BE MORE ACTIVE! UNDERSTANDING THE DIFFERENCES BETWEEN MEAN AND SAMPLED REPRESENTATIONS OF VARIATIONAL AUTOENCODERS

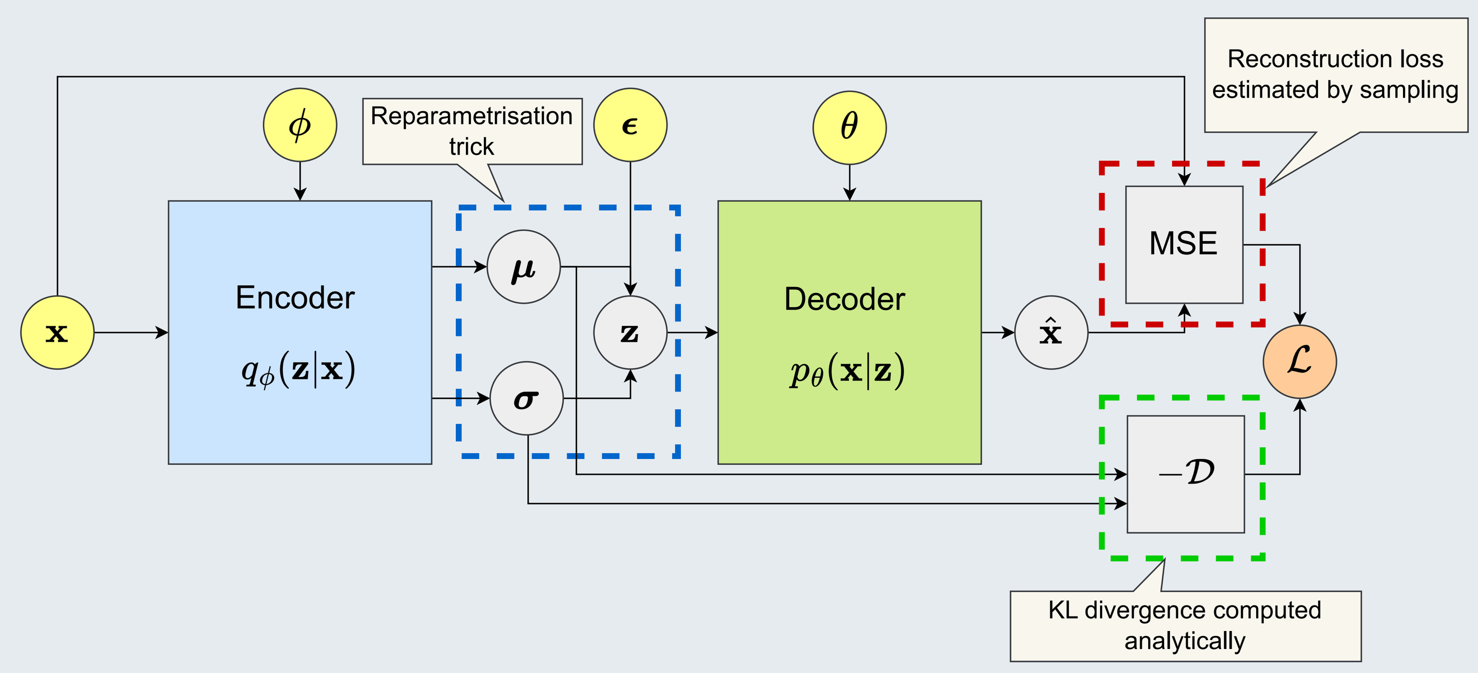
Lisa Bonheme
Marek Grzes

Summary

Locatello et al. [3] observed a lower disentanglement in mean than sampled representations of Variational Autoencoders (VAEs).
In this paper we:

- Analyse the problem through the lens of the polarised regime
- Show that the lower disentanglement of mean representations is due to (uninformative) passive variables
- Provide some recommendation for using mean representations on downstream tasks

What are Variational Autoencoders?

