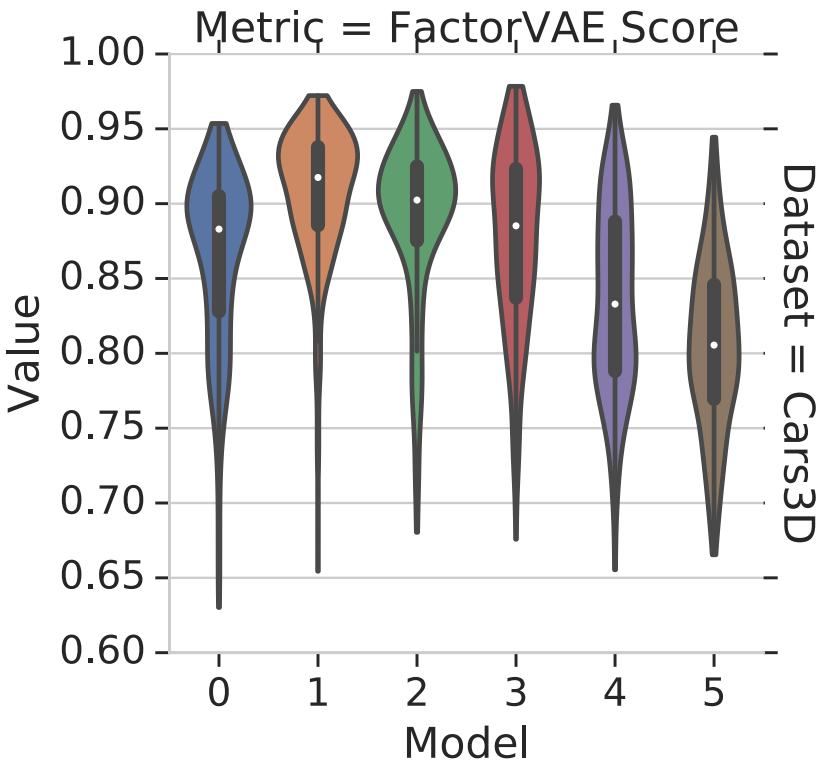
Dataset = Noisy-dSprites

	(A)	(B)	(C)	(D)	(E)	(F)
BetaVAE Score (A)	100	80	44	41	46	37
FactorVAE Score (B)	80	100	49	52	25	38
MIG (C)	44	49	100	76	6	42
DCI Disentanglement (D)	41	52	76	100	-8	38
Modularity (E)	46	25	6	-8	100	13
SAP (F)	37	38	42	38	13	100

Noisy-dSprites



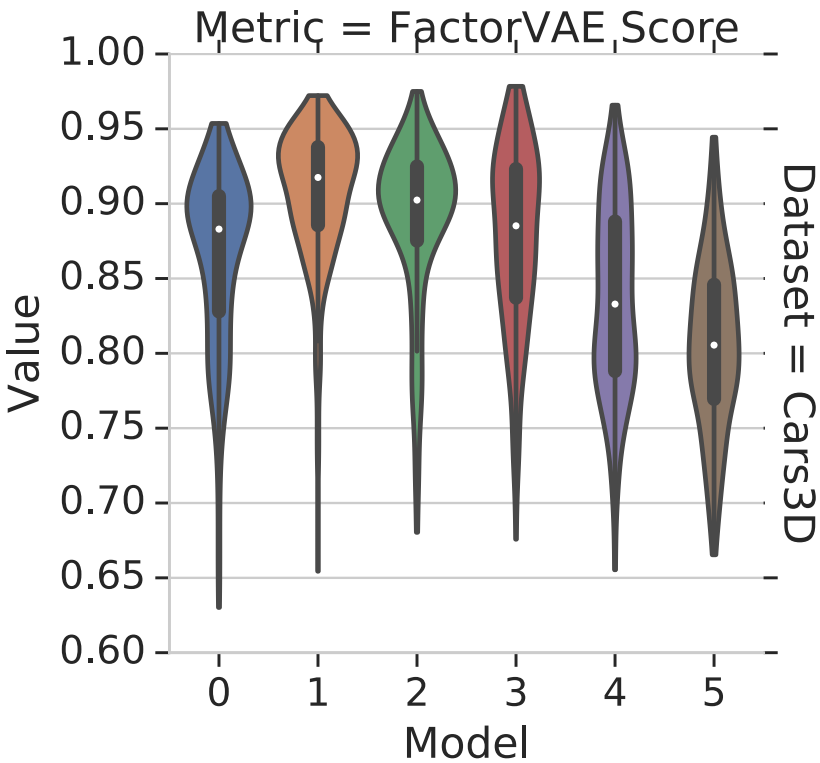
Disentangling depends on the hyperparameter and random seed!

