Disentangling Disentanglement in Variational Autoencoders

ICML 2019

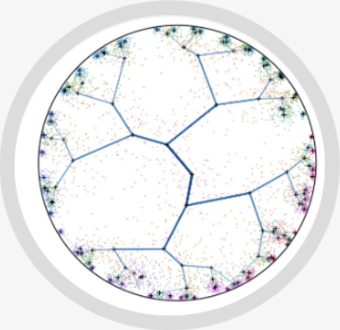
Emile Mathieu*, Tom Rainforth*, N. Siddharth*, Yee Whye Teh

June 12, 2019

Departments of Statistics and Engineering Science, University of Oxford

HIERARCHICAL REPRESENTATIONS WITH POINCARÉ VARIATIONAL AUTO-ENCODERS

On-going work on using hyperbolic geometry for modelling data with underlying hierarchical structure, within the variational auto-encoder setting.



DISENTANGLING DISENTANGLEMENT FOR VARIATIONAL AUTO-ENCODERS

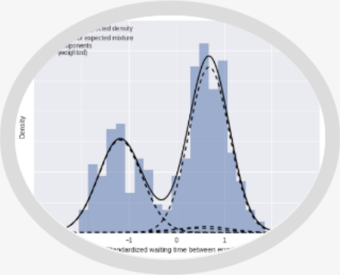
Recent work on *disentanglement* for variational auto-encoders, aiming at generalising such a concept to other than independent representations.

SAMPLING AND INFERENCE FOR BETA NEUTRAL-TO-THE-LEFT MODELS OF SPARSE NETWORKS

09/07/2018

Our paper has been accepted at [UAI 2018](#) for an oral presentation!

Check out the wonderful [slides](#).



NIPS ADVANCES IN APPROXIMATE BAYESIAN INFERENCE WORKSHOP

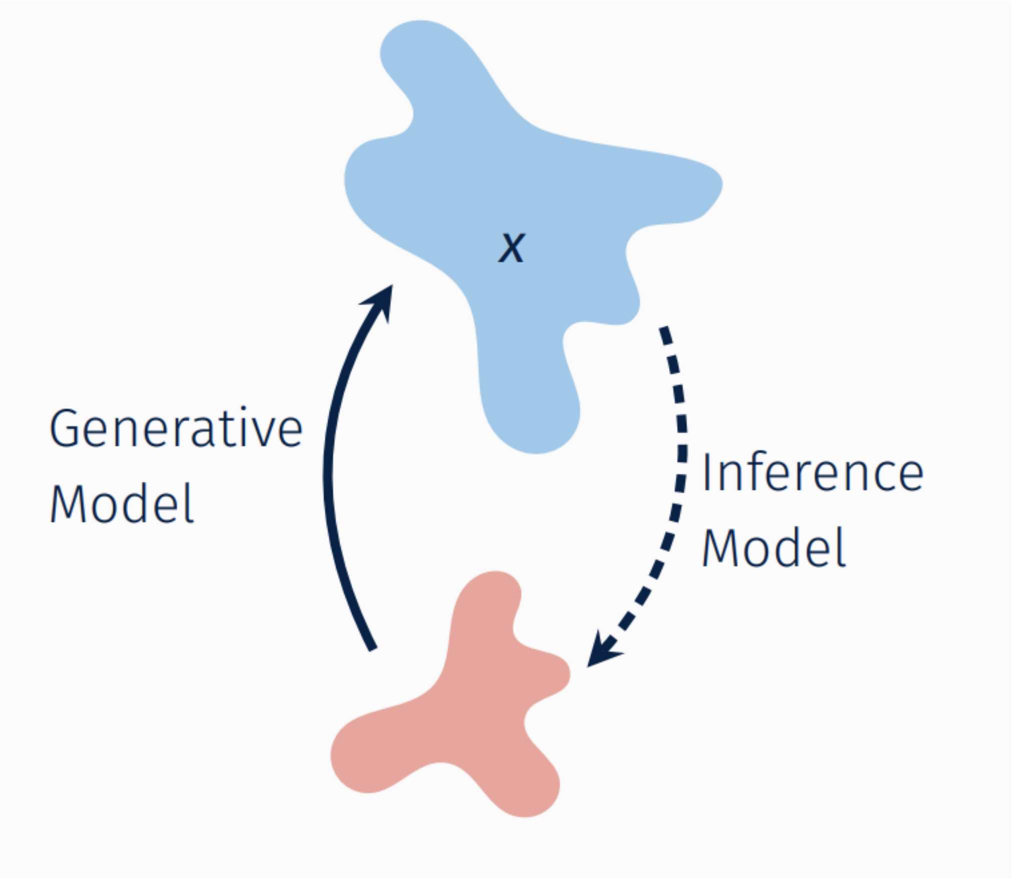
09/12/2017

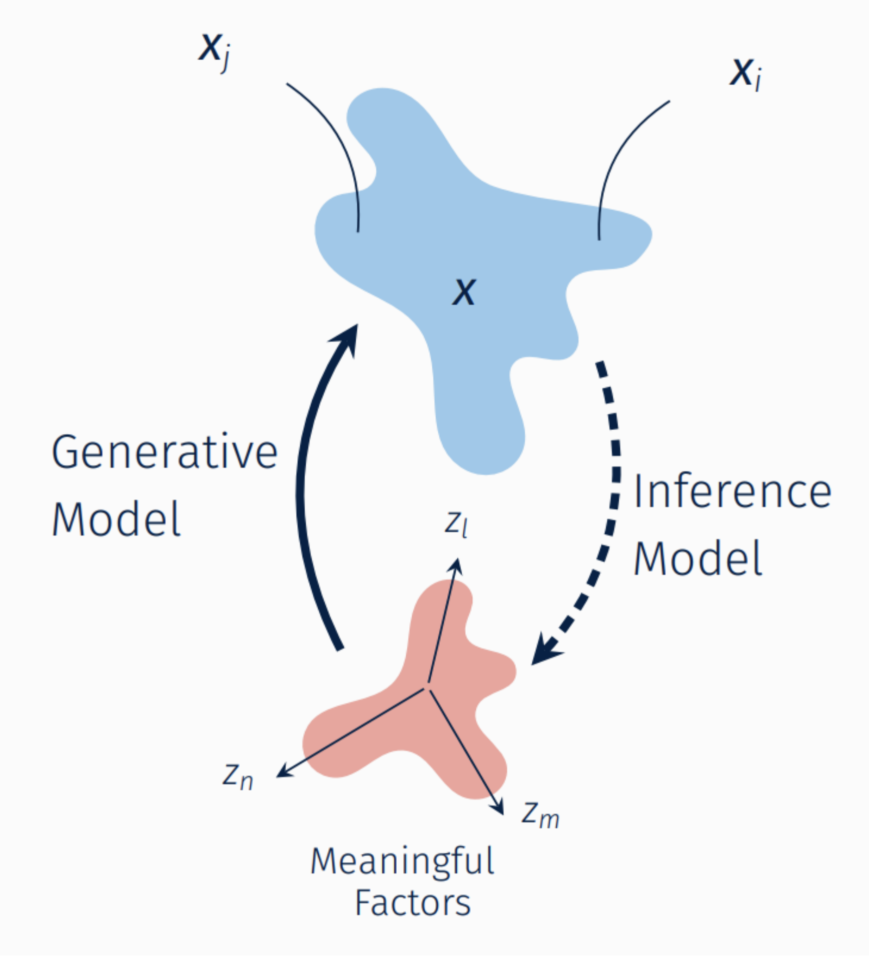
Our paper [*Sampling and inference for discrete random probability measures in probabilistic programs*](#) presented at NIPS.

