Ladder VAE

Hantao Zhang

Introduction

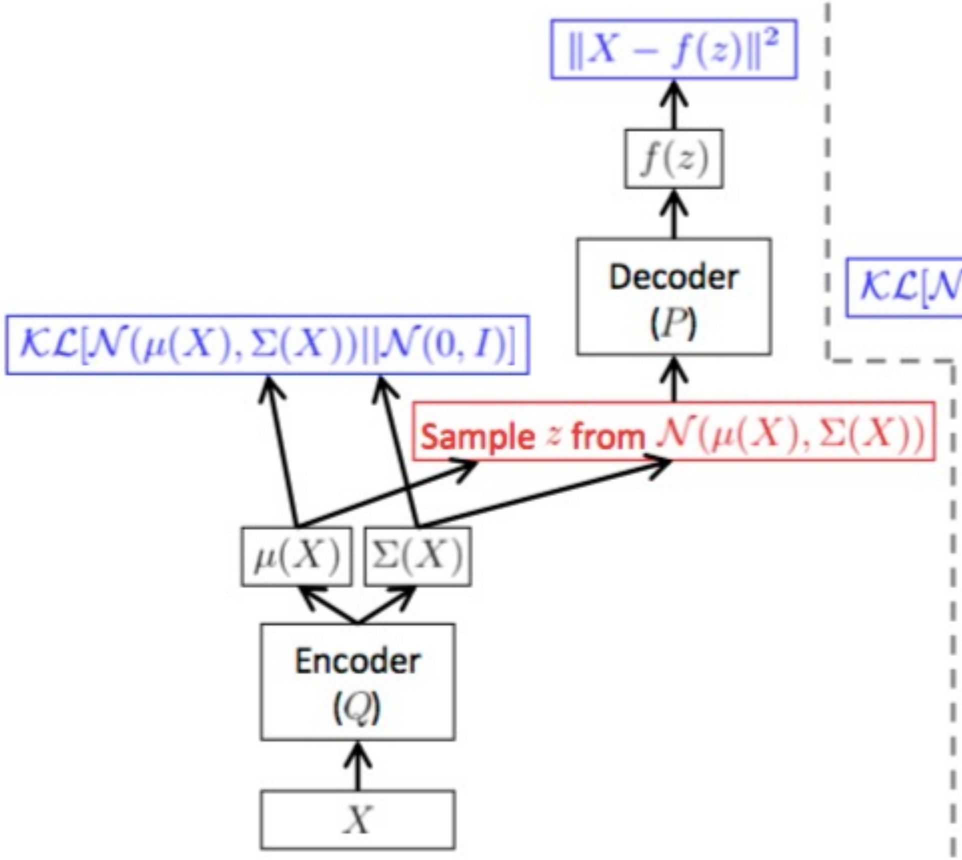
- Ladder VAE (LVAE) was introduced in 2016, just after VAE.
- Explores variational inference part of VAE model

