Auto-Encoding Variational Bayes

Philip Ball, Alexandru Coca, Omar Mahmood, Richard Shen

UNIVERSITY OF CAMBRIDGE
DEPARTMENT OF ENGINEERING

March 16, 2018

Motivation

Latent variable models (LVMs) are a class of statistical models that aim to represent the structure of complex, high-dimensional data in a compact manner. Such models can facilitate classification tasks, and can be used for knowledge discovery [1] or data compression.

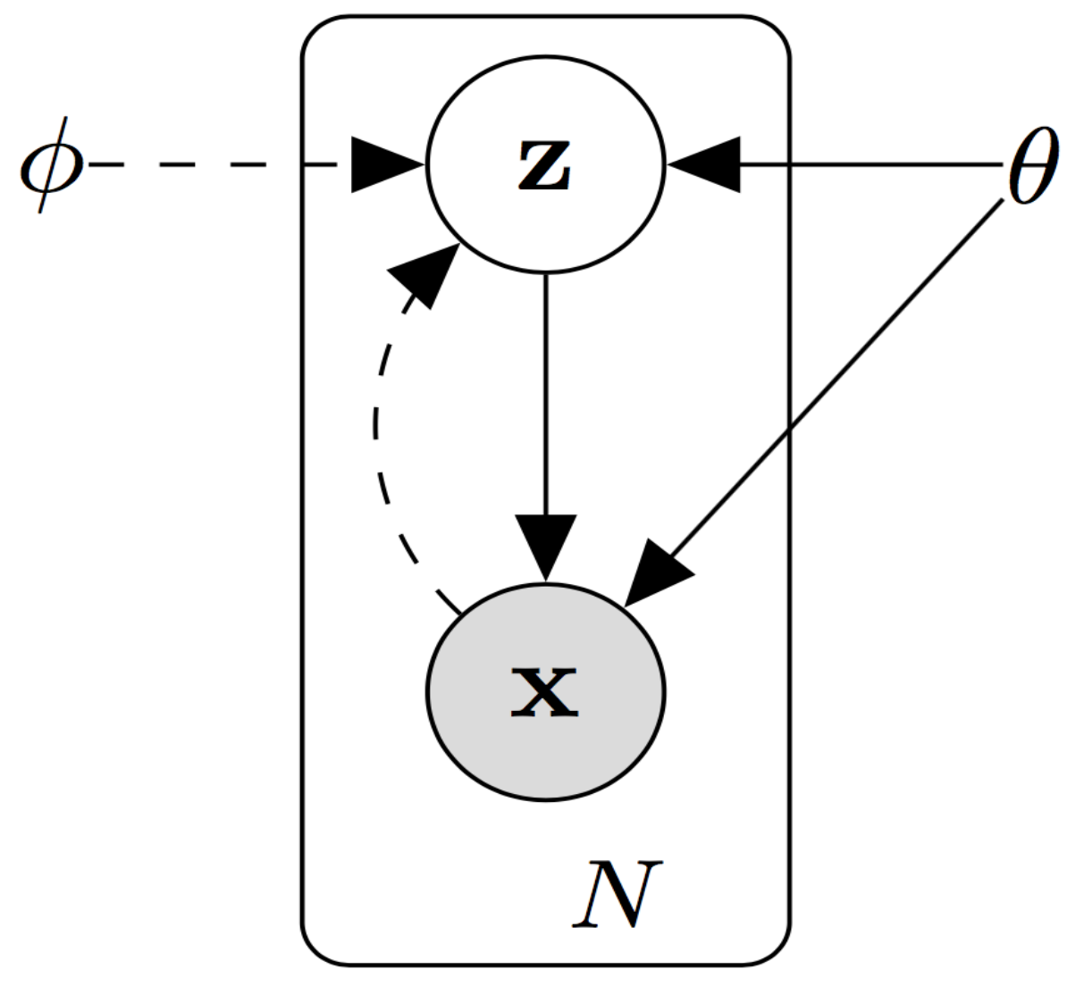

Figure 1: LVM structure

Learning Autoencoder Structure

In a *variational autoencoder*, one aims to jointly learn a data generating process ($p_\theta(\mathbf{x}|\mathbf{z})$) and the posterior distribution over the latent variables ($p(\mathbf{z}|\mathbf{x})$). Using density networks to model the data likelihood yields expressive models but with intractable marginal likelihoods and posteriors over the latent variables. Setting the optimisation objective in terms of Kullback-Leibler (KL) divergences $\mathcal{D}$ and a tractable (approximate) posterior $q_\phi(\mathbf{z}|\mathbf{x})$

$$\underbrace{\log p_\theta(\mathbf{x})}_{\text{marg. likelihood}} - \underbrace{\mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})]}_{\text{approximation error} \geq 0} = \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}_{\text{expected reconstruction error}} - \underbrace{\mathcal{D}\left[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right]}_{\text{regularisation term}} \quad (1)$$

allows optimisation of a lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ on the marginal likelihood $p_\theta(\mathbf{x})$ equal to the RHS of (1). The approximate posterior is given by $q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}; \mu(\mathbf{x}), \sigma^2(\mathbf{x}))$ where the mean and variance are nonlinear functions of ϕ modeled by a neural network.

The Reparameterisation Trick

Optimising the expectation in the RHS of (1) with respect to ϕ involves backpropagating the error term through a layer of samples of q which is not differentiable (Figure 2, left).