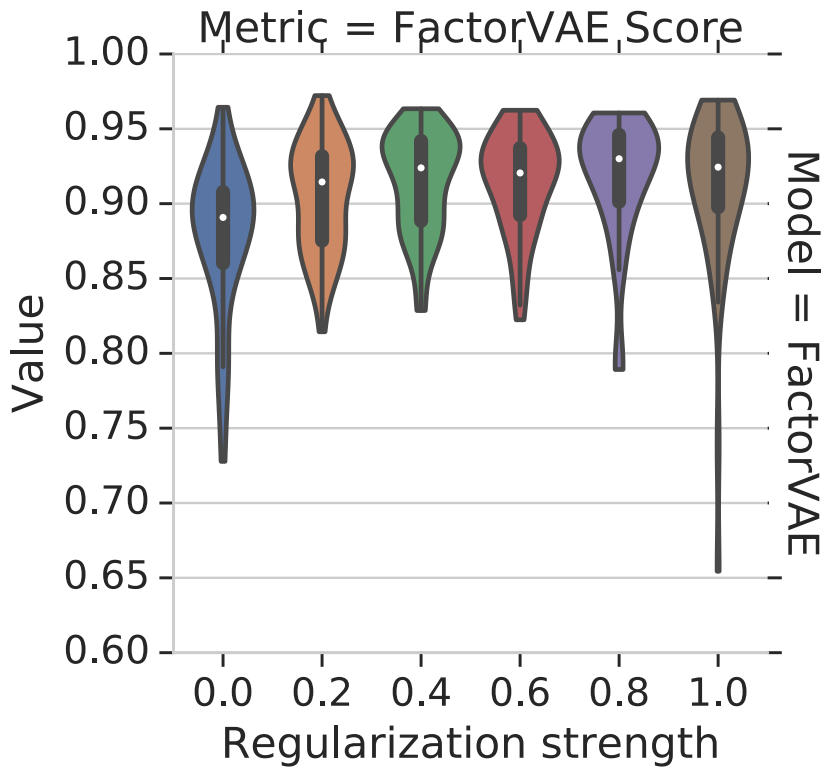


0 = β-VAE, 1 = FactorVAE, 2 = β-TCVAE
3 = DIP-VAE-I, 4 = DIP-VAE-II, 5 = AnnealedVAE

Disentangling depends on the hyperparameter and random seed!

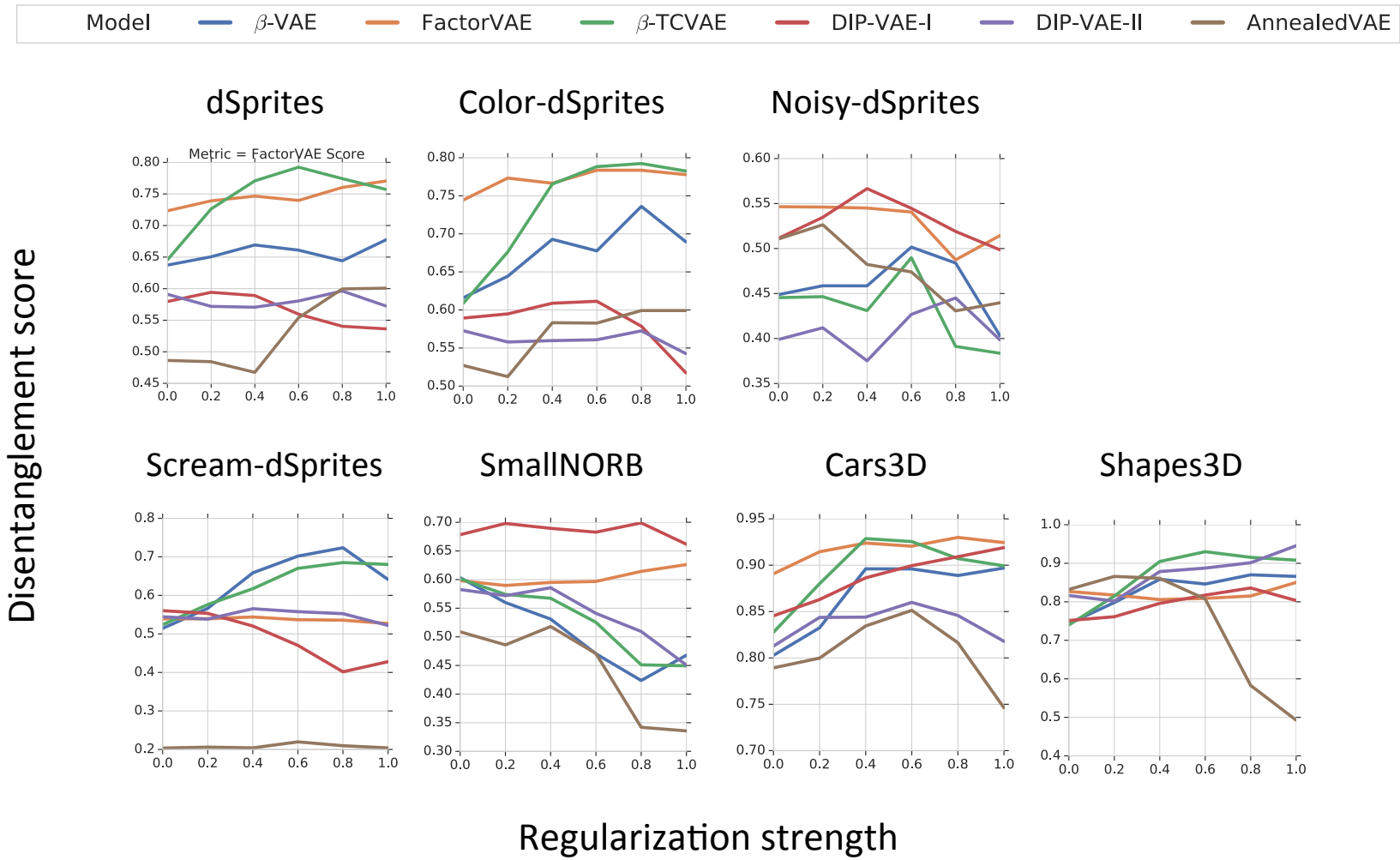
Random seeds have a huge impact!
A good run with a bad hyperparameter can beat a bad run with a good hyperparameter