Introduction

- Motivation for learning disentangled representations with deep generative models:
  - Desire to achieve interpretability.
  - The decomposability of latent representations to admit intuitive explanations.
- Conventional view of disentanglement:
  - recovering independence has subsequently motivated the development of formal evaluation metrics for independence
  - Employing regularizer explicitly encouraging independence in the representations
- Shortcomings:
  - Such an approach is not generalizable, and potentially even harmful, to learning interpretable representations for more complicated problems
  - simplistic representations cannot accurately mimic the generation of high dimensional data from low dimensional latent spaces, and more richly structured dependencies are required.

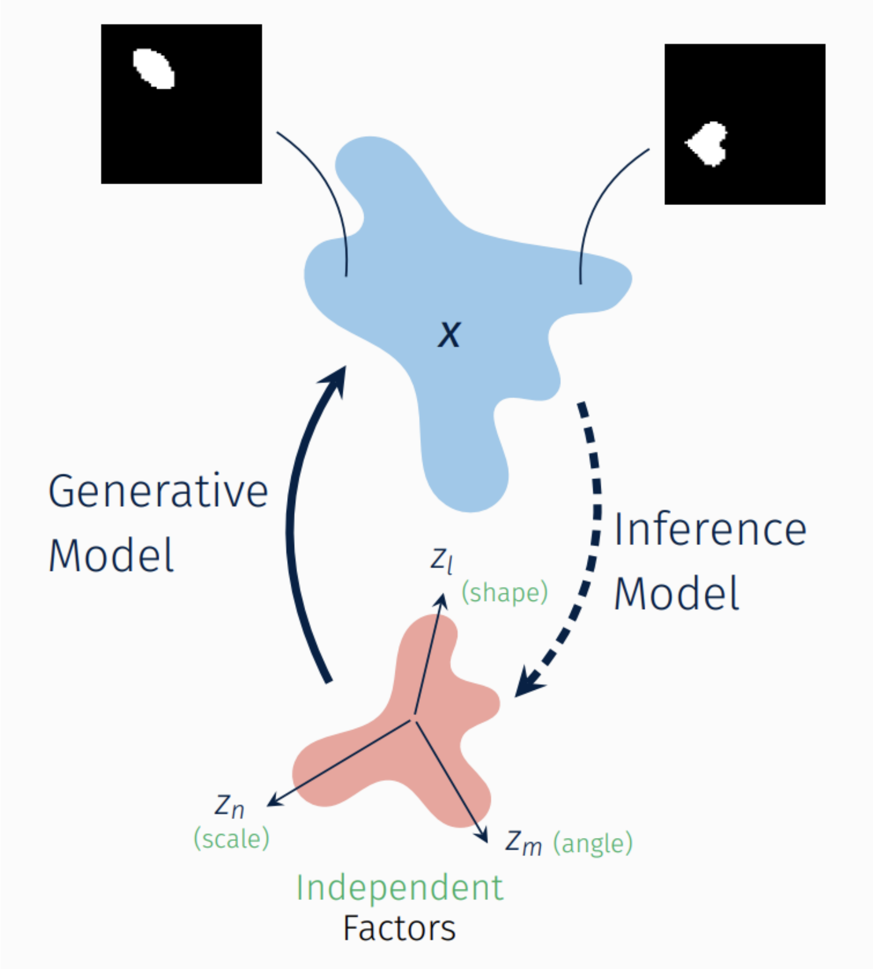
Introduction



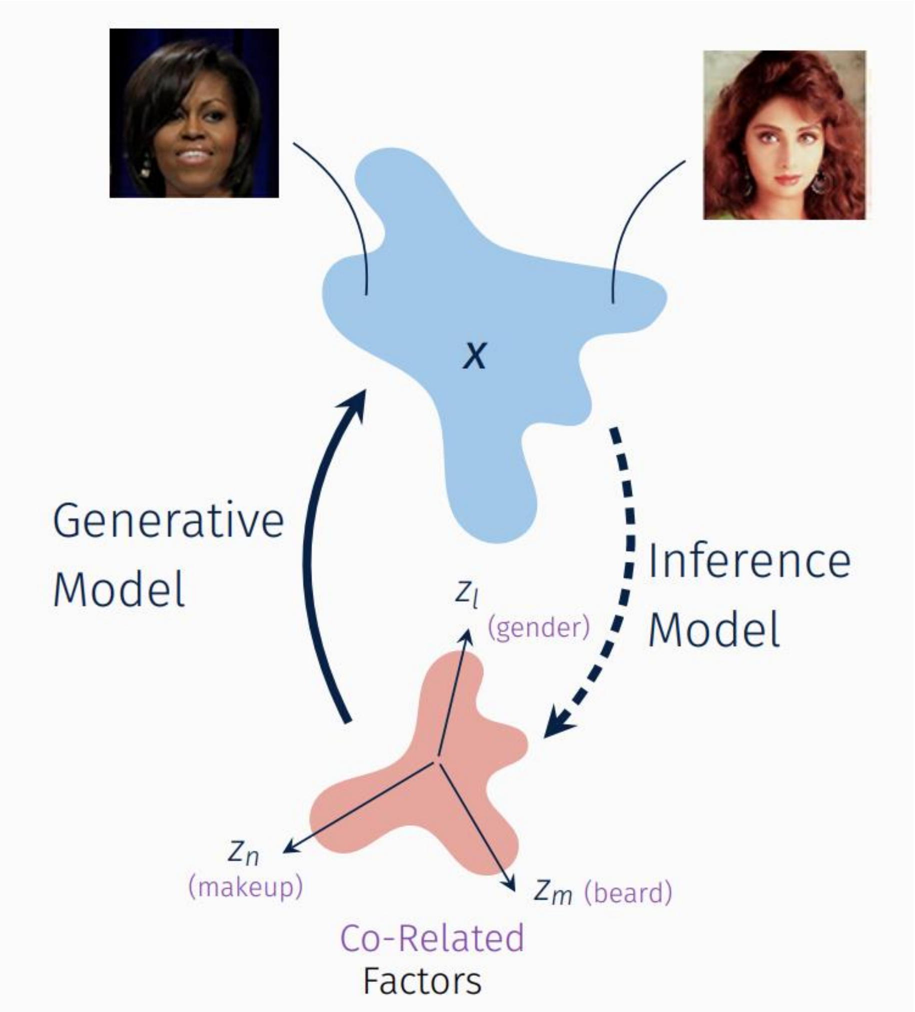


- Variational Autoencoders
- Disentanglement

Introduction



- Disentanglement = Independence



- Decomposition ∈ {Independence, Clustering, Sparsity, …}

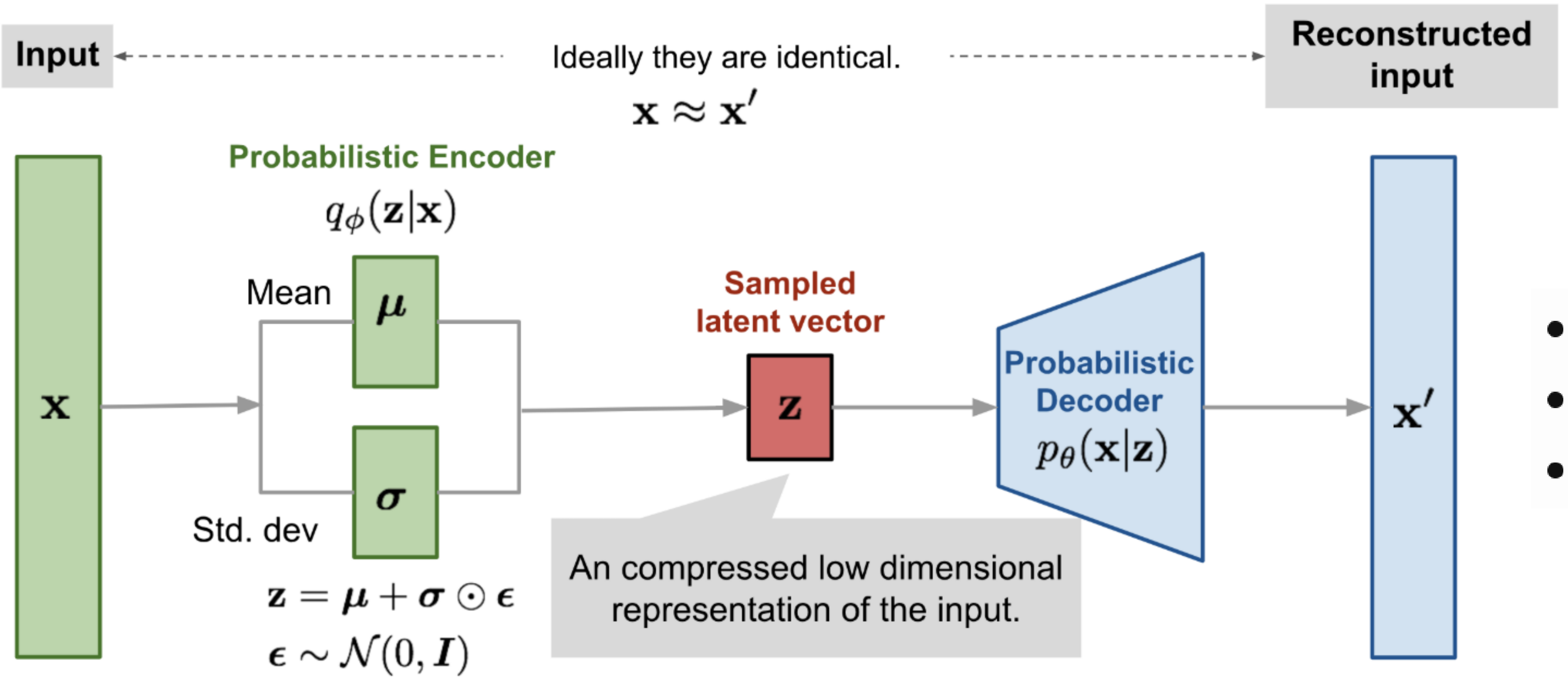
Decomposition: A Generalization of Disentanglement

- Generalization of disentanglement in VAEs:
  - decomposing their latent representations
- Characterize decomposition as the fulfilment of two factors:
  - The latent encodings of data having an appropriate level of overlap
  - The aggregate encoding of data conforming to a desired structure, represented through the prior.
- Sufficient:
  - without an appropriate level of overlap, encodings can degrade to a lookup table where the latents convey little information about data.
  - without the aggregate encoding of data following a desired structure, the encodings do not decompose as desired.
- Beta-VAE: (only allows control of latent overlap)
  - is equivalent to the VAE plus a regularizer encouraging higher encoder variance.

Contributions

- A theoretical analysis of the β-VAE objective showing it contributes to control overlap.
- Propose an alternative objective that takes into account the distinct needs of the two factors of decomposition, and use it to learn clustered and sparse representations as demonstrations of alternative forms of decomposition.
- Show that simple manipulations to the prior can improve disentanglement, and other decompositions, with little or no detriment to the reconstruction accuracy.

Related work——VAE



Ideally they are identical.