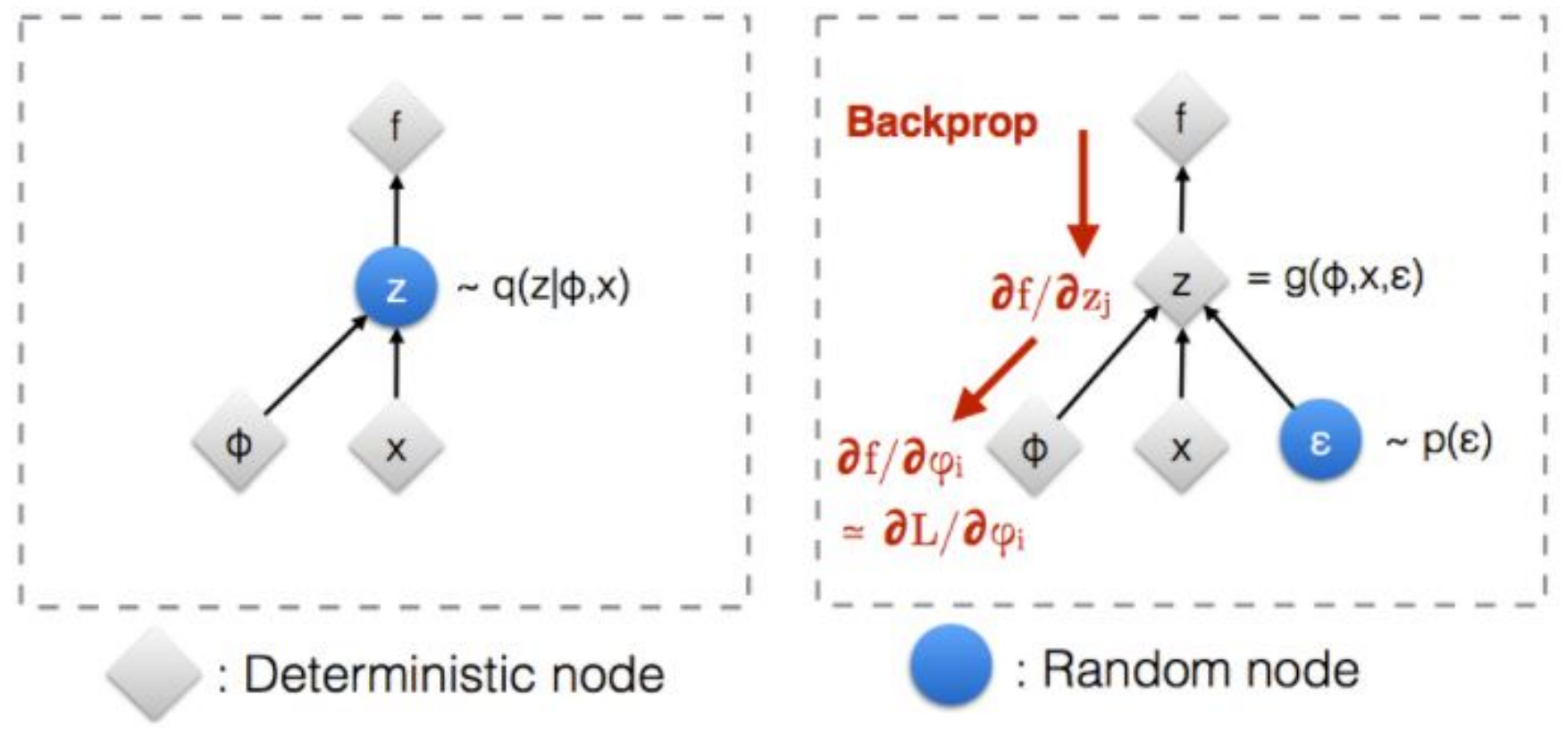

Figure 2: The reparametrisation trick

This is overcome (Figure 2, right) by expressing $\mathbf{x}$ as a deterministic function g of an *auxiliary variable*, $\epsilon \sim p(\epsilon)$, continuous with respect to ϕ, which for Gaussian q is

$$\mathbf{z} = g_\phi(\epsilon, \mathbf{x}) = \mu + \sigma \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I})$$

MNIST Training Curves