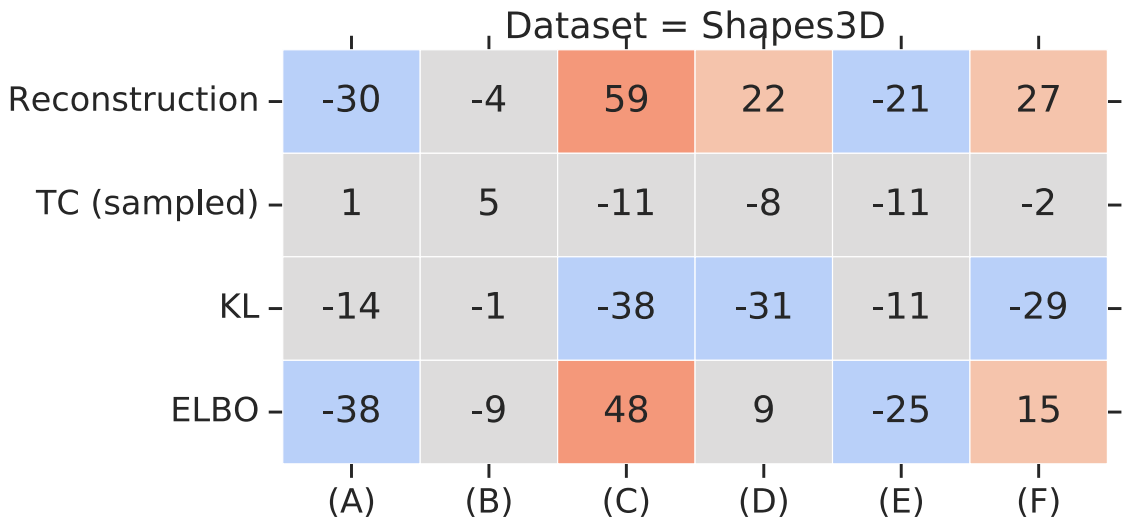
0 = β-VAE, 1 = FactorVAE, 2 = β-TCVAE
3 = DIP-VAE-I, 4 = DIP-VAE-II, 5 = AnnealedVAE

Are there universally good hyperparameters?



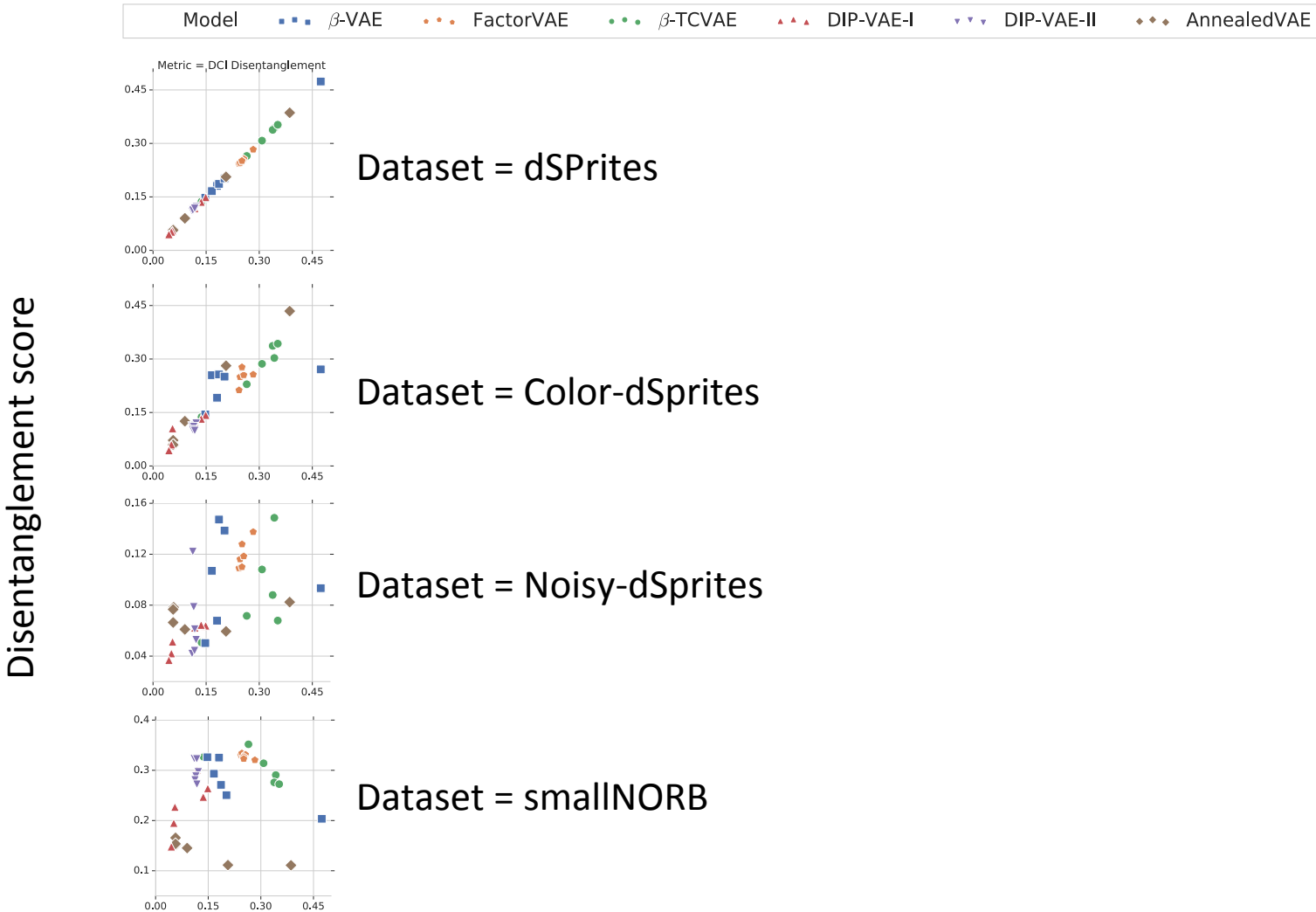
Can we find good hyperparameters using unsupervised metrics?