$$\mathbf{x} \approx \mathbf{x}'$$

$$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$$

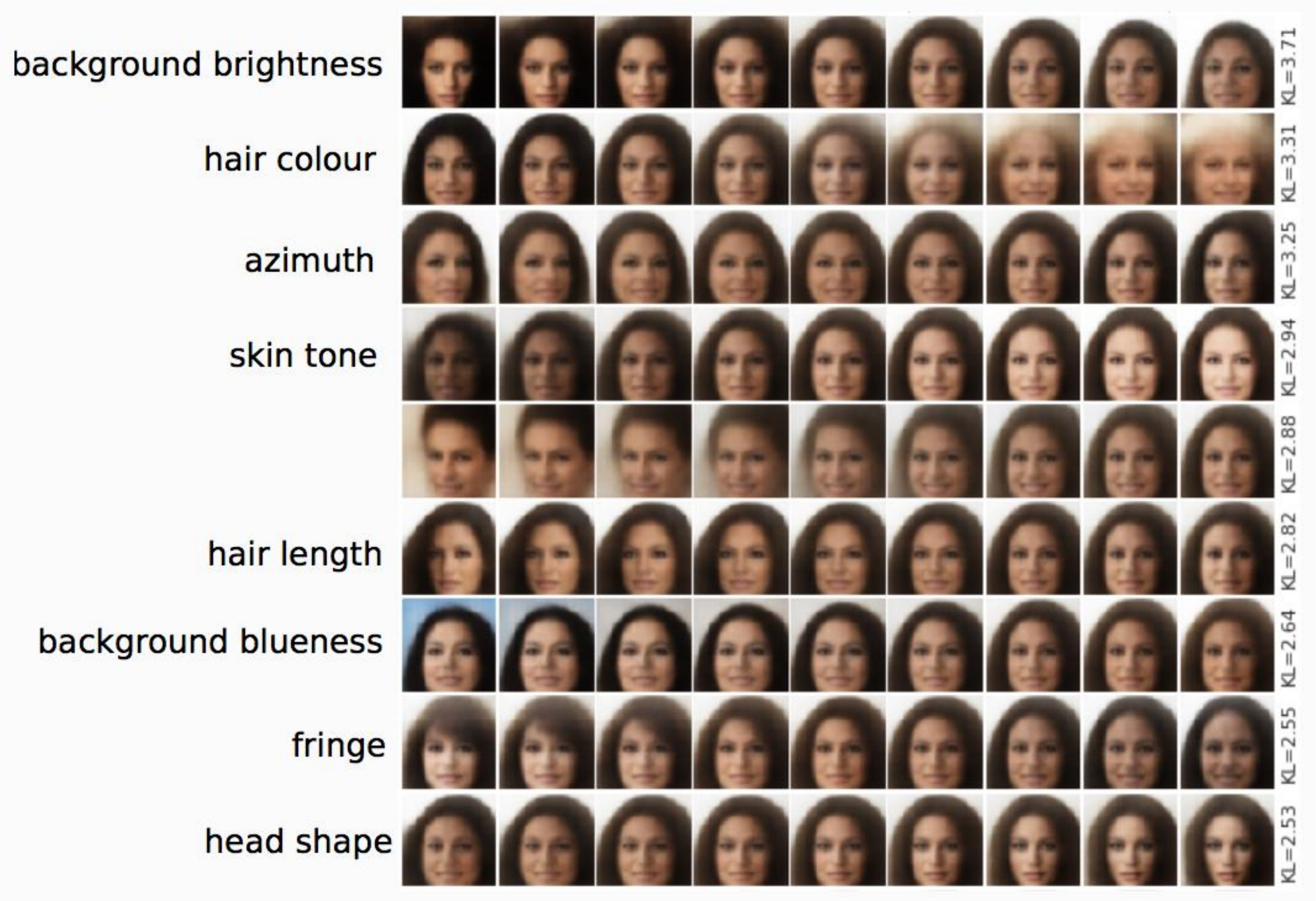
$$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$$

An compressed low dimensional representation of the input.

- Prior $p_\theta(\mathbf{z})$
- Likelihood $p_\theta(\mathbf{x}|\mathbf{z})$
- Posterior $p_\theta(\mathbf{z}|\mathbf{x})$

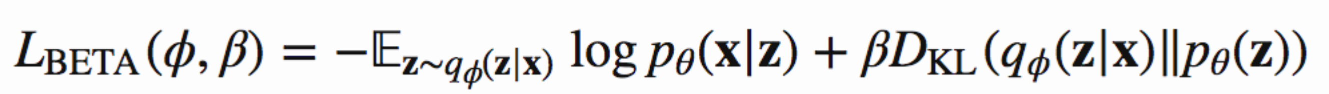
- Instead of mapping the input into a *fixed* vector, VAE want to map it into a distribution.
- The conditional probability $p_\theta(\mathbf{x}|\mathbf{z})$ defines a generative model, also known as *probabilistic decoder.*
- The approximation function $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the *probabilistic encoder.*
- The estimated posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ should be very close to the real one$p_\theta(\mathbf{x}|\mathbf{z})$
- In our case we want to minimize $D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z}|\mathbf{x}))$ with respect to ϕ.

Related work——Disentanglement



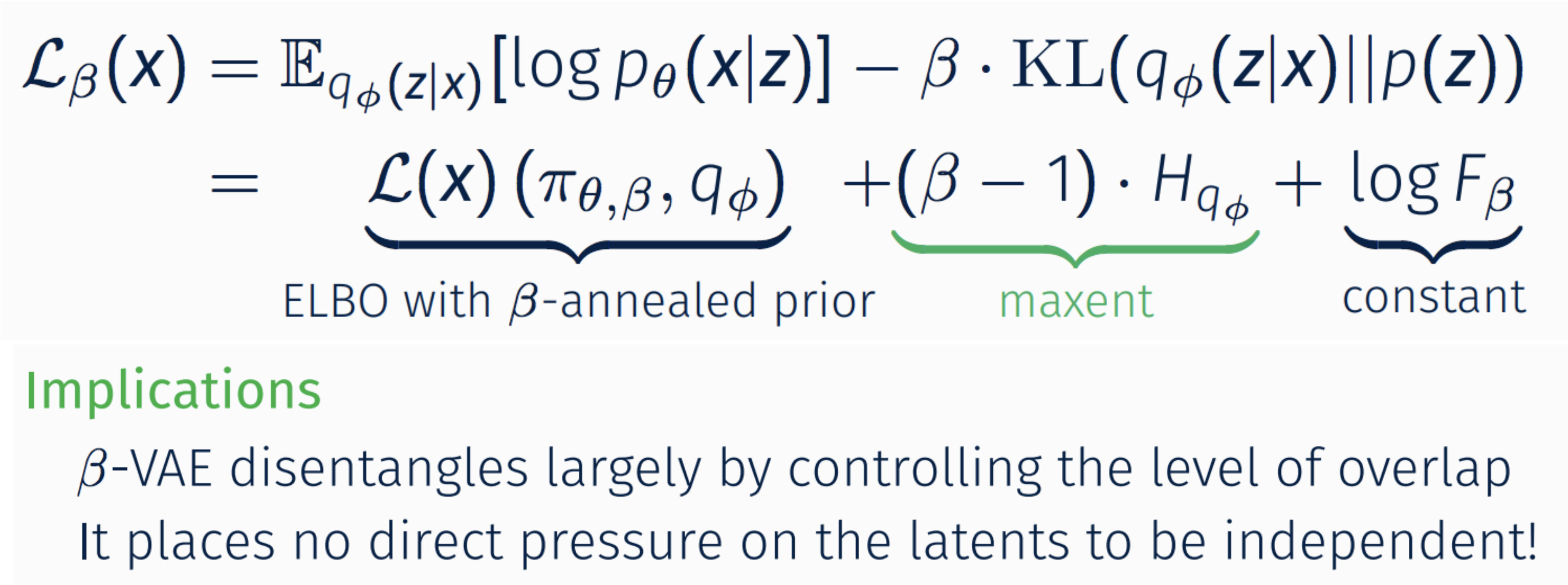
Beta-VAE

- If each variable in the inferred latent representation **z** is only sensitive to one single generative factor and relatively invariant to other factors, we will say this representation is disentangled or factorized.
- The loss function of β-VAE is defined as:



$$L_{\text{BETA}}(\phi, \beta) = -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})} \log p_\theta(\mathbf{x}|\mathbf{z}) + \beta D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p_\theta(\mathbf{z}))$$