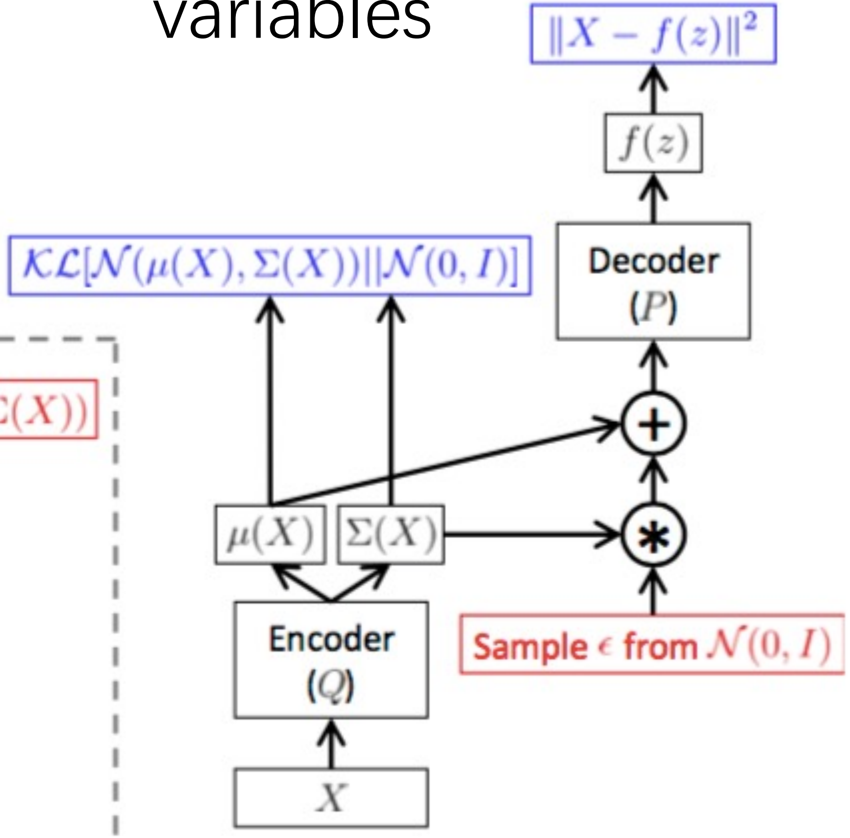
Main change

- recursively corrects the generative distribution by a data dependent approximate likelihood

Review of VAE

- variational inference -> generative
- hierarchies of conditional stochastic variables





The Problem

- VAE models are hierarchical
- Difficult to optimize when num_layers++
- （high order layers learns nothing）
- Constrained complexity

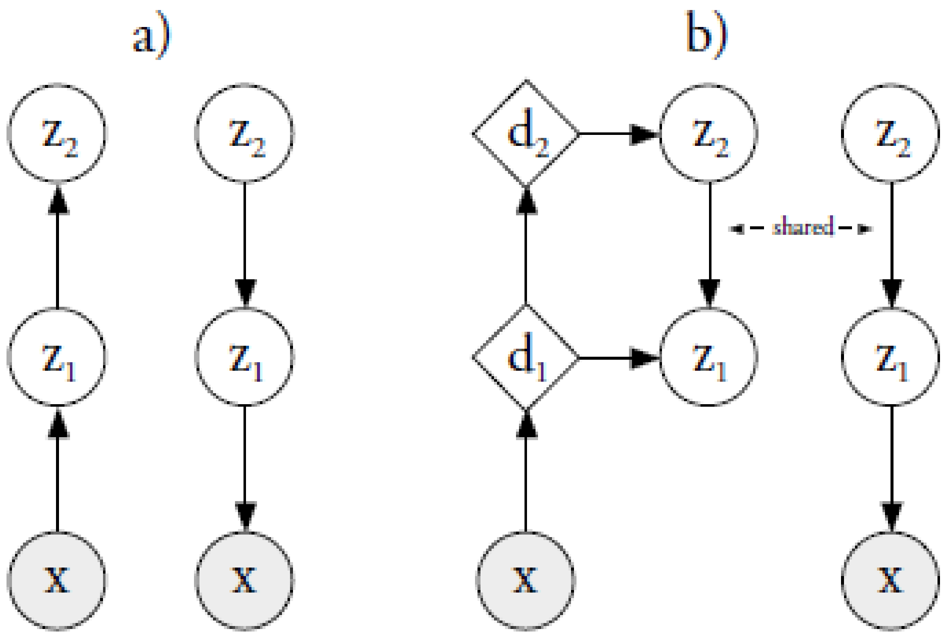
Main Contribution

- Proposed Ladder VAE architecture to support deep hierarchical encoder.
- Verified the importance of BatchNorm (BN) and Warm-Up (WU)

Model Architecture

VAE

LVAE



- Shared information between encoder and decoder
- Deterministic upward pass
- Followed by stochastic downward pass

Model cont.