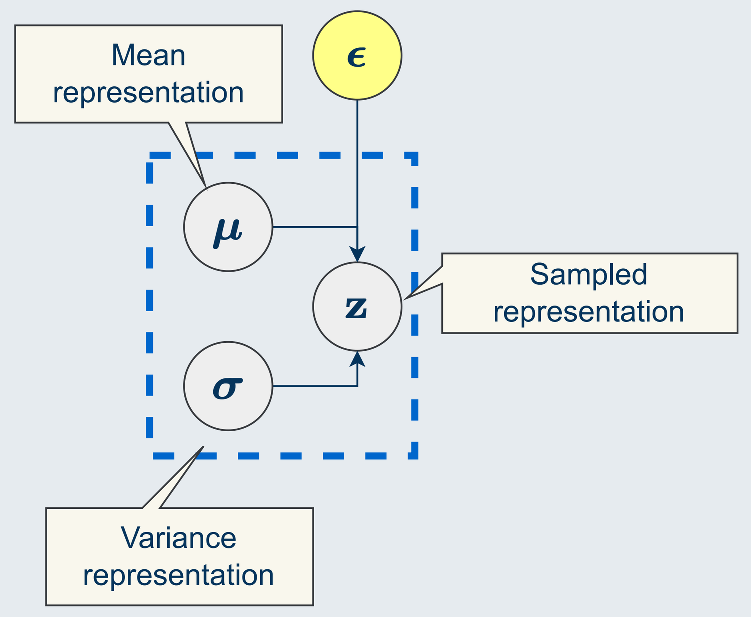


$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\right)}_{\text{regularisation term}}.$$

The samples from the learned latent representation are obtained using the reparameterisation trick [2] such that $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

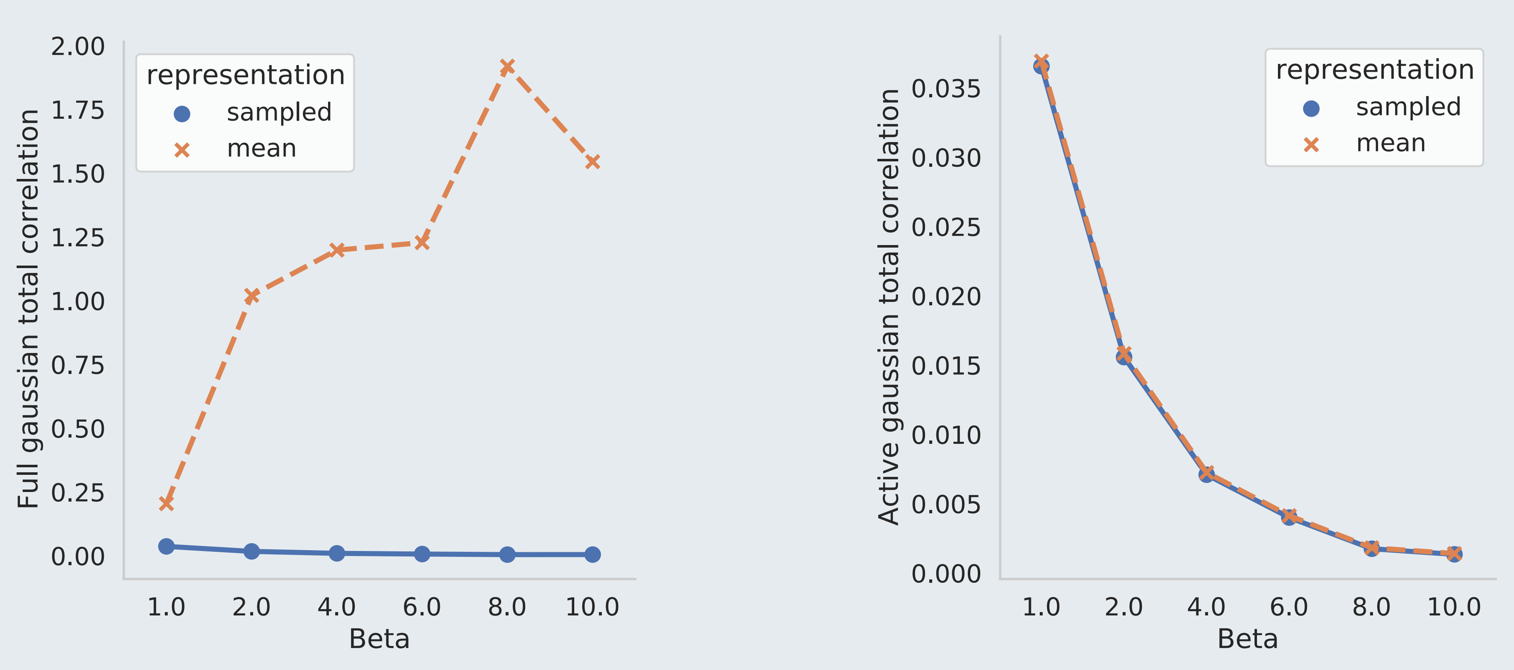
The polarised regime [1, 4]

