Learning Manifold Dimensions with Conditional Variational Autoencoders

Yijia Zheng Purdue/AWS

Tong He AWS

Yixuan Qiu SUFE

David Wipf AWS

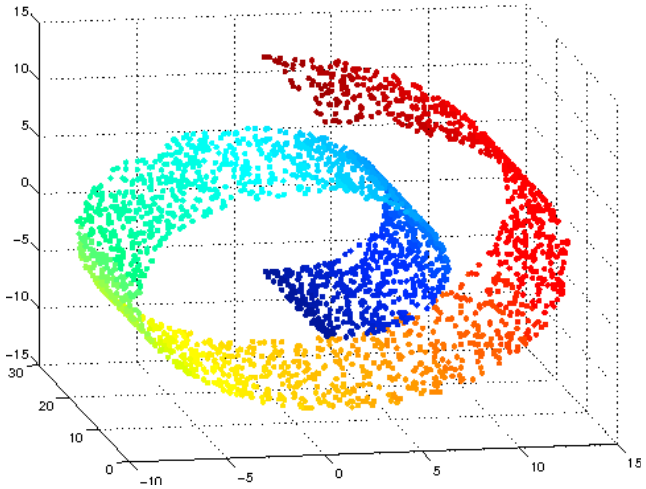
March 2023

Outline

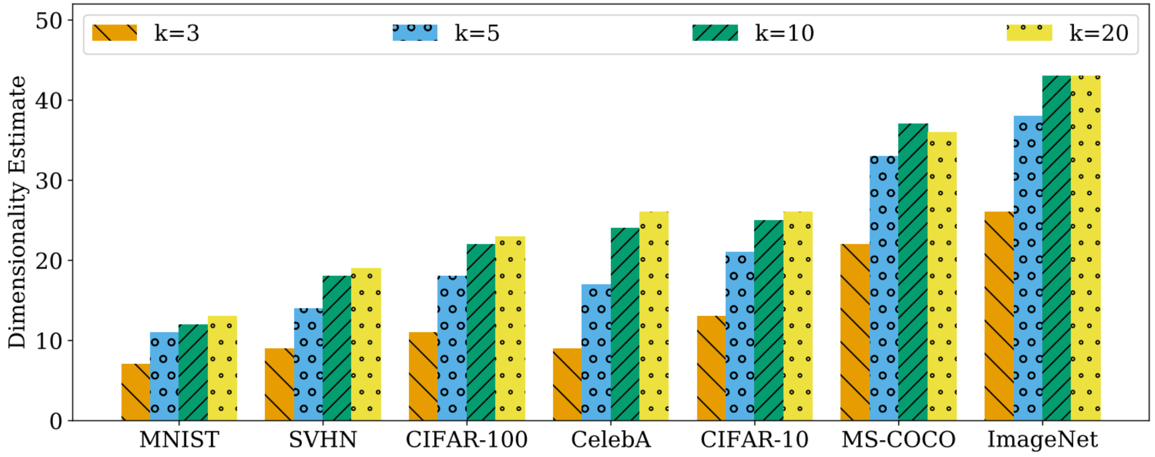
- VAE in learning manifold dimensions
- Extension to CVAE
- Model design diagnoses

Manifold

- **Data** lies on a low-dimensional manifold, which is a mathematical object that can be curved but looks flat locally



A Swiss Roll



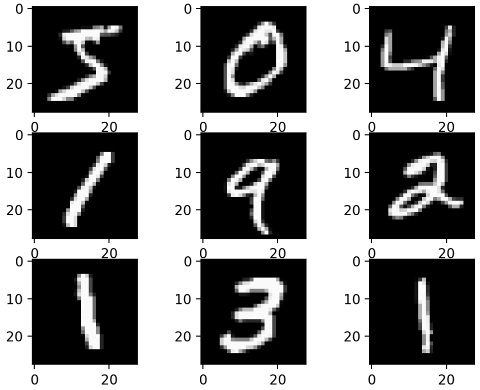
Estimates of the intrinsic dimension of commonly used datasets obtained using the MLE method with k = 3, 5, 10, 20 nearest neighbors[1]

[1] The Intrinsic Dimension of Images and Its Impact on Learning, Pope et al., ICLR 2021

Latent Variable Model[1]