We closely reproduce results on MNIST reported in [2] for the average variational lower bound $\mathcal{L}$ for VAEs with the specified latent space dimensions. We observe that increasing the number of latent variables from 20 to 200 does not lead to overfitting.

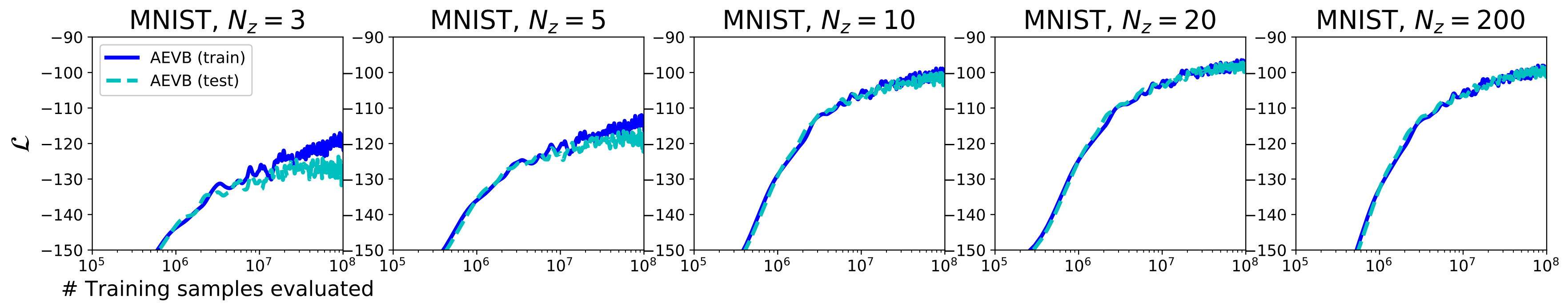

Figure 3: MNIST data set training and testing $\mathcal{L}$ for different latent variable dimensionality