- higher β encourages more efficient latent encoding and further encourages the disentanglement.



$$\begin{aligned}\mathcal{L}_\beta(x) &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot \mathrm{KL}(q_\phi(z|x) \| p(z)) \\ &= \underbrace{\mathcal{L}(x)\,(\pi_{\theta,\beta}, q_\phi)}_{\text{ELBO with } \beta\text{-annealed prior}} + \underbrace{(\beta - 1) \cdot H_{q_\phi}}_{\text{maxent}} + \underbrace{\log F_\beta}_{\text{constant}}\end{aligned}$$

Implications

β-VAE disentangles largely by controlling the level of overlap

It places no direct pressure on the latents to be independent!

Overlap

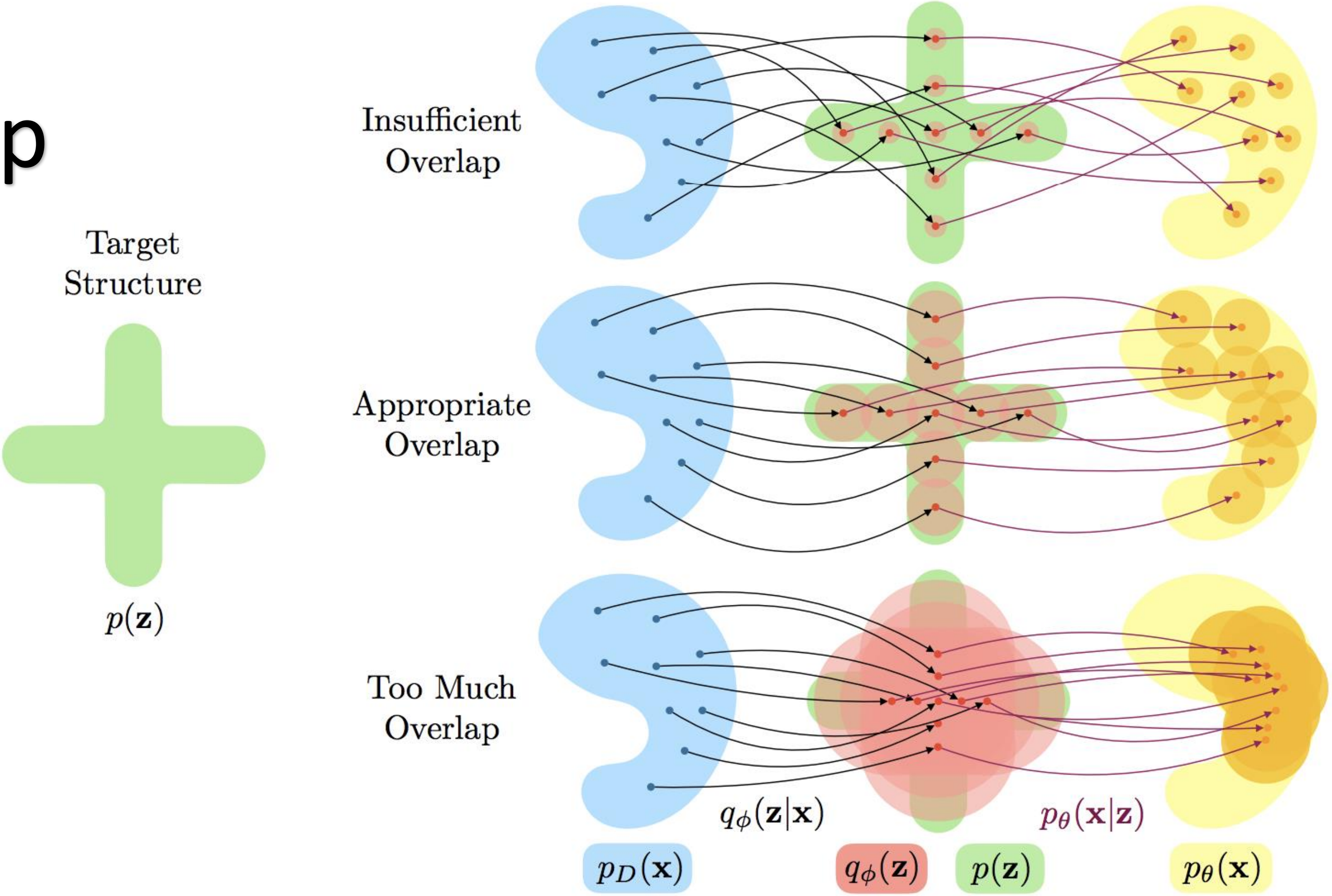
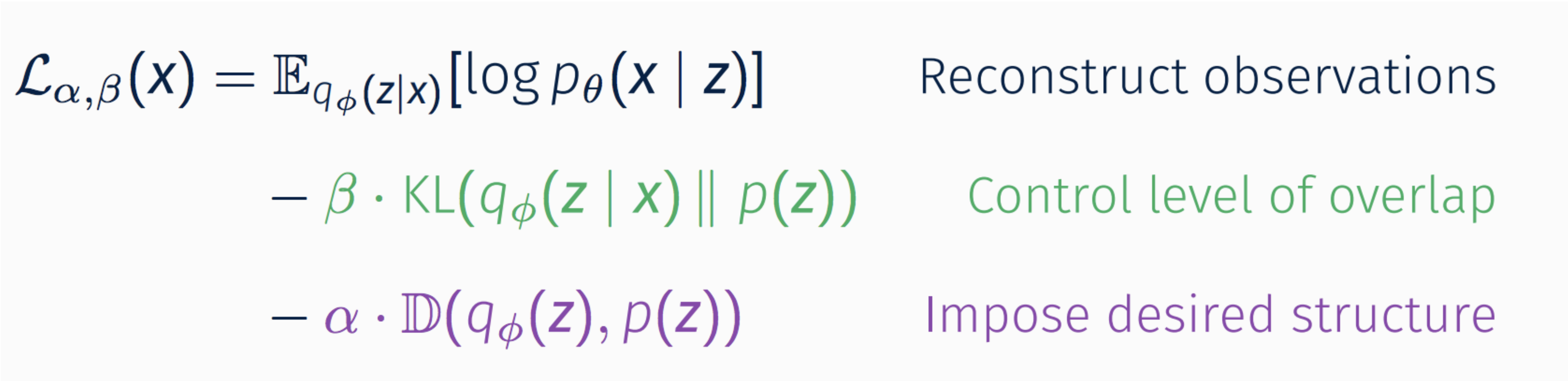


Figure 1. Illustration of decomposition where the desired structure is a cross shape (enforcing *sparsity*), expressed through the prior $p(\boldsymbol{z})$ as shown on the left. In the scenario where there is insufficient overlap [top], we observe a lookup table behavior: points that are close in the data space are not close in the latent space and so the latent space loses meaning. In the scenario where there is too much overlap [bottom], the latent variable and observed datapoint convey little information about one another, such that the latent space again loses meaning. Note that if the distributional form of the latent distribution does not match that of the prior, as is the case here, this can also prevent the aggregate encoding matching the prior when the level of overlap is large.

