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

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

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

$$p_U(x, z) = p(z)p(x | U^\top z)$$

For every $$x^i \in X$$ we have $$p(x^i) = p_U(x^i)$$.

Proof. We simply compute

$$p_U(x^i) = \int p_U(x^i, z) \, dz$$

$$= \int p(z)p(x^i | U^\top z) \, dz$$

$$= \int p(Uz)p(x^i | z) \, dz \quad \text{change of variables}$$

$$= \int p(z)p(x^i | z) \, dz = p(x^i) \quad \text{rotational symmetry of prior } p(z)$$

Rolinek, Zietlow et al. 2018 “Variational Autoencoders Pursue PCA Directions (by Accident)”
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

\[
\mathcal{L}(x; \theta, \phi) \triangleq \log p_\theta(x) - \text{KL}(q_\phi(z|x) \| p_\theta(z|x)) \\
= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \text{KL}(q_\phi(z|x) \| p(z)).
\]

\[
p_U(x, z) = p(z)p(x \mid U^\top z) \\
q_U(x, z) = q(x)q(U^\top z \mid x) \]

\[
P_U(Z \mid x^i) \quad D_{KL}(q_U(Z \mid x^i) \| p_U(Z \mid x^i))
\]

\[
= \int q_U(z \mid x^i) \log \frac{q_U(z \mid x^i)}{p_U(z \mid x^i)} \, dz

\]

Bayes’ Rule

\[
= \int q_U(z \mid x^i) \log \frac{q_U(z \mid x^i) \cdot p_U(x^i)}{p_U(z) \cdot p_U(x^i \mid z)} \, dz

\]

change of variables

\[
= \int q(U^\top z \mid x^i) \log \frac{q(U^\top z \mid x^i) \cdot p(x^i)}{p(z) \cdot p(x^i \mid U^\top z)} \, dz

\]

rotational symmetry of prior \( p(z) \)

\[
= \int q(z \mid x^i) \log \frac{q(z \mid x^i) \cdot p(x^i)}{p(Uz) \cdot p(x^i \mid U^\top z)} \, dz

\]

\[
= \int q(z \mid x^i) \log \frac{q(z \mid x^i) \cdot p(x^i)}{p(z) \cdot p(x^i \mid z)} \, dz

\]

\[
= D_{KL}(q(z \mid x^i) \| p(z \mid x^i)),
\]

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 $p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$. Then, there exists an infinite family of bijective functions $f : \text{supp}(\mathbf{z}) \to \text{supp}(\mathbf{z})$ such that $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all $i$ and $j$ (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled) and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \text{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).

Can current methods enforce an uncorrelated representation?

![Graph showing the total correlation metric against regularization strength for different methods.](image)
Can current methods enforce an uncorrelated representation?

![Graph showing the total correlation (TC) of sampled representations vs. regularization strength for different models.](image)
Can current methods enforce an uncorrelated representation?

![Graph showing the total correlation (TC) metric for different methods as a function of regularization strength. The x-axis represents the regularization strength, ranging from 0.0 to 1.0, and the y-axis shows the value of the TC metric. The graph includes lines for VAE, β-VAE, FactorVAE, β-TCVAE, DIP-VAE-I, and DIP-VAE-II, each with a different color. The results indicate that the total correlation decreases as the regularization strength increases.]
Can current methods enforce an uncorrelated representation?

\[
\mathbb{E}_{q_{\phi}(z|x)} \left[ \log p_\theta(x|z) \right] - \beta \text{KL}(q_{\phi}(z|x)||p(z))
\]

\[
p(z) = \mathcal{N}(0, I)
\]
Do different methods for quantifying disentanglement agree?

<table>
<thead>
<tr>
<th>Dataset = Noisy-dSprites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beta VAE Score (A)</td>
</tr>
<tr>
<td>Factor VAE Score (B)</td>
</tr>
<tr>
<td>MIG (C)</td>
</tr>
<tr>
<td>DCI Disentanglement (D)</td>
</tr>
<tr>
<td>Modularity (E)</td>
</tr>
<tr>
<td>SAP (F)</td>
</tr>
</tbody>
</table>

(A) 100  80  44  41  46  37
(B)  80 100  49  52  25  38
(C)  44  49 100  76   6  42
(D)  41  52  76 100 -8  38
(E)  46  25   6 -8 100  13
(F)  37  38  42  38  13 100
Disentangling depends on the hyperparameter and random seed!

\( \text{Metric} = \text{FactorVAE, Score} \)

0 = \( \beta \)-VAE, 1 = FactorVAE, 2 = \( \beta \)-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.

\[ \text{Metric} = \text{FactorVAE,Score} \]

Dataset = Cars3D

Value

Model

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

<table>
<thead>
<tr>
<th>Dataset = Shapes3D</th>
<th>Reconstruction</th>
<th>TC (sampled)</th>
<th>KL</th>
<th>ELBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) = BetaVAE Score</td>
<td>-30</td>
<td>-4</td>
<td>59</td>
<td>22</td>
</tr>
<tr>
<td>(B) = FactorVAE Score</td>
<td>-21</td>
<td>-8</td>
<td>-11</td>
<td>-11</td>
</tr>
<tr>
<td>(C) = MIG</td>
<td>27</td>
<td>-2</td>
<td>-29</td>
<td>-25</td>
</tr>
<tr>
<td>(D) = DCI Disentanglement</td>
<td>15</td>
<td>9</td>
<td>48</td>
<td>15</td>
</tr>
<tr>
<td>(E) = Modularity</td>
<td>-38</td>
<td>-9</td>
<td>-11</td>
<td>-29</td>
</tr>
<tr>
<td>(F) = SAP</td>
<td>-38</td>
<td>-9</td>
<td>-31</td>
<td>-31</td>
</tr>
</tbody>
</table>

(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?