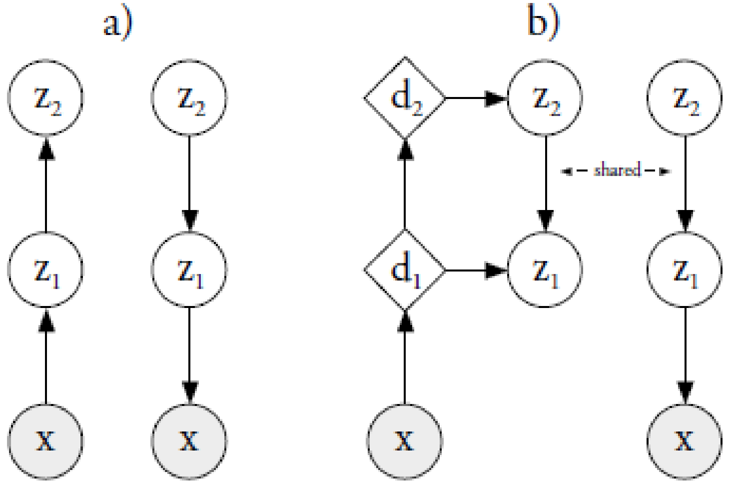
Variational Regularization Term

- **Objective**
- $\log p(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta, \phi; x) = \boxed{-KL\left(q_\phi(z|x)||p_\theta(z)\right)} + \boxed{E_{q_\phi(z|x)}[\log p_\theta(x|z)]}$

Reconstruction Error

- Generative arch (Decoder)
- $p_\theta(z) = p_\theta(z_L)\prod_{i=1}^{L-1} p_\theta(z_i|z_{i+1})$
- $p_\theta(z_i|z_{i+1}) = N\left(z_i\middle|\mu_{p,i}(z_{i+1}), \sigma_{i+1}^2(z_{i+1})\right), p_\theta(z_L) = N(z_L|0, \mathrm{I})$
- $p_\theta(x|z_1) = N\left(x\middle|\mu_{p,0}(z_1), \sigma_{p,0}^2(z_1)\right)$

Model cont. (Inference arch)



- VAE
- $q_\phi(z|x) = q_\phi(z_1|x) \prod_{i=2}^{L} q_\phi(z_i|z_{i-1})$
- $q_\phi(z_1|x) = N\left(z_1 \middle| \mu_{q,1}(x), \sigma_{q,1}^2(x)\right)$
- $q_\phi(z_i|z_{i-1}) = \mathrm{N}\left(z_i \middle| \mu_{q,i}(z_{i-1}), \sigma_{q,i}^2(z_{i-1})\right), \mathrm{i} = 2 \ldots L$

- $d(y) = \mathrm{MLP}(y)$
- $\mu(y) = \mathrm{Linear}(d(y))$
- $\sigma^2(y) = \mathrm{Softplus}\left(\mathrm{Linear}(d(y))\right)$

- LVAE
- $\sigma_{q,i} = \frac{1}{\hat{\sigma}_{q,i}^{-2} + \sigma_{p,i}^{-2}}$
- $\mu_{q,i} = \frac{\hat{\mu}_{q,i}\hat{\sigma}_{q,i}^{-2} + \mu_{p,i}\sigma_{p,i}^{-2}}{\hat{\sigma}_{q,i}^{-2} + \sigma_{p,i}^{-2}}$
- $\sigma_{q,L} = \hat{\sigma}_{q,L}, \ \mu_{q,L} = \hat{\mu}_{q,L}$
- $q_\phi(Z_i|\cdot) = N\left(z_i \middle| \mu_{q,i}, \sigma_{q,i}^2\right)$