Rank correlation between model rankings
using unsupervised scores and disentangling scores

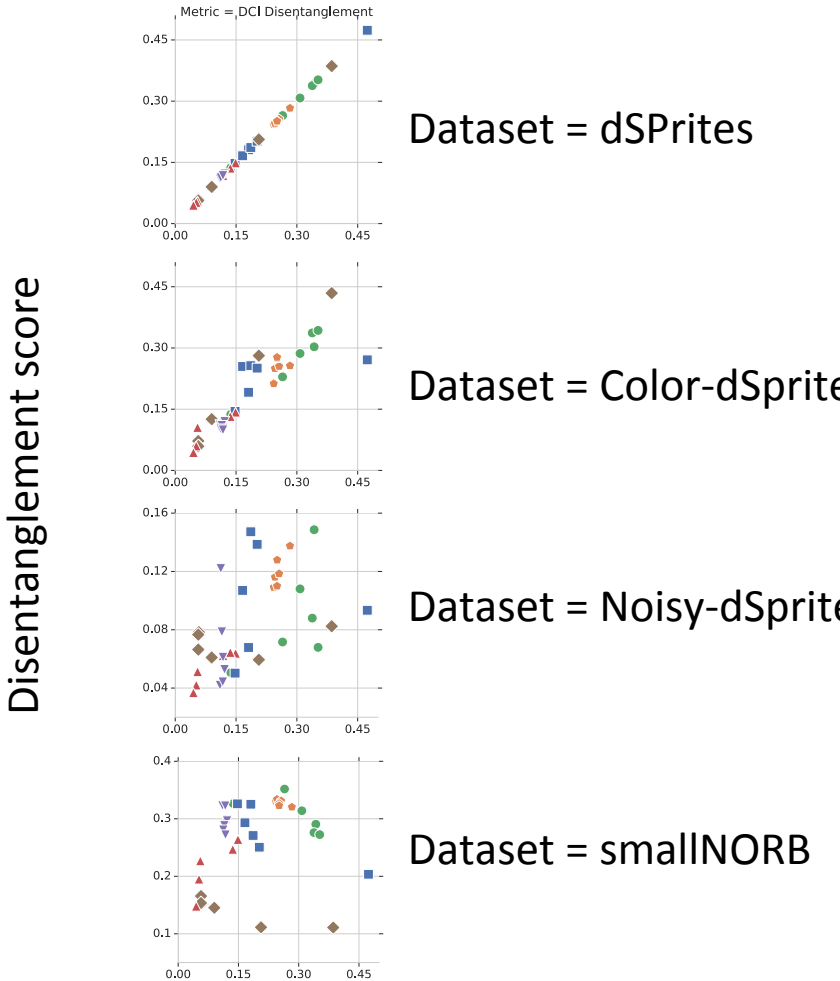


(A)=BetaVAE Score, (B)=FactorVAE Score, (C)=MIG
(D)=DCI Disentanglement, (E)=Modularity, (F)=SAP

Do hyperparameters learned on one task transfer to other tasks?



Do hyperparameters learned on one task transfer to other tasks?

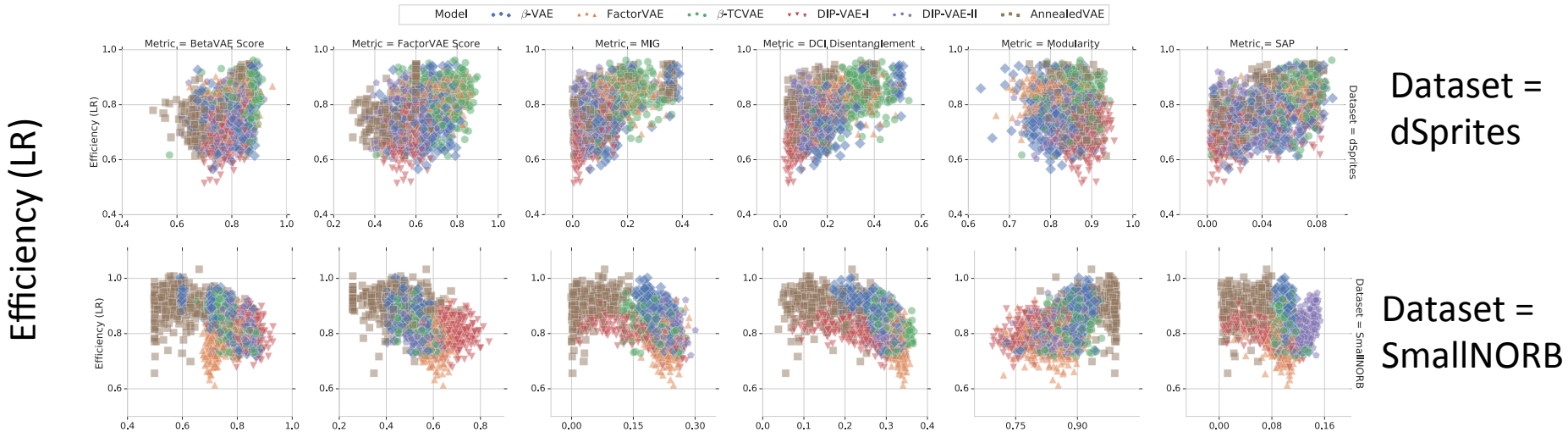


Rank correlation

Metric = DCI Disentanglement

	(I)	(II)	(III)	(IV)	(V)	(VI)	VII
dSprites (I)	100	95	65	65	34	64	46
Color-dSprites (II)	95	100	61	60	21	63	47
Noisy-dSprites (III)	65	61	100	68	17	64	59
Scream-dSprites (IV)	65	60	68	100	36	93	69
SmallNORB (V)	34	21	17	36	100	21	-9
Cars3D (VI)	64	63	64	93	21	100	85
Shapes3D (VII)	46	47	59	69	-9	85	100

Does data efficiency increase as disentangling increases?