Decomposition: Objective



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

Reconstruct observations

$$- \beta \cdot \mathrm{KL}(q_\phi(z \mid x) \parallel p(z))$$

Control level of overlap

$$- \alpha \cdot \mathbb{D}(q_\phi(z), p(z))$$

Impose desired structure

- To incorporate direct control over the regularization (b) between the marginal posterior and the prior.
- allowing control over how much factors (a) and (b) are enforced, through appropriate setting of β and α respective

Prior for Disentanglement

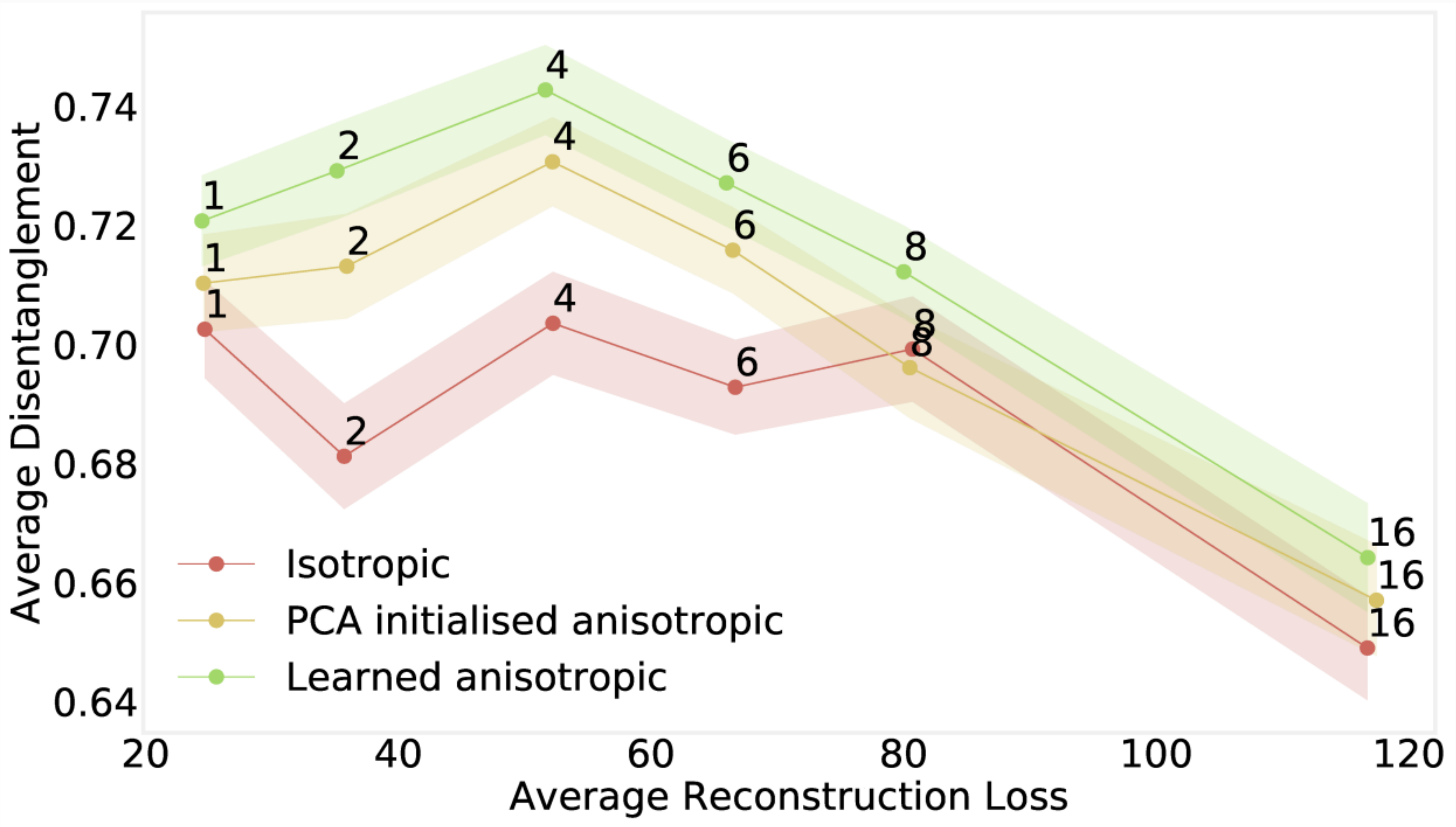


Figure 1: β-VAE trained on *2D Shapes*[1] computing disentanglement[2].