Observed Data: $x \in \mathcal{X} \subseteq \mathbb{R}^d$

Assumed Latent Vector: $z \in \mathcal{Z} \subseteq \mathbb{R}^\kappa$



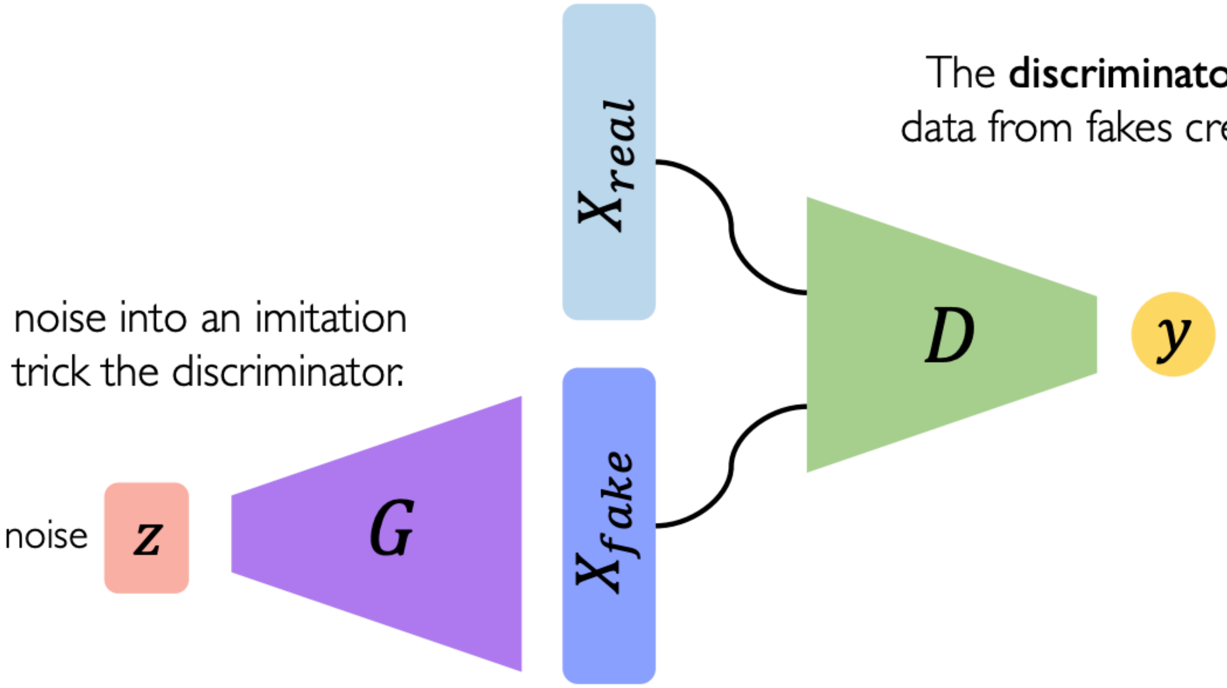
Each sample is 28*28=784 dim

$\kappa < 20 \ll d = 784$ is sufficient

z is a low-dimensional representation of significant factors in x

[1] The images of latent variable models part are borrowed from the slides of MIT 6.S191 and ICASSP 2019 tutorial of David Wipf.

GAN



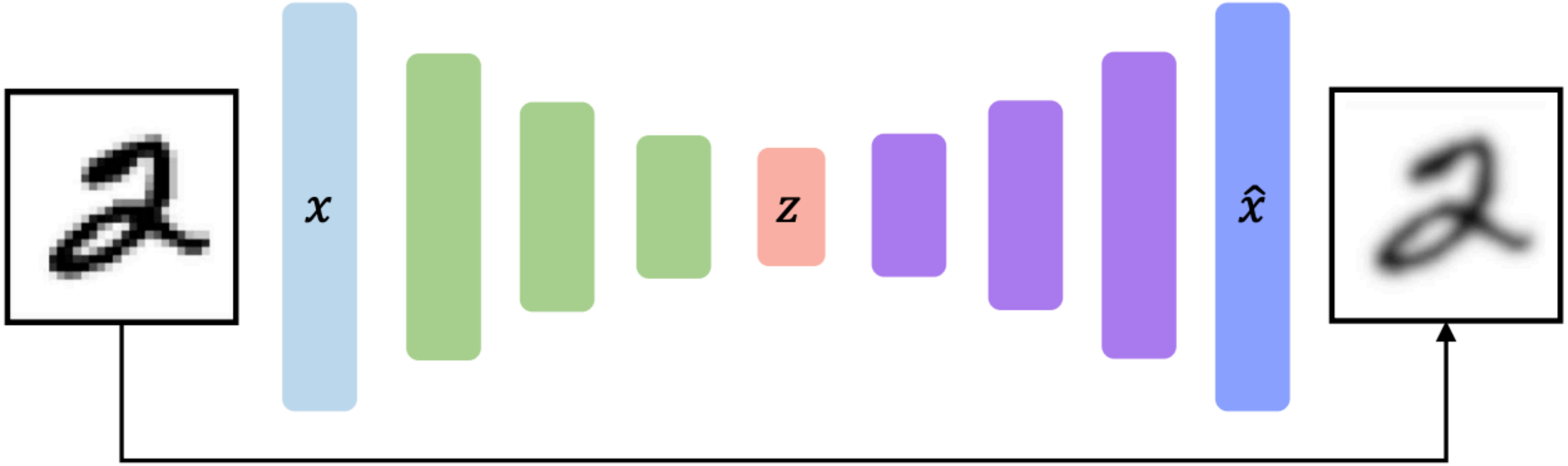
The **discriminator** tries to identify real data from fakes created by the generator.

The **generator** turns noise into an imitation of the data to try to trick the discriminator.