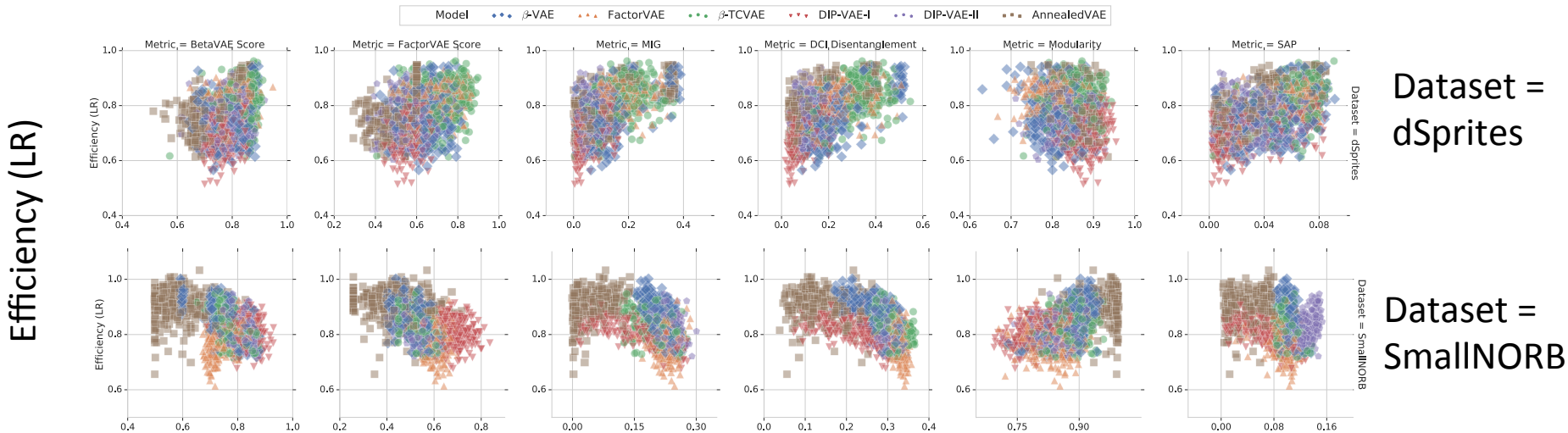
Efficiency (LR)

Dataset = dSprites

Dataset = SmallNORB

Value of disentanglement metric

Does data efficiency increase as disentangling increases?



"While prior work successfully applied disentanglement methods such as β-VAE on a variety of downstream tasks, it is not clear to us that these approaches and trained models performed well *because of disentanglement*."

Locatello et al. 2019 "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations

Part II: Disentangling disentanglement

Disentangling disentanglement

Disentangling disentanglement in variational auto-encoders - Mathieu et al. 2019

- Generalize disentanglement as a decomposition of the latent space into two factors:
  - a.) latent encodings of data having appropriate amount of overlap
  - b.) aggregate encoding of the data conforming to a desired structure.
- β-VAE is simply optimizing the standard VAE ELBO with an exponentially annealed prior and regularized variance of the encoding distribution (degree of overlap).
  - Special case: with gaussian prior and encoding distribution, the β-VAE optimizes the standard VAE ELBO with a $\sqrt{\beta}$-scaled latent space.
- Decomposition enforcing objective allows direct control over a.) degree of overlap and b.) conformation to posterior.