Latent Manifold Visualisation

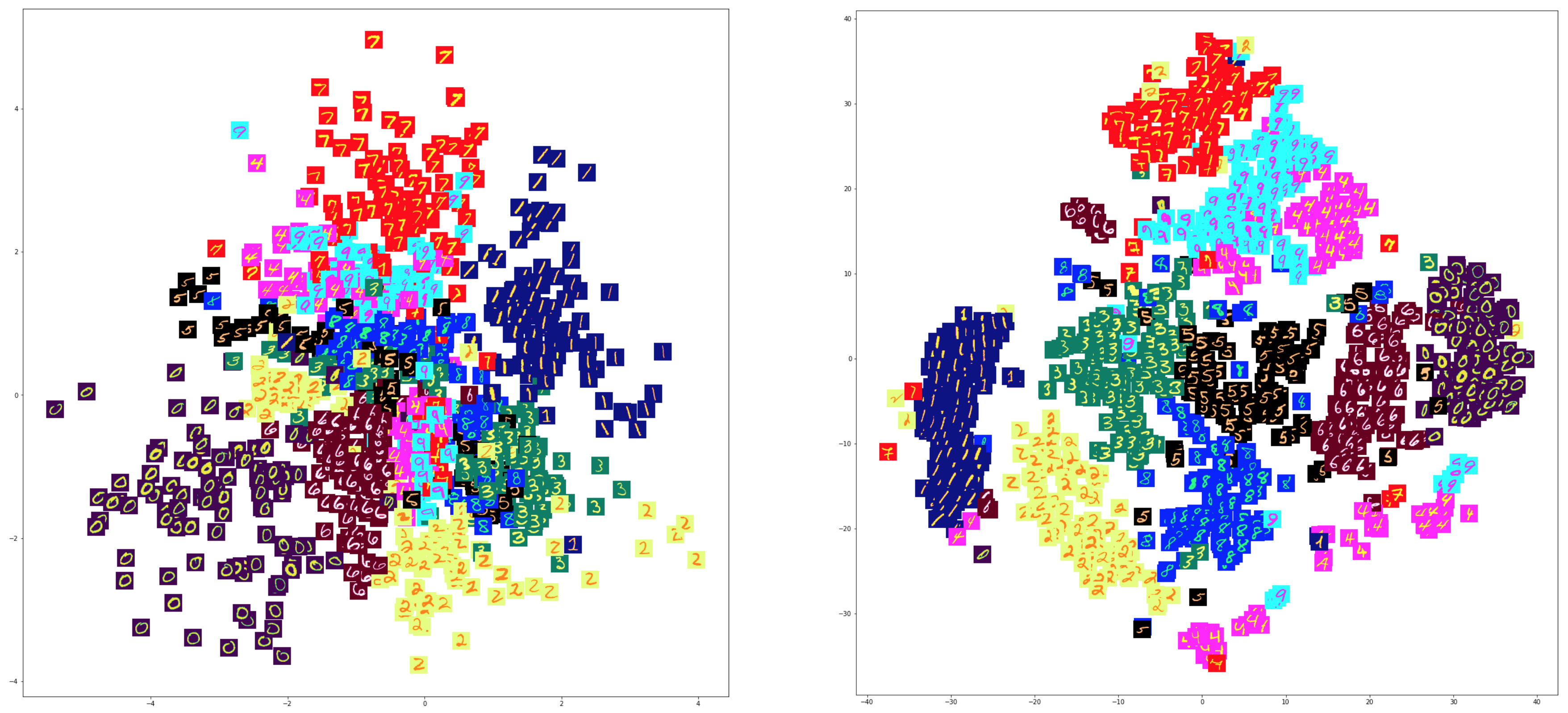

Figure 4: **Left**: 2D latent space **Right**: t-SNE projection of 20D latent space

We observe the effect of the regularisation term in (1) by comparing plots of input data mapped through the encoder onto the latent space. The 20D latent space shows better separation than the 2D space due to the additional degrees of freedom whilst maintaining a valid Gaussian distribution.