We need to train it via a minimax game:

$$\min_{\theta_g} \max_{\theta_d} [\mathbb{E}_{x \sim p_{data}} \log D(x; \theta_d) + \mathbb{E}_{z \sim p(z)} \log(1 - D(G(z; \theta_g); \theta_d))]$$

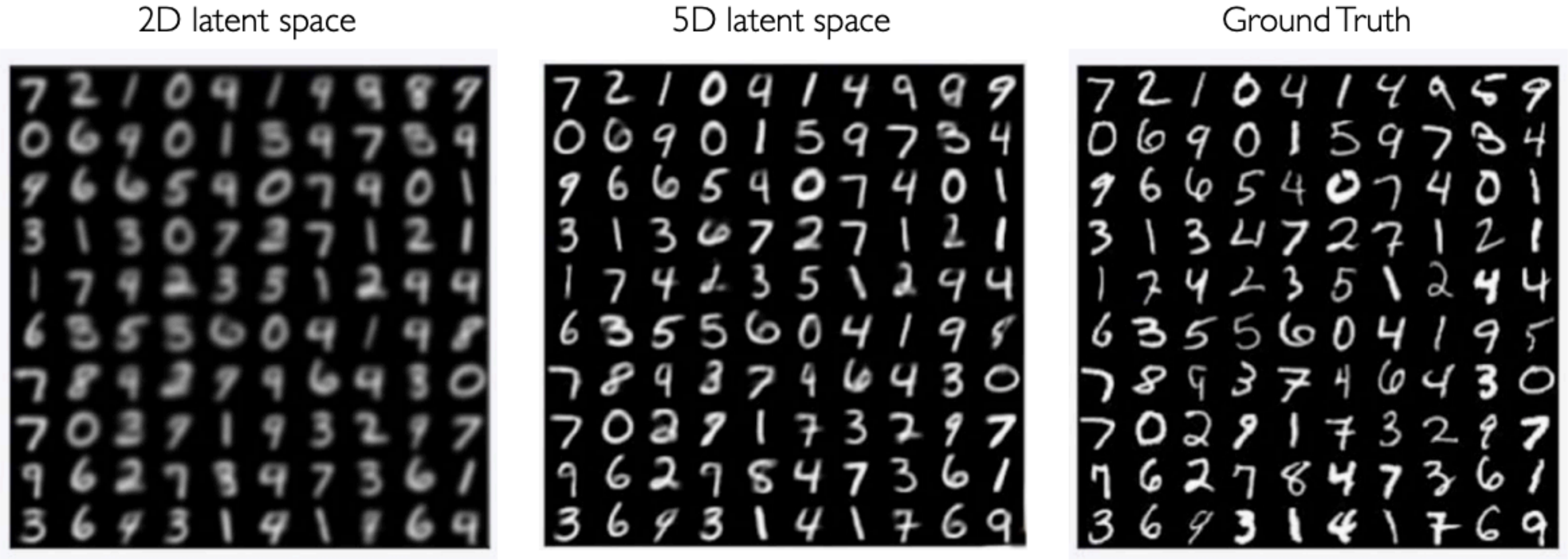
Autoencoders



$$\mathcal{L}(x, \hat{x}) = ||x - \hat{x}||^2$$

No labels are used in the loss!