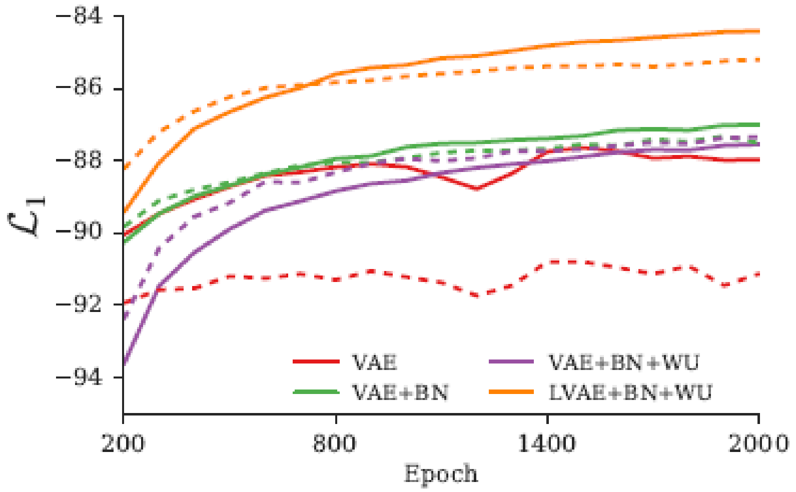
- $d_n = \mathrm{MLP}(d_{n-1}), d_0 = x$
- $\hat{\mu}_{q,i} = \mathrm{Linear}(d_i), i = 1 \ldots L$
- $\hat{\sigma}_{q,i}^2 = \mathrm{Softplus}\left(\mathrm{Linear}(d_i)\right), i = 1 \ldots L$

Warm-Up

- Motivation
- Large number of units becomes inactive in early stage of training
- Solution
- Initialize training using reconstruction error only

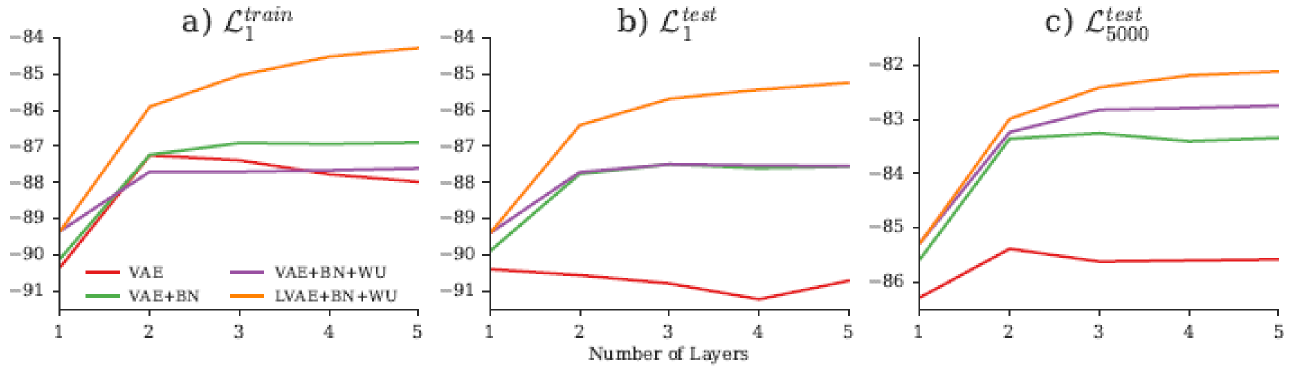
- $\log p\,(x) \geq E_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = L(\theta, \phi; x)$
- $= -\beta KL\left(q_\phi(z|x)||p_\phi(z)\right) + E_{q_\phi(z|x)}[\log p_\theta\,(x|z)]$