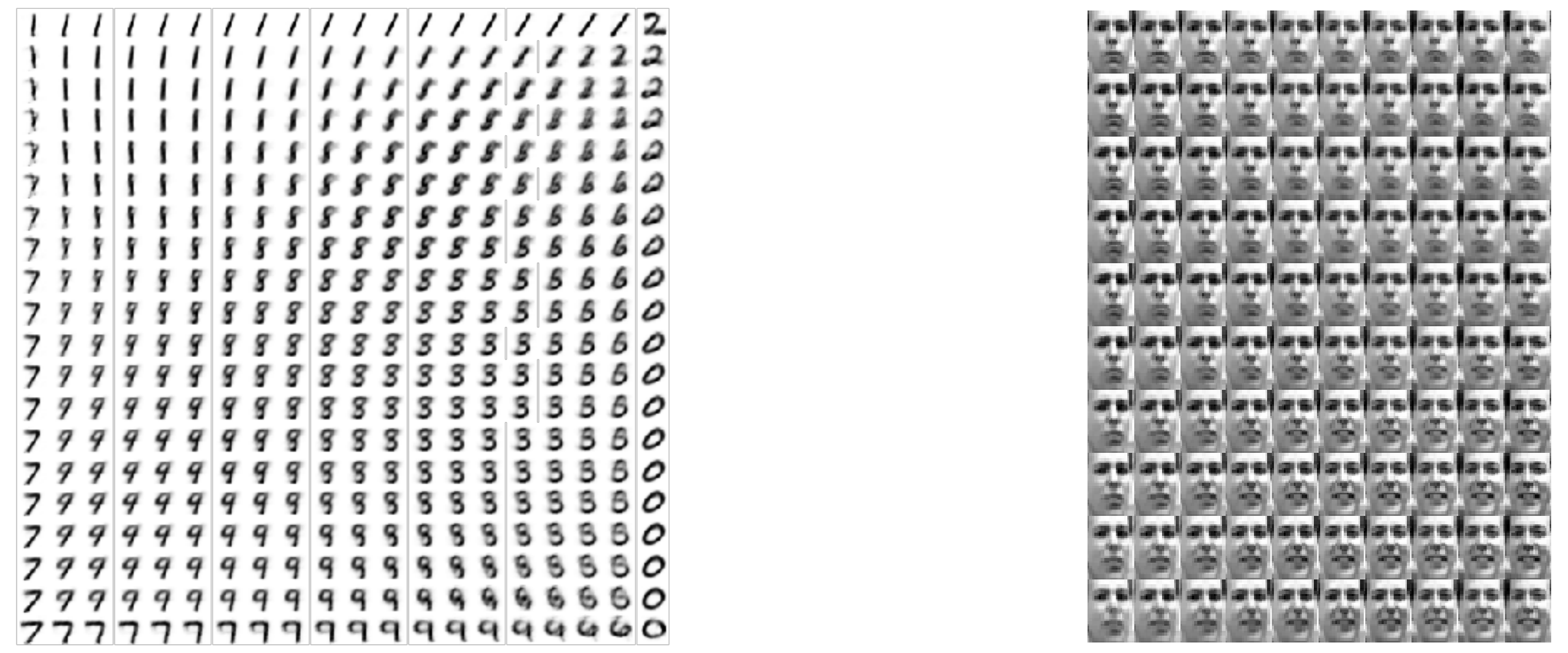

Figure 5: **Left**: Learnt 2D MNIST manifold **Right**: Learnt 2D Frey face manifold

Mapping a grid of values on the unit square through the inverse Gaussian CDF and the trained decoder (2D latent space) allows visualisation of the learned manifolds. These show the ability to change one underlying data property (i.e. rotation) by varying along a single latent dimension.

Importance Weighted Autoencoder (IWAE)

The term $\mathcal{D}[q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x})]$ in (1) penalizes low-probability posterior samples. Thus, the VAE posterior is a good approximation only if the true posterior can be obtained by nonlinear regression.