learning a lower-dimensional feature representation from unlabeled training data



2D latent space

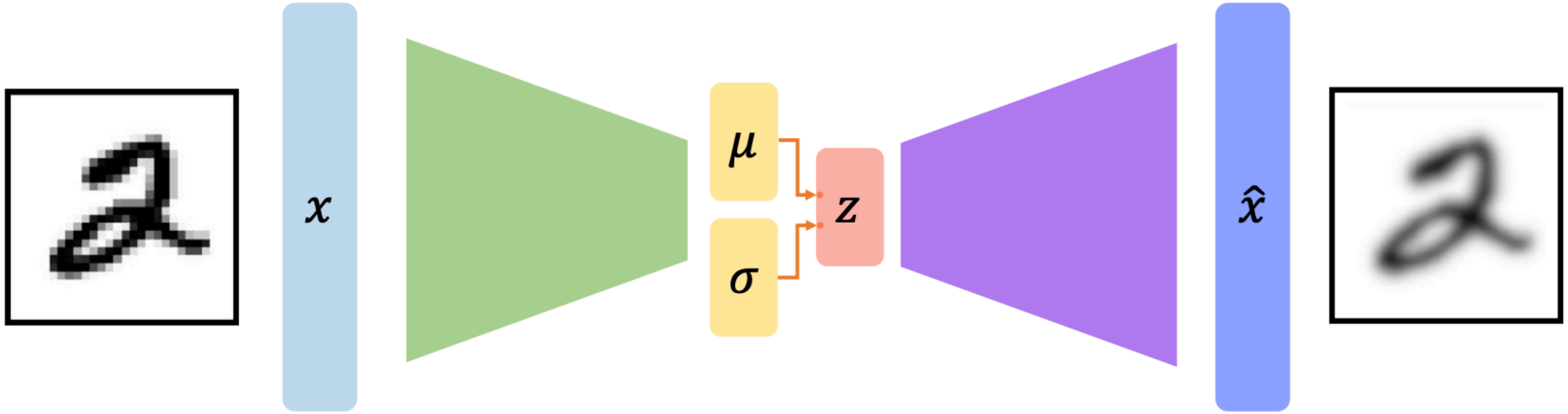
5D latent space

Ground Truth

Dimensions of latent space ⇒ Reconstruction quality

Smaller latent space will force a larger training bottleneck

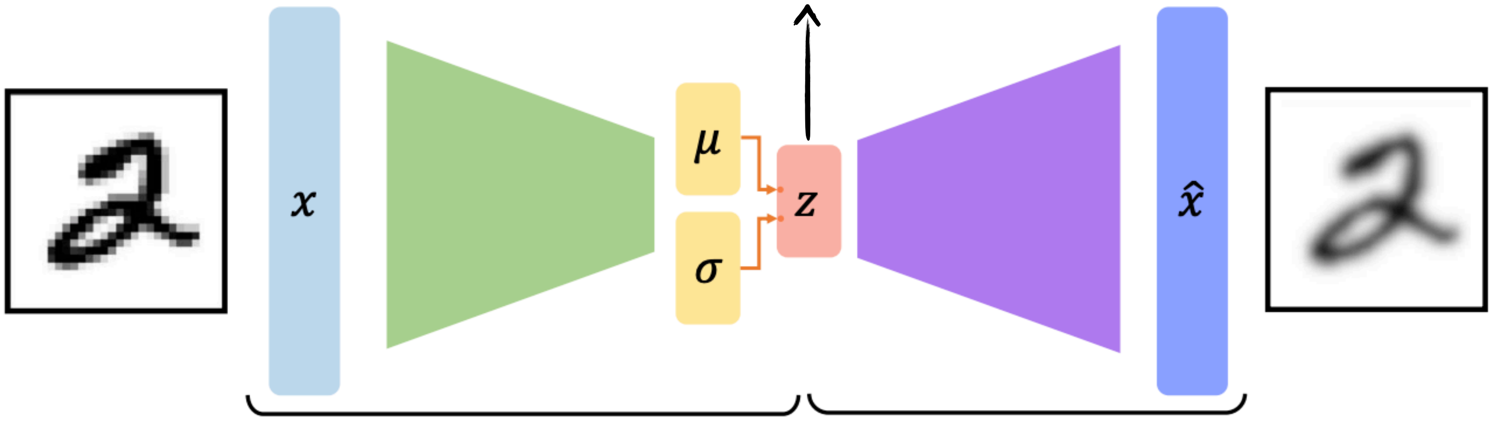
Variational Autoencoder



Variational autoencoders are a probabilistic twist on autoencoders!
Sample from the mean and standard deviation to compute latent sample

Variational Autoencoder

$$z = \mu_z(x) + \sigma_z(x) \cdot \varepsilon, \text{ where } \varepsilon \sim N(0,\ I_\kappa)$$



Encoder: $q_\phi(z \mid x)$

Decoder: $p_\theta(x \mid z)$

Data dimension d

Latent dimension κ

Intrinsic dimension r

Encoder $N(\mu_z(x), \text{diag}[\sigma_z^2(x)]; \phi)$

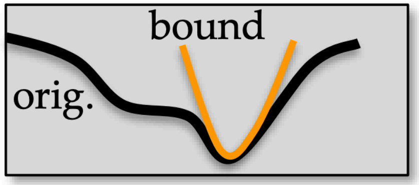
Decoder $N(\mu_x(z), \underset{\substack{\downarrow \\ \text{A scalar}}}{\gamma} I; \theta)$

Loss

Goal Given $x \sim p_{gt}(x)$, solve $\min_{\theta} -\int \log p_{\theta}(x)dx$

A naive approximation Sample $\{z^i\}_{i=1}^m \sim N(0,\ I)$, compute $\int p_{\theta}(x|z)N(0,I)dz \approx \frac{1}{m}\sum_{i=1}^{m} p_{\theta}(x|z^i)$

In this case, for most $z^i \sim N(0,\ I)$, $p_{\theta}(x|z^i) = 0$



Use the variational upper bound $\mathscr{L}$ (ELBO)

$$\mathscr{L}(\theta,\phi) = \int_{\mathscr{X}} \{-\mathbb{E}_{q_{\phi}(z|x)}[\log p_{\theta}(x|z)] + \mathbb{KL}[q_{\phi}(z|x)||\underset{\displaystyle\text{Prior } N(0,\ I_{\kappa})}{\underset{\big\downarrow}{p(z)}}]\}\omega_{gt}(dx)$$

When the decoder variance γ is trainable

γ goes to zero when the VAE model reaches its optimum[2]

We observed there are two behaviors of encoder variance $\sigma_z^2(x)$ in different dimensions:

1. $\sigma_z^2(x) \rightarrow 1$, unnecessary

**Reconstructions as we change latent code along this dimension*

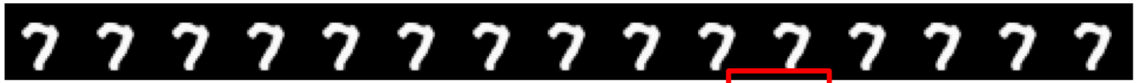


Image Variance = $\boxed{0.000}$ **no changes**

2. $\sigma_z^2(x) \rightarrow 0$, informative

Why would that happen?
How many informative dimensions there are?