- **Passive variables**
  $\boldsymbol{\mu}_i \approx 0$, $\boldsymbol{\sigma}_i \approx 1$, and $\mathbf{z}_i \sim \mathcal{N}(0, 1)$.
- **Active variables**
  $\boldsymbol{\sigma}_i \approx 0$ and $\mathbf{z}_i \approx \boldsymbol{\mu}_i$.



Truncation experiment

- Total correlation (TC) and averaged mutual information (MI) are higher in $\boldsymbol{\mu}$ than $\mathbf{z}$ [3].
- For active variables $\mathbf{z}_i \approx \boldsymbol{\mu}_i$.
- Discrepancies between $\boldsymbol{\mu}$ and $\mathbf{z}$ must come from passive variables.

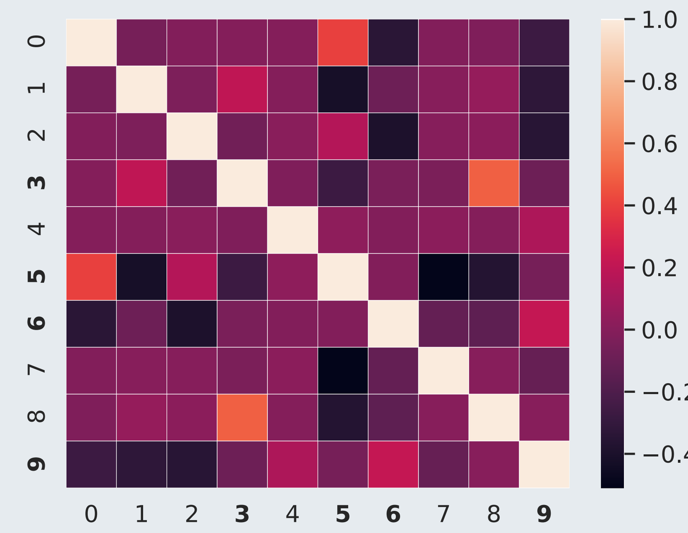


(a) Full representation (b) Only active variables

Figure: TC of β-TC VAE on noisy dSprites

Passive variables correlation

- Passive variables are more correlated in $\boldsymbol{\mu}$ than $\mathbf{z}$.
- They should be removed before using $\boldsymbol{\mu}$ on downstream tasks.



Where does this correlation come from?

- Are variables passive because they are correlated?
- Does the correlation occur because the variables are passive?