Clustered Prior

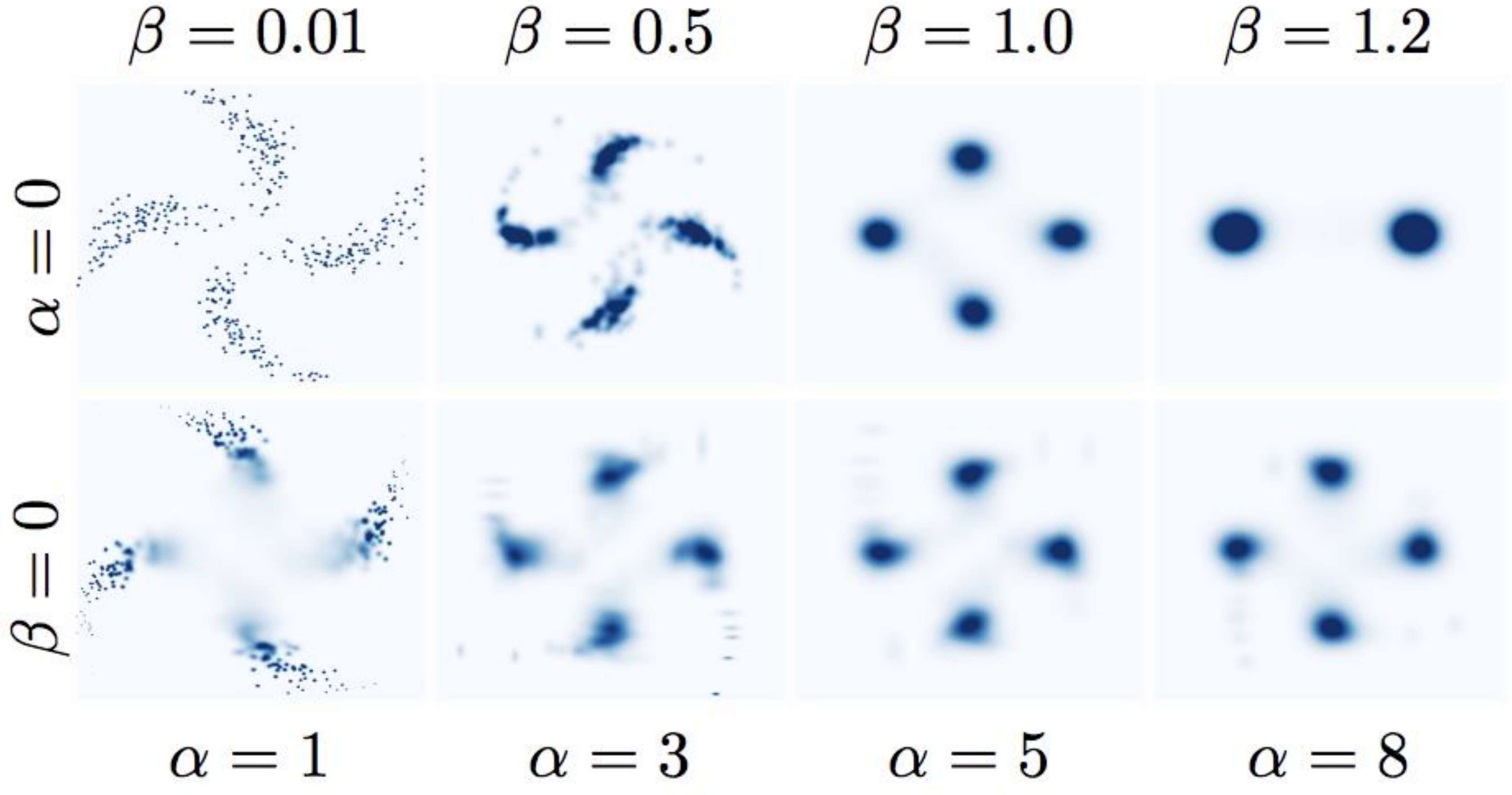


Figure 3. Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for spirals dataset with a mixture of Gaussian prior.

Prior for Sparsity

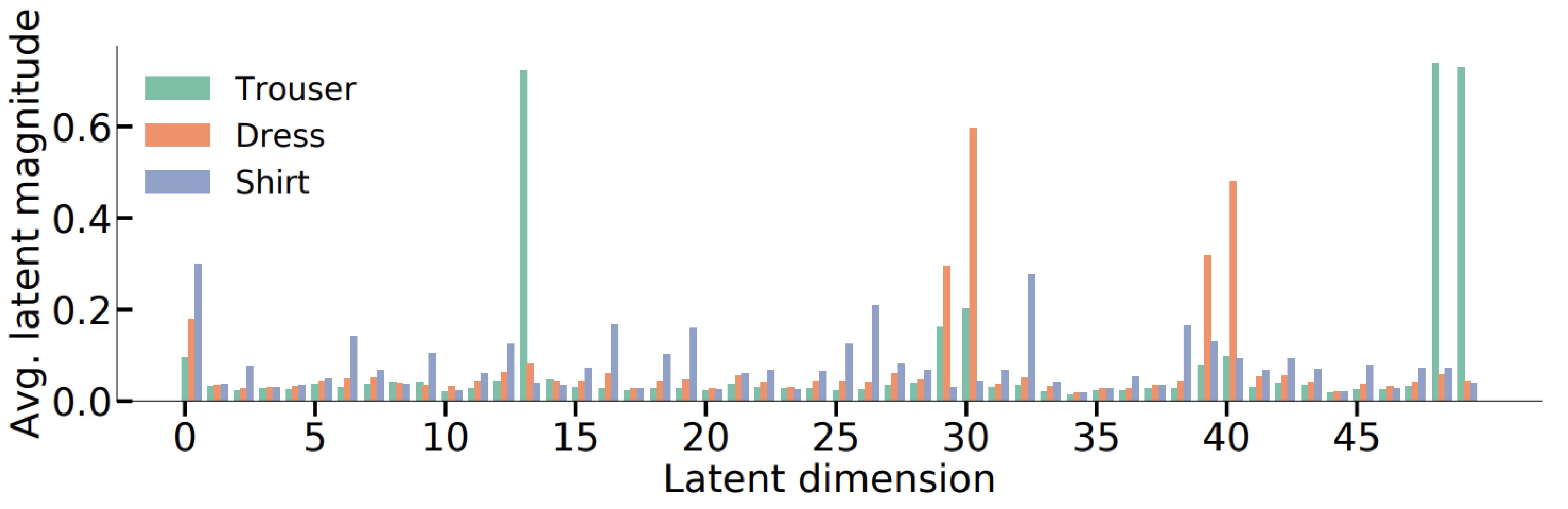
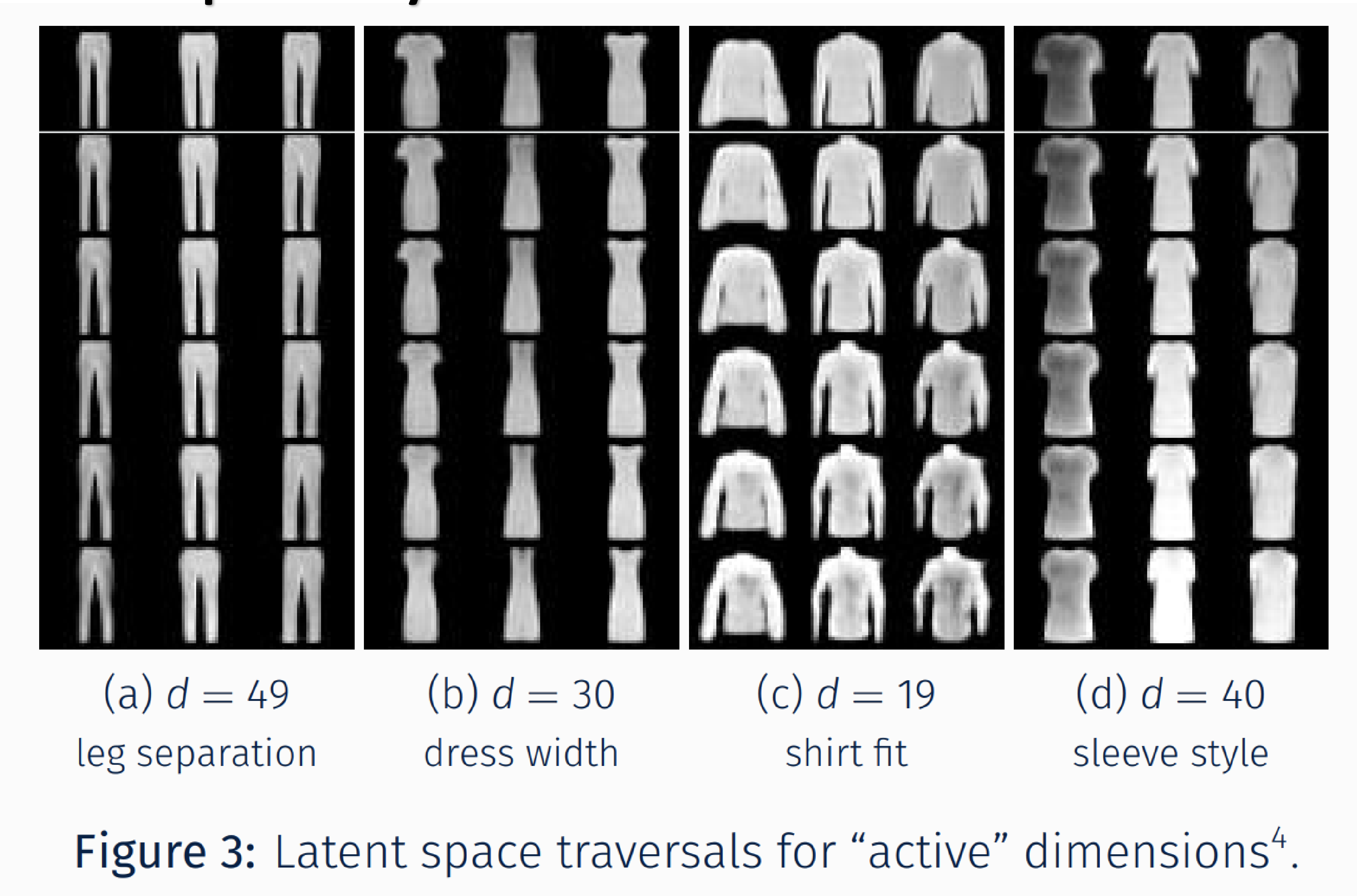