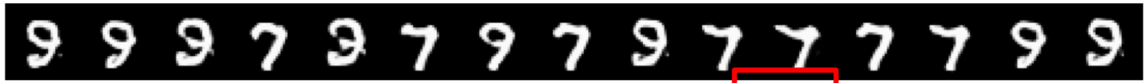
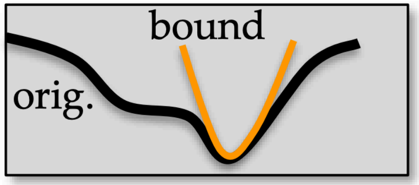


Image Variance = $\boxed{27.20}$ **large changes**

[2] Diagnosing and Enhancing VAE Models, Dai & Wipf, 2019

Loss



$$2\mathscr{L}(\theta,\phi) = 2\int_{\mathscr{X}} \{-\overset{\substack{\text{Reconstruction}\\ \uparrow}}{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]} + \overset{\substack{\text{Regularizer}\\ \uparrow}}{\mathbb{KL}[q_\phi(z|x)||p(z)]}\}\omega_{gt}(dx)$$

Remind that $q_\phi(z|x)$ and $p_\theta(x|z)$ are Gaussian

$$= \int_{\mathscr{X}} \{\log(2\pi\gamma) + \frac{1}{\gamma}\underbrace{\mathbb{E}_{q_\phi(z|x)}(||x-\mu_x(z)||^2)}_{\downarrow} + 2\mathbb{KL}[q_\phi(z|x)||p(z)]\}\omega_{gt}(dx)$$

We want $\mu_x(z)$ to reconstruct x. This expectation will go to zero.

γ will also go to zero.

In our paper, γ is a trainable scalar!

We want $||x-\mu_x(z)||^2 \to 0$ at a higher rate than $\gamma \to 0$. Otherwise $\mathscr{L}$ will go infinity.

How about the KL term?

KL term

$$\int_{\mathcal{X}} \{\log(2\pi\gamma) + \frac{1}{\gamma}\mathbb{E}_{q_\phi(z|x)}(||x-\mu_x(z)||^2) + \underbrace{2\mathbb{KL}[q_\phi(z|x)||p(z)]}_{\big\downarrow}\}\omega_{gt}(dx)$$

$$= \mu_z(x)'\mu_z(x) + tr(\sigma_z^2(x)) - \kappa - \log(|\sigma_z^2(x)|)$$

$$= -\log(|\sigma_z^2(x)|) + O(1)$$

Leave out the terms which will not get
unbounded values when $\gamma \to 0$,
since $\lim_{\gamma\to 0}\sigma_z^2(x)$ for informative dimensions

To perfectly reconstruct x which is a r-dimensional manifold, we need ***r* dimensions of information**.
We assume the first r dimensions of z are used for the decoder to do reconstruction.

Reconstruction Term

Assume the mean function $\mu_z(x;\phi)$ is L-Lipschitz continuous, we can get an upper bound of the norm

$\mathbb{E}_{z\sim q_{\phi_\gamma}(z|x)}[||x-\mu_x(z)||^2] = \mathbb{E}_{z\sim q_{\phi_\gamma}(z|x)}[||x-\mu_x(z)_{1:r}||^2] \leq \mathbb{E}_{\varepsilon\sim N(0,I)}[||L\sigma_z(x)_{1:r}\varepsilon||^2]$, where $\varepsilon \sim N(0,I)$

The upper bound of $\mathscr{L}$:

$$\tilde{\mathscr{L}} = \int_{\mathscr{X}} \{\log(2\pi\gamma) + \frac{1}{\gamma}\mathbb{E}_{\varepsilon\sim N(0,I)}[||L\sigma_z(x)_{1:r}\varepsilon||^2] - \log(|\sigma_z^2(x)_{1:r}|) - \log(|\sigma_z^2(x)_{r+1:\kappa}|) + O(1)\}\omega_{gt}(dx)$$

By taking the derivatives of $\sigma_z^2(x)$ and γ respectively, a relation shows