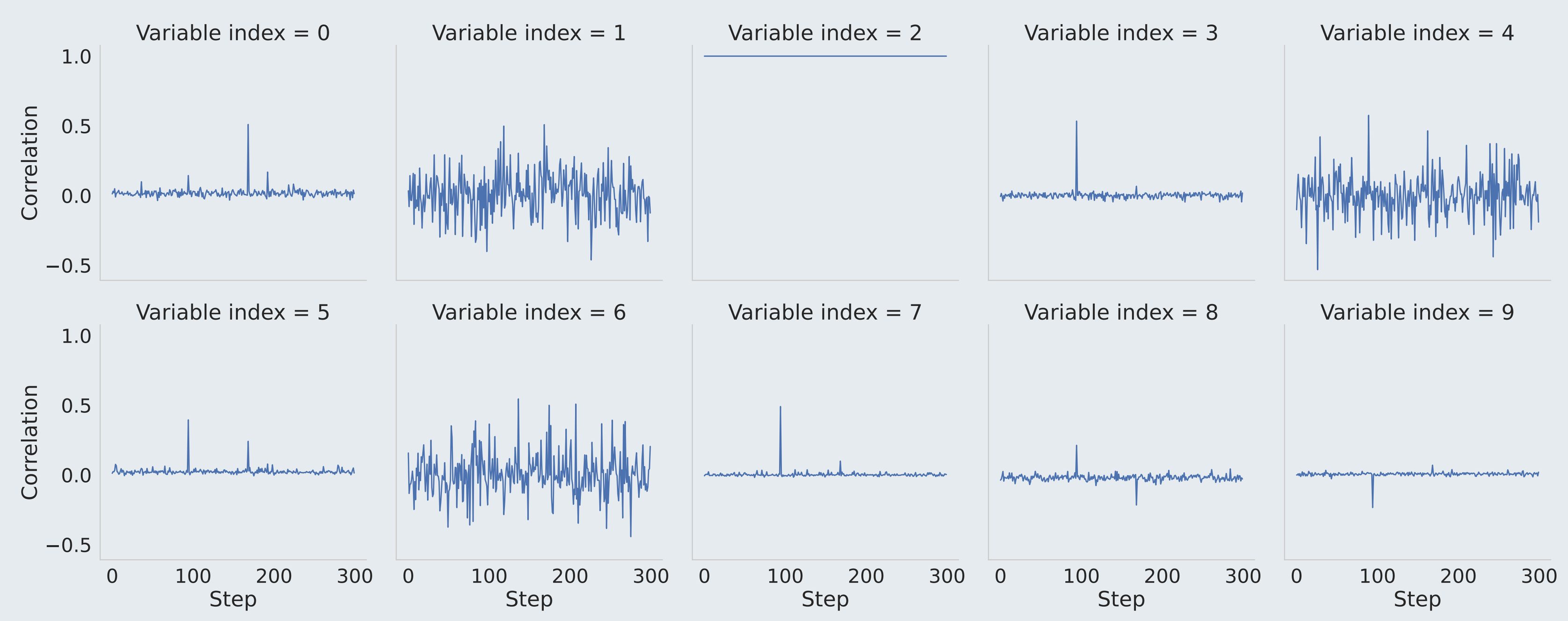


Figure: The correlation scores of the active variable at index 2 of the mean representation with all the other variables during the 300K training steps of a β-VAE with $\beta = 8$ trained on dSprites. We can see an increased correlation with all the passive variables (indexes 1, 4, and 6).

Conclusion

- Active variables are as disentangled in mean as in sampled representations
- Passive variables are highly correlated with various active variables
- An in-depth study of the learning dynamics of VAEs would be needed to explain this phenomenon
- Passive variables should be removed from mean representations before downstream tasks

More about the paper

References

[1] B. Dai, Y. Wang, J. Aston, G. Hua, and D. Wipf. Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models. *Journal of Machine Learning Research*, 19 (41):1–42, 2018.

[2] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, volume 2, 2014.

[3] F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019.

[4] M. Rolinek, D. Zietlow, and G. Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.