Figure 3: Sparsity of learnt representations for the *Fashion-MNIST*[4] dataset.

Prior for Sparsity



(a) $d = 49$
leg separation

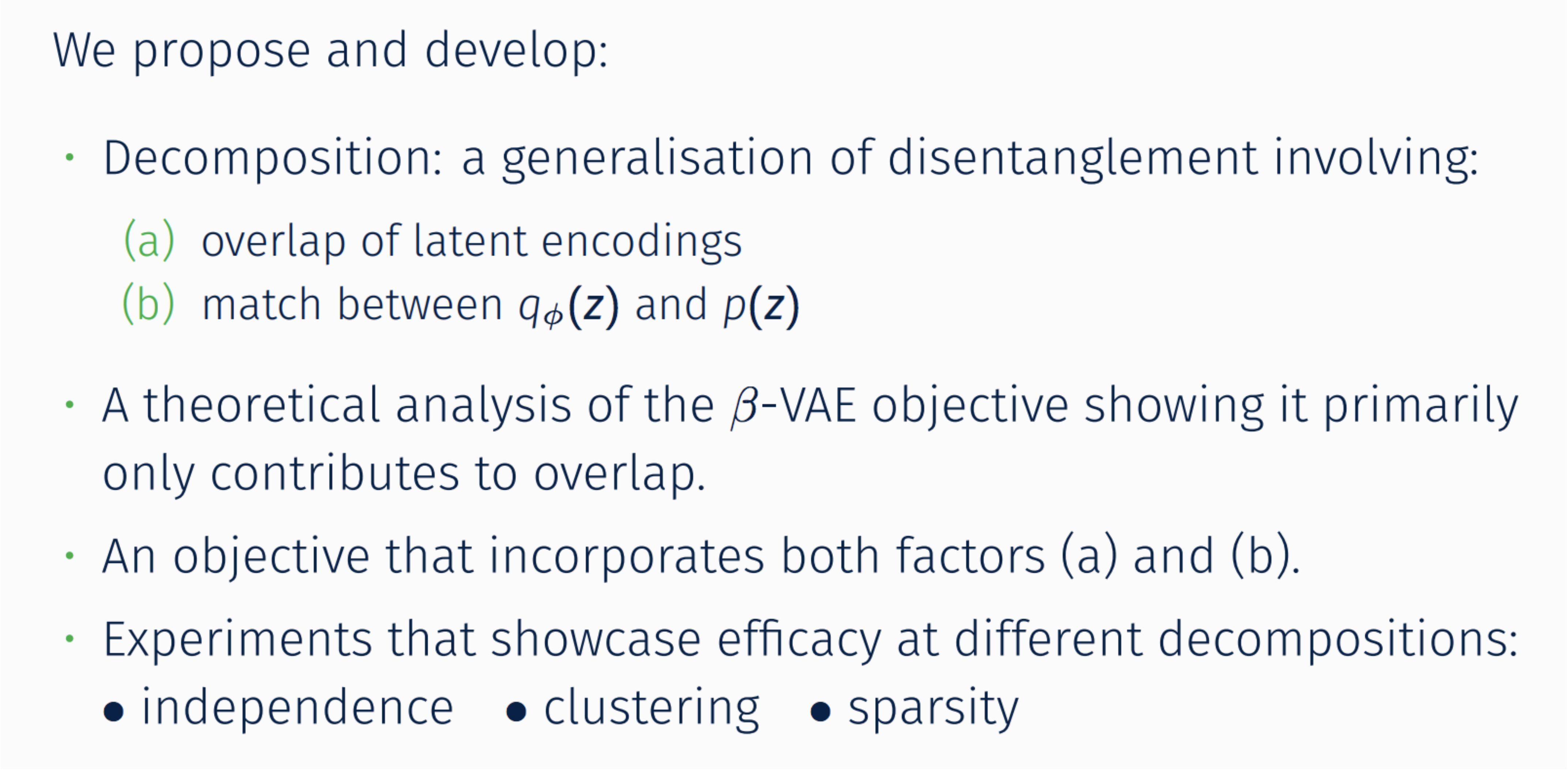
(b) $d = 30$
dress width

(c) $d = 19$
shirt fit

(d) $d = 40$
sleeve style

Figure 3: Latent space traversals for “active” dimensions[4].

Conclusion



We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) overlap of latent encodings
  (b) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$
- A theoretical analysis of the β-VAE objective showing it primarily only contributes to overlap.
- An objective that incorporates both factors (a) and (b).
- Experiments that showcase efficacy at different decompositions: • independence • clustering • sparsity