<table>
<thead>
<tr>
<th>Model</th>
<th>β-VAE</th>
<th>FactorVAE</th>
<th>β-TCVAE</th>
<th>DIP-VAE-I</th>
<th>DIP-VAE-II</th>
<th>AnnealedVAE</th>
</tr>
</thead>
</table>

**Dataset = dSPrites**

**Dataset = Color-dSPrites**

**Dataset = Noisy-dSPrites**

**Dataset = smallNORB**

**Dataset = dSPrites**

Disentanglement score
Do hyperparameters learned on one task transfer to other tasks?

![Disentanglement scores on dSprites vs other data sets. Good hyperparameters only seem to transfer consistently from dSprites to Color-dSprites.](image)

<table>
<thead>
<tr>
<th>Model</th>
<th>β-VAE</th>
<th>FactorVAE</th>
<th>β-TVAE</th>
<th>DIP-VAE-I</th>
<th>DIP-VAE-II</th>
<th>AnnealedVAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>dSprites (I)</td>
<td>100</td>
<td>95</td>
<td>65</td>
<td>65</td>
<td>34</td>
<td>64</td>
</tr>
<tr>
<td>Color-dSprites (II)</td>
<td>95</td>
<td>100</td>
<td>61</td>
<td>60</td>
<td>21</td>
<td>63</td>
</tr>
<tr>
<td>Noisy-dSprites (III)</td>
<td>65</td>
<td>61</td>
<td>100</td>
<td>68</td>
<td>17</td>
<td>64</td>
</tr>
<tr>
<td>Scream-dSprites (IV)</td>
<td>65</td>
<td>60</td>
<td>68</td>
<td>100</td>
<td>36</td>
<td>93</td>
</tr>
<tr>
<td>SmallNORB (V)</td>
<td>34</td>
<td>21</td>
<td>17</td>
<td>36</td>
<td>100</td>
<td>21</td>
</tr>
<tr>
<td>Cars3D (VI)</td>
<td>64</td>
<td>63</td>
<td>64</td>
<td>93</td>
<td>21</td>
<td>100</td>
</tr>
<tr>
<td>Shapes3D (VII)</td>
<td>46</td>
<td>47</td>
<td>59</td>
<td>69</td>
<td>-9</td>
<td>85</td>
</tr>
</tbody>
</table>

Rank correlation

Dataset = dSprites

Dataset = Color-dSprites

Dataset = Noisy-dSprites

Dataset = smallNORB
Does data efficiency increase as disentangling increases?

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."

Part II: 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.

- $\beta$-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 $\beta$-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 $q_\phi(z | x_i)$, 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 $q_\phi(z)$ and the desired prior $p(z)$. 
Deconstructing the $\beta$-VAE

- **$\beta$-VAE objective**

$$L_\beta(x) = E_{q_\phi(z|x)}[\log p_\theta(x | z)] - \beta KL(q_\phi(z | x) || p(z))$$

- Hoffman et al. 2017 show that the $\beta$-VAE has an implicit prior of $r(x) = q_\phi(z)^{(1-\beta)} p(z)^\beta$.

- **Theorem 1.** The $\beta$-VAE target $L_\beta(x)$ an be interpreted in terms of the standard ELBO, $L(x; \pi_\theta, \beta, q_\phi)$, for an adjusted target $\pi_\theta, \beta(x, z) \triangleq p_\theta(x | z) f_\beta(z)$ with annealed prior $f_\beta(z) \triangleq p_\theta(z)^\beta / F_\beta$ as

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

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

- **Proof.**

$$L_\beta(x) = E_{q_\phi(z|x)}[\log p_\theta(x | z)] + \beta H_{q_\phi} + \beta E_{q_\phi(z|x)}[\log p_\theta(z)]$$

$$= E_{q_\phi(z|x)}[\log p_\theta(x | z)] + (\beta - 1) H_{q_\phi} + H_{q_\phi}$$

$$+ E_{q_\phi(z|x)} \left[\log p_\theta(z)^\beta - \log F_\beta\right] + \log F_\beta$$

$$= E_{q_\phi(z|x)}[\log p_\theta(x | z)] - KL(q_\phi(z | x) || f_\beta(z)) + (\beta - 1) H_{q_\phi} + \log F_\beta$$

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

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

where \( \theta' \) and \( \phi' \) 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 \( \beta \)-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).
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.

$$L_{\alpha,\beta}(x) = E_{q_\phi(z|x)} \left[ \log p_\theta(x \mid z) \right] - \beta KL(q_\phi(z \mid x) \mid \mid p(z)) - \alpha D(q_\phi(z), p(z))$$
Priors for axis-aligned disentangling: isotropy versus anisotropy ($\alpha = 0$)

- 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 $\nu$ incurs minor reconstruction penalty, while conferring better disentangling, until $\nu$ is too low and we see effects of heavy tails.
Tweaking knobs \( \alpha \) and \( \beta \) while learning pinwheels with a clustered prior

- Increasing \( \beta \) (\( \alpha = 0 \)) 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 \( \alpha \) (\( \beta = 0 \)) forces the aggregate posterior to be the prior.

*Figure 3. Density of aggregate posterior \( q_\phi(z) \) with different \( \alpha, \beta \) for spirals dataset with mixture of Gaussian prior.*
Learning sparse disentangled latents using a sparse prior

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

Figure 4. [Left] Sparsity vs regularisation strength $\alpha$ (c.f. (7)). [Center] Reconstruction loss vs $\alpha$. [Right] Divergence (MMD) vs $\alpha$. Note here that the different values of $\gamma$ 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.
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).