Two factors of decomposition

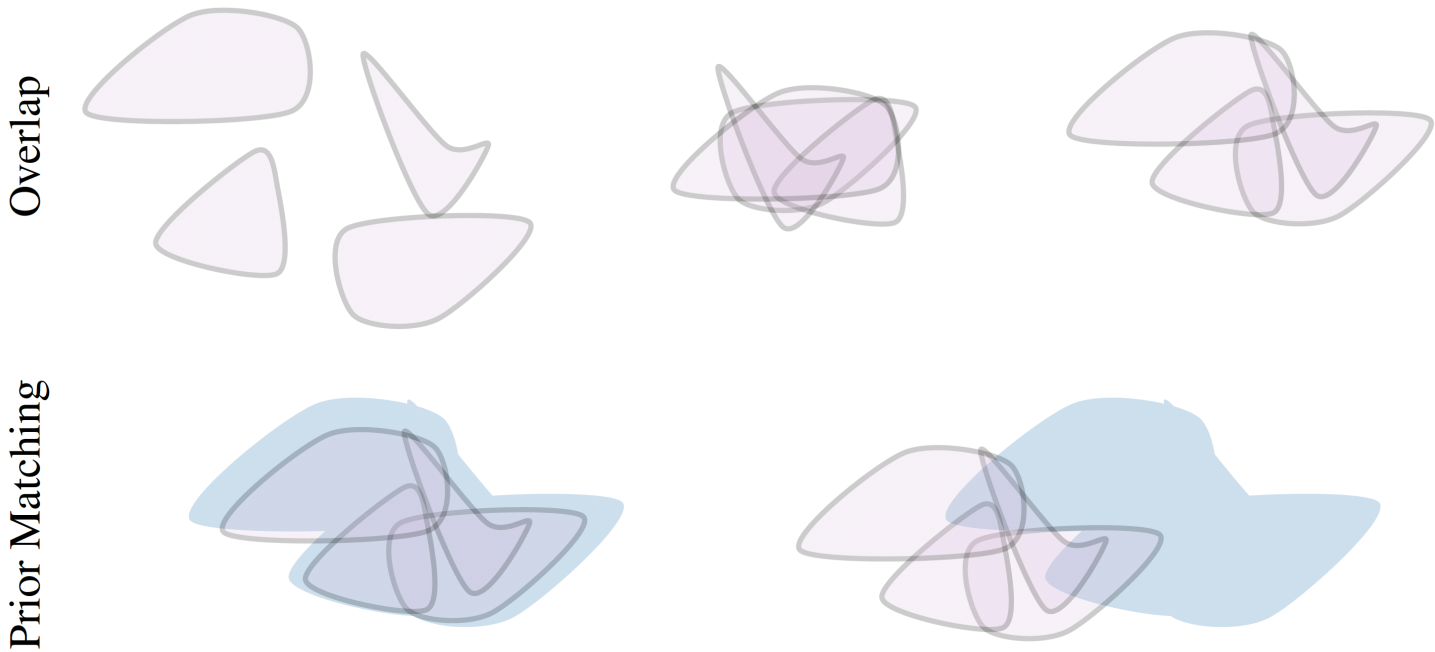


Figure 1. The two factors of decomposition. [Top] Overlap between encodings $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}_i)$, showing cases with (l) too little overlap, (m) too much overlap, and (r) an "appropriate" level of overlap. [Bottom] Illustration of (l) good and (r) bad regularisation between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

Deconstructing the β-VAE

- β-VAE objective

$$\mathcal{L}_\beta(x) = E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \beta KL(q_\phi(z \mid x)||p(z))$$

- Hoffman et al. 2017 show that the β-VAE has an implicit prior of $r(x) = q_\phi(z)^{(1-\beta)}p(z)^\beta$.
- **Theorem 1.** *The β-VAE target $\mathcal{L}_\beta(x)$ an be interpreted in terms of the standard ELBO, $\mathcal{L}(x; \pi_\theta, \beta, q_\phi)$, for an adjusted target $\pi_{\theta,\beta}(x, z) \triangleq p_\theta(x \mid z)f_\beta(z)$ with annealed prior $f_\beta(z) \triangleq p_\theta(z)^\beta / F_\beta$ as*

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

where $F_\beta \triangleq \int_z p_\theta(z)^\beta dz$ is constant given β, and H_{q_ϕ} is the entropy of $q_\phi(z \mid x)$.

- *Proof.*

$$\begin{aligned}
\mathcal{L}_\beta(x) &= E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + \beta H_{q_\phi} + \beta E_{q_\phi(z|x)}\left[\log p_\theta(z)\right] \\
&= E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] + (\beta - 1)H_{q_\phi} + H_{q_\phi} \\
&\quad + E_{q_\phi(z|x)}\left[\log p_\theta(z)^\beta - \log F_\beta\right] + \log F_\beta \\
&= E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - KL(q_\phi(z \mid x)||f_\beta(z)) + (\beta - 1)H_{q_\phi} + \log F_\beta \\
&= \mathcal{L}(x; \pi_\theta, \beta, q_\phi) + (\beta - 1)H_{q_\phi} + \log F_\beta
\end{aligned}$$

Special Case: Gaussian VAEs

- *If* $p_\theta(z) = \mathcal{N}(z; 0, \Sigma)$ *and* $q_\phi(z \mid x) = \mathcal{N}(z; \mu_\phi(x), S_\phi(x))$*, then,*

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

where θ' *and* ϕ' *represent rescaled networks such that*

$$p_{\theta'}(x \mid z) = p_\theta(x \mid z/\sqrt{(\beta)})$$

$$q_{\phi'}(z \mid x) = \mathcal{N}(z; \mu_{\phi'}(x), S_{\phi'}(x))$$

$$\mu_{\phi'}(x) = \sqrt{\beta}\mu_\phi(x)$$

$$S_{\phi'}(x) = \beta S_\phi(x)$$

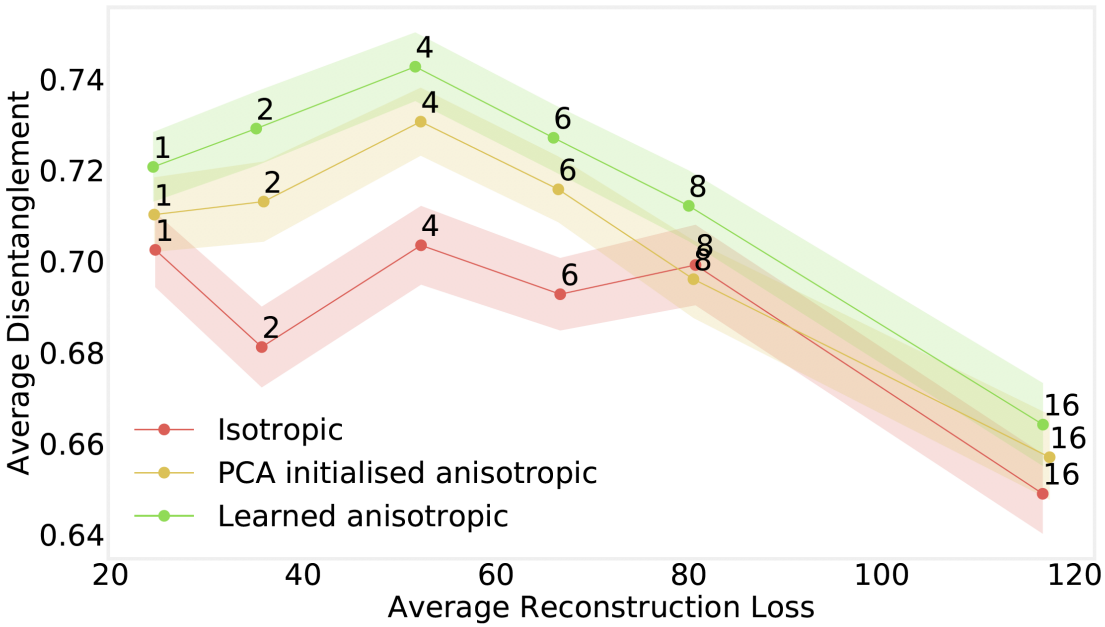
- Noting c's irrelevance to the optimization, we see that β-VAE in the Gaussian case corresponds to optimizing the standard ELBO with a $\sqrt{B}$-scaled latent space with maximum entropy regularization.
  - This is formalized by showing equivalence of stationary points of those two objectives (Corollary 2).