Experiments

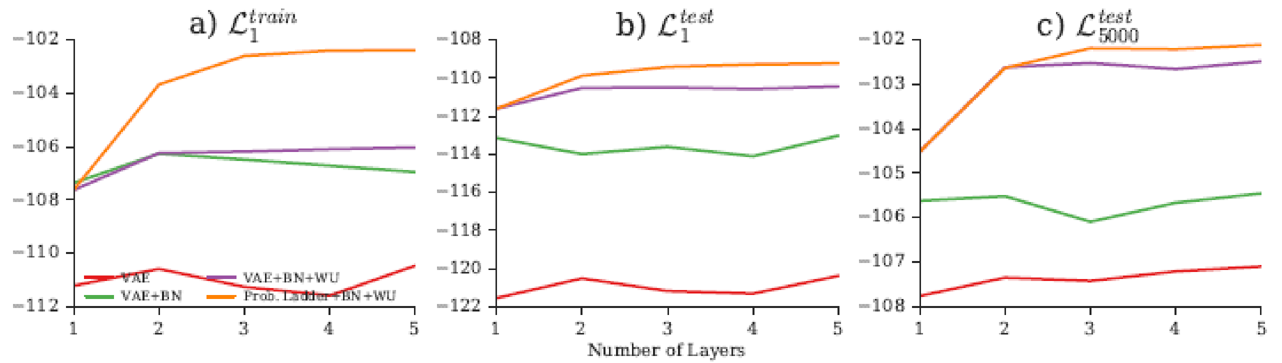


MNIST

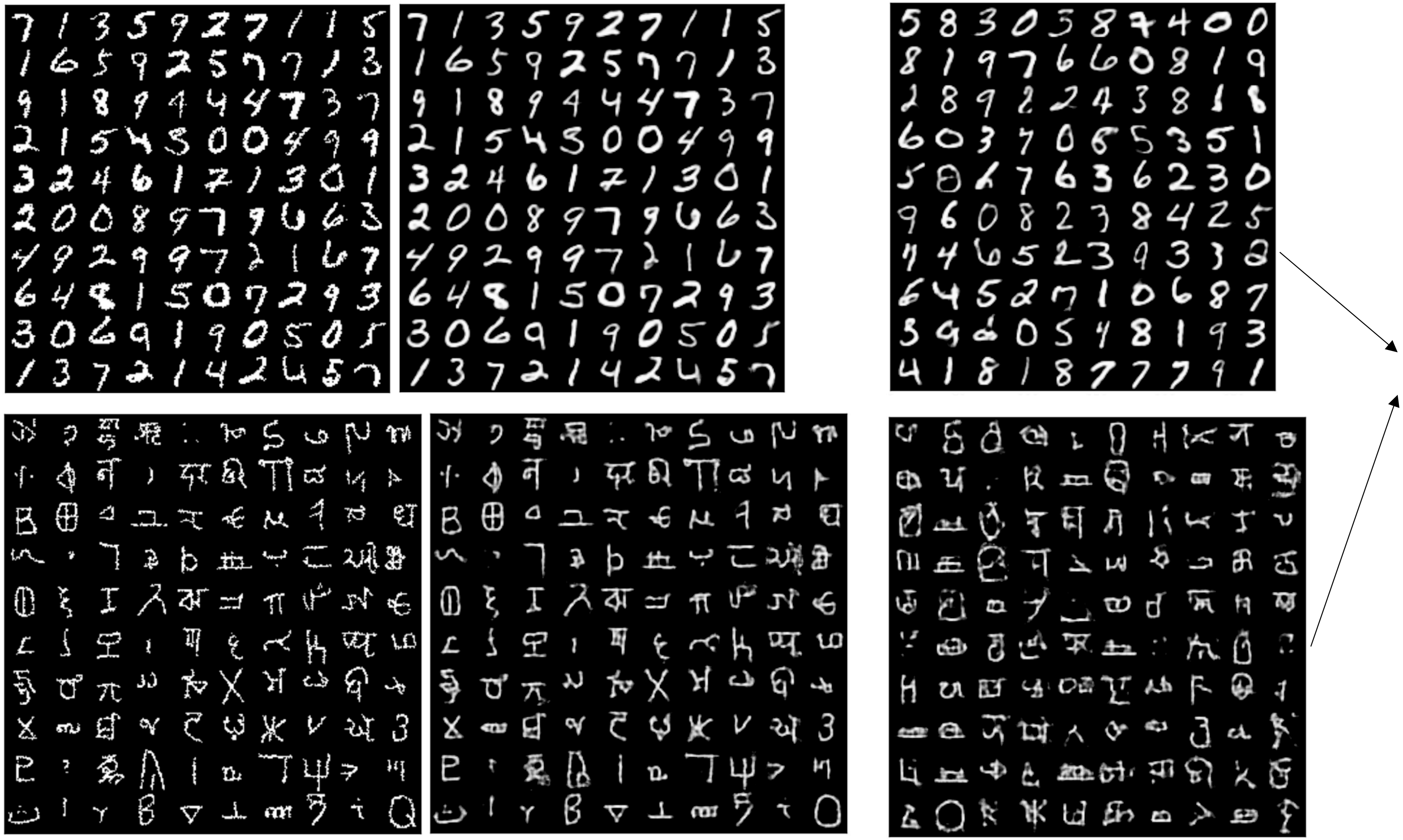
MNIST



OMNIGLOT