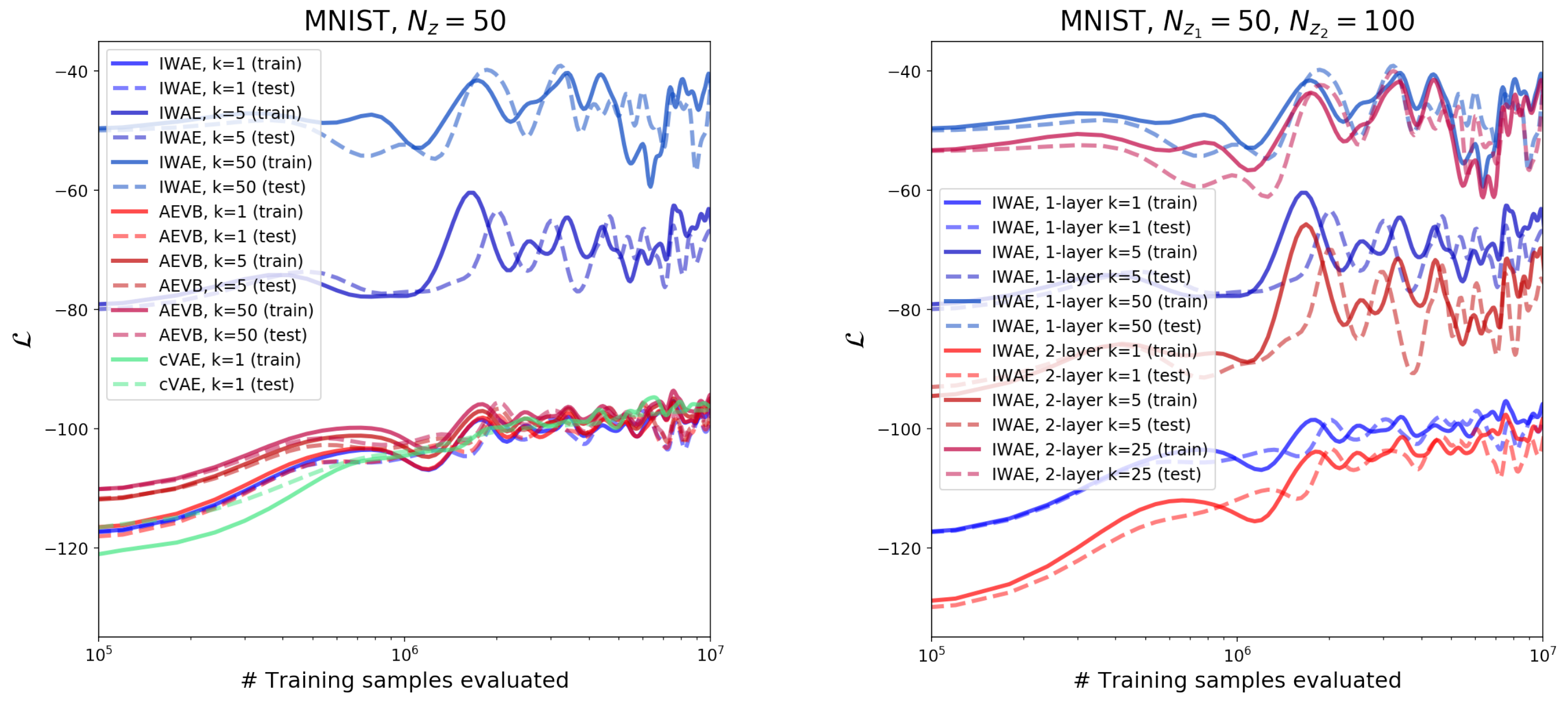

Figure 6: **Left**: VAE and IWAE **Right**: IWAE 1- and 2-stochastic layers

This assumption can be relaxed by sampling low-probability posterior regions using *importance sampling*, which yields a tighter $\mathcal{L}_k$ on the marginal likelihood [3]:

$$\mathcal{L}_k = \mathbb{E}_{\mathbf{z}_1,\ldots,\mathbf{z}_k \sim q(\mathbf{z}|\mathbf{x})}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(\mathbf{x}, \mathbf{z}_i)}{q(\mathbf{z}_i|\mathbf{x})}\right] \quad (2)$$