Enforcing VAE Decomposition

- Decomposition of VAE latent space into two factors
  - 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 stochasticity of the encoder.
  - b.) the aggregate encoding $q_\phi(z)$ matching the prior $p_\theta(z)$ where the latter expresses the desired dependency between latents.

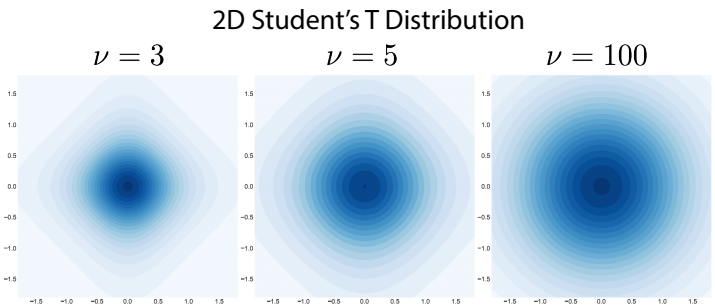
$$L_{\alpha,\beta}(x) = E_{q_\phi(z|x)}\left[\log p_\theta(x \mid z)\right] - \beta KL(q_\phi(z \mid x)||p(z)) - \alpha\mathcal{D}(q_\phi(z), p(z))$$

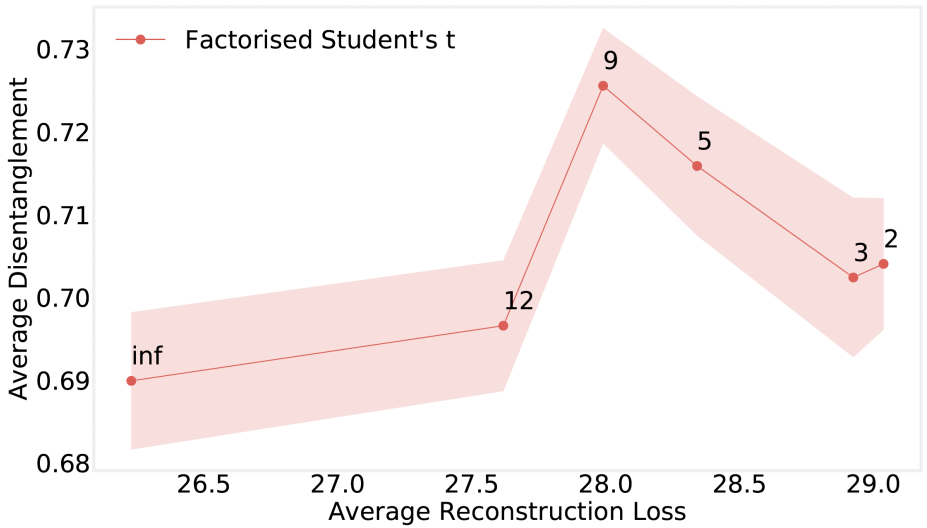
Priors for axis-aligned disentangling: iso- versus aniso-tropy ($\alpha = 0$)



- Anisotropic priors (not rotationally invariant) confer better disentangling, especially if the degree of anisotropy is learned.

Priors for axis-aligned disentangling: Student's. T $\alpha = 0$





- Reducing ν incurs minor reconstruction penalty, while conferring better disentangling, until ν is too low and we see effects of heavy tails.