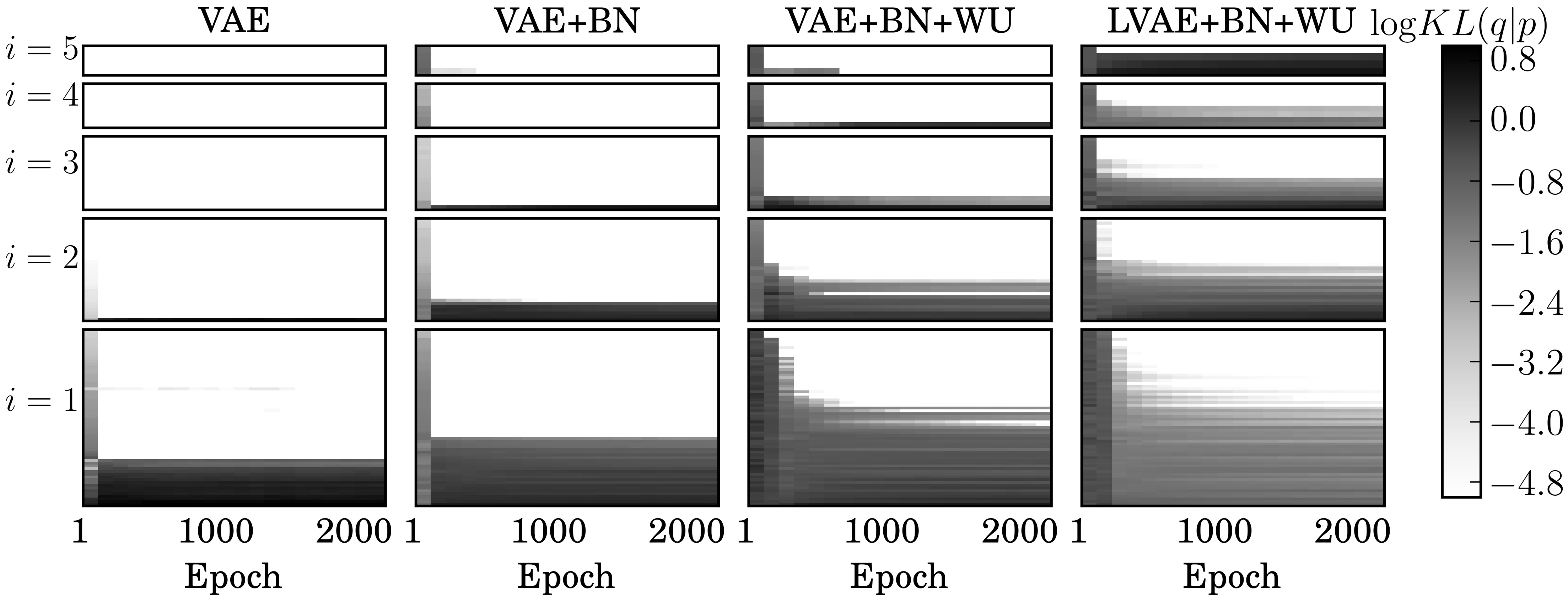


Experiments



Samples from Prior

Experiments: active unit comparison



Experiments: PCA analysis