Conditional VAEs (cVAEs)

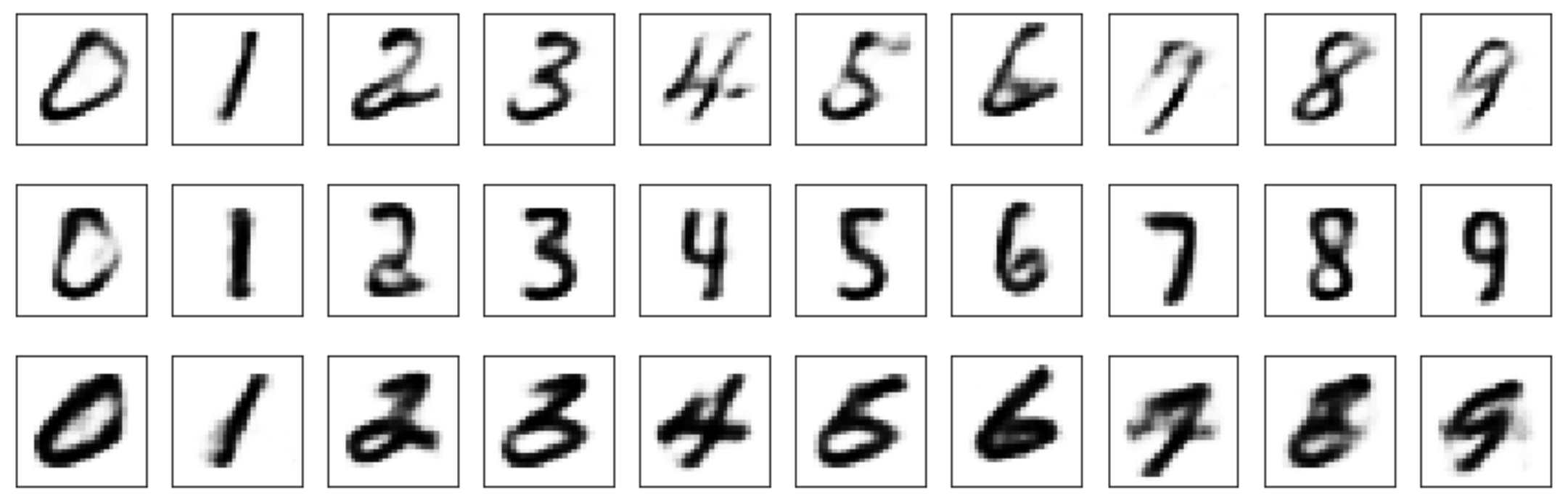

Figure 7: Samples generated from the same noise sample but different labels

Conditional VAEs (cVAEs) [4] include additional information (i.e. labels) at the input and stochastic layers.

Future Work

- Convolutional/recurrent encoder and decoder architectures
- Generalise to colour images
- Different prior distributions over the latent space

References

[1] M. Kusner, B. Paige, J. M. Hernández-Lobato. *Grammar variational autoencoder.* arXiv preprint arXiv:1703.01925 (2017)

[2] D. Kingma, M. Welling. *Auto-Encoding Variational Bayes.* arXiv preprint arXiv:1312.6114 (2014)

[3] Y. Burda, R. Grosse, R. Salakhutdinov. *Importance weighted autoencoders.* arXiv preprint arXiv:1509.00519 (2015).

[4] J. Walker, C. Doersch, A. Gupta, M. Herbert. *An Uncertain Future: Forecasting from Static Images using Variational Autoencoders.* arXiv preprint arXiv:1606.07873 (2016)