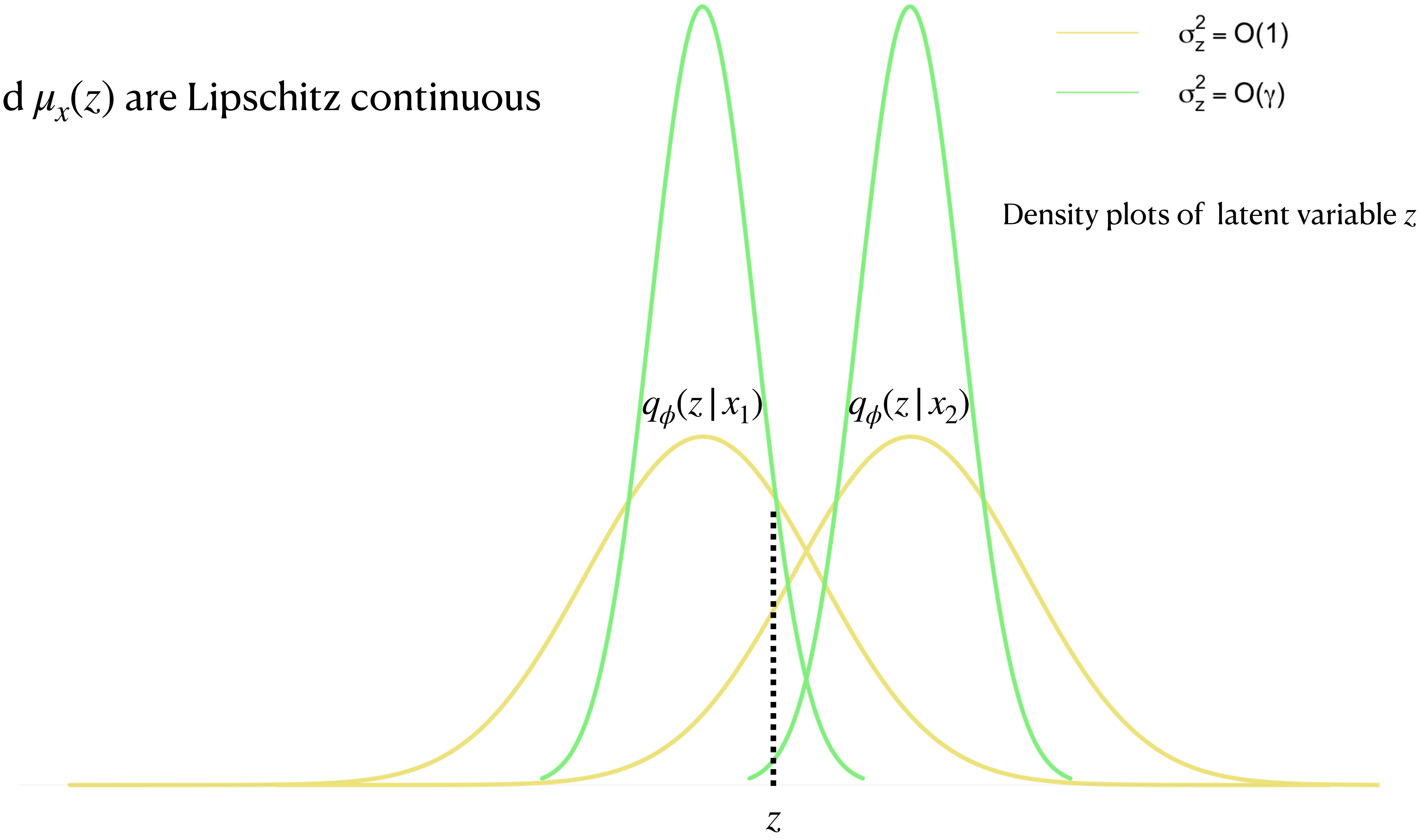
$$\sigma_z^*(x)^2_{1:r} = \gamma\frac{I}{L^2}$$

γ goes to zero when the VAE model reaches its optimum

At least r dimensions of $\sigma_z^2(x)$ goes to zero at optimum

Intuitively: why $\sigma_z^2(x)$ should be small for r dimensions?

$\mu_z(x)$ and $\mu_x(z)$ are Lipschitz continuous



KL term

Assume we have $\hat{r}$ dimensions of $\sigma_z^2(x)$ goes to zero with γ, i.e. $\sigma_z^2(x)_{1:\hat{r}} = O(\gamma)$, where $\hat{r} \geq r$

$$\mathbb{KL}(q_\phi(z|x)||p(z)) = -\log(|\sigma_z^2(x)_{1:r}|) - \log(|\sigma_z^2(x)_{r+1:\hat{r}}|) \; -\log(|\sigma_z^2(x)_{\hat{r}+1:\kappa}|) + O(1)$$

If we do not constrain $\sigma_z^2(x)_{r+1:\hat{r}}$, these dimensions will try to match the prior's variance, i.e. 1

To minimize $\mathscr{L}(\phi, \theta)$, $\hat{r} = r$ when model converges

Remind that $\sigma_z^*(x)_{1:r}^2 = \gamma \dfrac{I}{L^2}$, we have the final form of the KL term is $-r\log(\gamma) + O(1)$

Loss (continued)

$$\underbrace{-\mathbb{E}_{z\sim q_\phi(z|x)}[\log p_\theta(x\,|\,z)]}_{\downarrow\; = \frac{1}{2\gamma}\mathbb{E}_{q_\phi(z|x)}||x-\mu_x(z)||^2 + \frac{1}{2}d\log(2\pi\gamma)} + \underbrace{\mathbb{KL}(q_\phi(z\,|\,x)\,||\,p(z))}_{\searrow\; = -\frac{1}{2}r\log(\gamma) + O(1)} \longrightarrow$$