Tweaking knobs α and β while learning pinwheels with a clustered prior

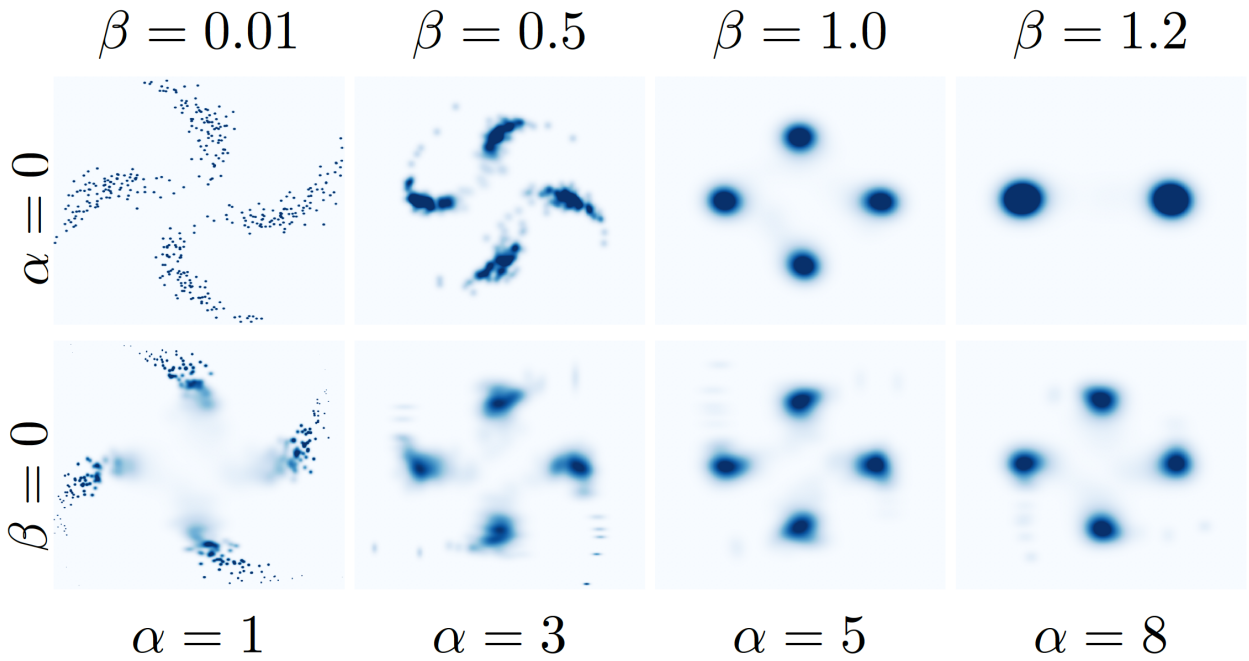


Figure 3. Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for spirals dataset with mixture of Gaussian prior.

- increasing β ($\alpha = 0$) increases overlap through encoder increased encoder variance, and the aggregate posterior does not have to match the prior $p_\theta(z)$ as $\beta \to \infty$.
- increasing α ($\beta = 0$) forces the aggregate posterior to be the prior.

Learning sparse disentangled latents using a sparse prior

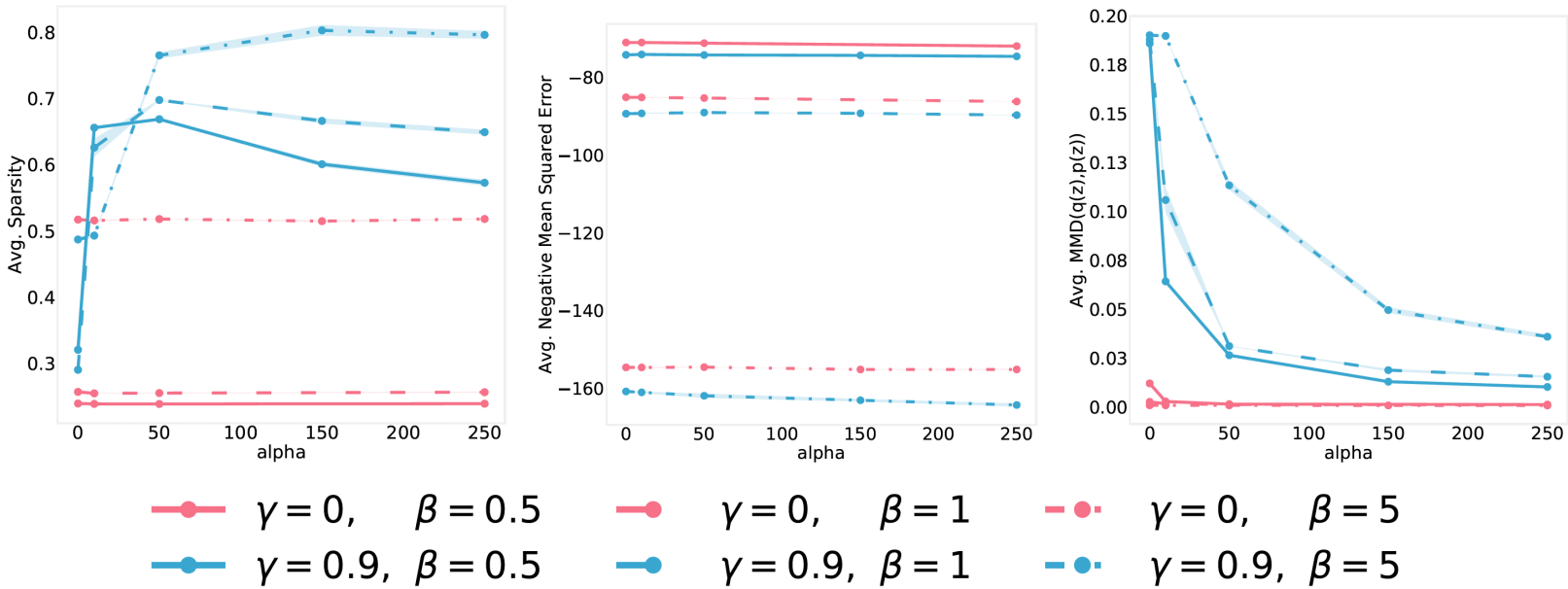


Figure 4. [Left] Sparsity vs regularisation strength α (c.f. (7)). [Center] Reconstruction loss vs α. [Right] Divergence (MMD) vs h α. Note here that the different values of γ represent regularizations to different distributions, with regularization to a Gaussian (i.e. $\gamma = 0$) much easier to achieve than the sparse prior, hence the lower divergence. Shaded areas represent 95% confidence intervals calculated using 3 separately trained networks. See Appendix B for details.

- Sparsity is most effectively induced by changing the prior to be more sparse, rather than increasing β.
- The use of a sparse prior induces far less reconstruction loss than is caused by increasing β.

Summary

- Unsupervised learning of disentangled representations is impossible without using inductive bias.
- We should not expect axis-aligned disentangled latents by enforcing the aggregated posterior to be an isotropic gaussian.
- By decomposing characterization of the latent space into fullfillment of two features
  - degree of overlapping representations
  - conformation of aggregated posterior to the desired structure

  and directly controlling these features through Lagrange multipliers in the learning objective, we can obtain better disentangled representations than more naive approaches (e.g. β-VAE).