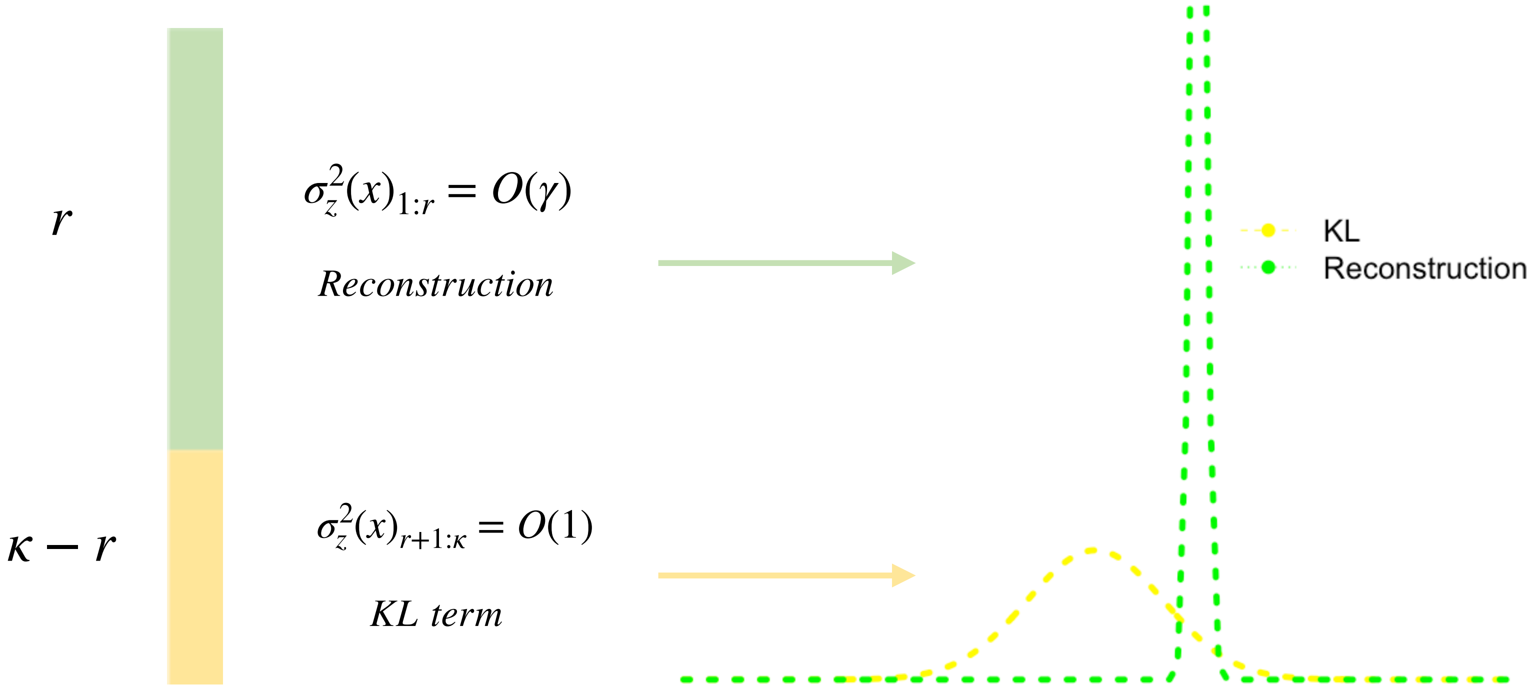
An additional coefficient β
is added for KL term in β-VAE

$$= (d-r)\log(\gamma) + O(1)$$

Active Dimensions

The dimensions of $\sigma_z^2(x)$ that are used for reconstruction.

Such $\sigma_z^2(x)$ will go to **zero** when the model reaches its optimality!



r

$\sigma_z^2(x)_{1:r} = O(\gamma)$

Reconstruction

$\kappa - r$

$\sigma_z^2(x)_{r+1:\kappa} = O(1)$

KL term

Results of VAE models

κ	d	r	AD	Recon	$\mathbb{KL}$	γ	-ELBO
20	10	2	2	3×10^{-4}	18.31	1.625×10^{-5}	-58.26
		4	4	2.6×10^{-3}	24.22	5.654×10^{-5}	-29.83
		6	6	9.2×10^{-3}	24.14	3×10^{-4}	-17.39
		8	7	1.27×10^{-2}	27.91	1.4×10^{-3}	-10.38
		10	8	5.99×10^{-2}	16.39	2.5×10^{-3}	-6.40
	20	2	2	1.6×10^{-3}	17.98	5.052×10^{-5}	-114.52
		4	4	1.75×10^{-2}	23.11	2×10^{-4}	-60.90
		6	6	3.09×10^{-2}	28.96	6×10^{-4}	-43.75
		8	8	3.42×10^{-2}	33.83	1.2×10^{-3}	-36.82
		10	10	4.74×10^{-2}	35.81	1.1×10^{-3}	-28.34
	30	2	2	2.6×10^{-3}	18.42	7.221×10^{-5}	-176.74
		4	4	2.73×10^{-2}	24.60	2×10^{-4}	-100.28
		6	6	4.74×10^{-2}	31.89	9×10^{-4}	-76.46
		8	8	5.68×10^{-2}	37.28	1.6×10^{-3}	-65.66
		10	10	1.13×10^{-1}	35.13	2.5×10^{-3}	-47.00
5	20	6	5	1.299×10^{-1}	22.53	2.1×10^{-3}	-36.97
		8	5	3.719×10^{-1}	16.618	8.8×10^{-3}	-22.60
		10	5	3.564×10^{-1}	15.966	1.113×10^{-2}	-16.96

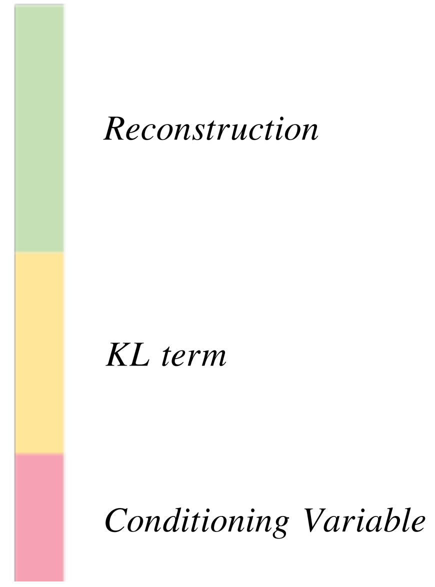
0.0080	0.0018	1.0000	1.0000	1.0000
0.0027	0.0031	1.0000	1.0000	1.0000
1.0000	0.0087	1.0000	0.0141	1.0000
1.0000	1.0000	1.0000	1.0000	1.0000

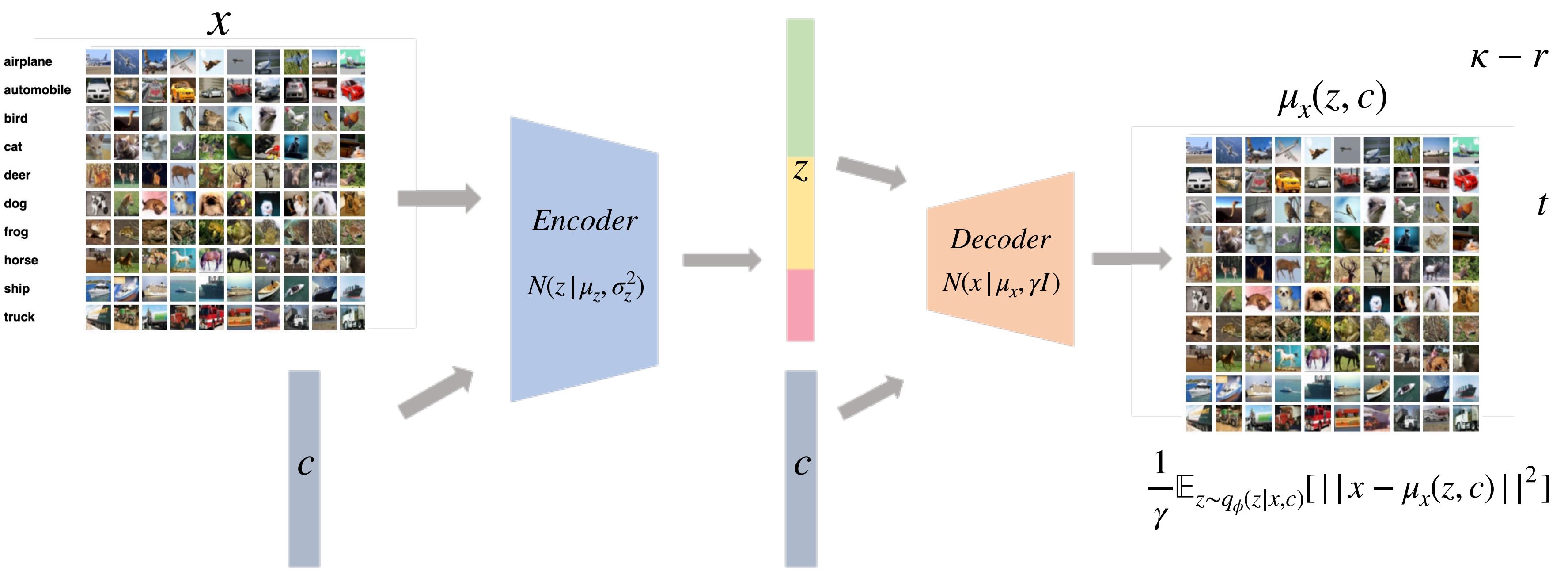
Visual of $\sigma_z^2(x)$ with $\kappa = 20,\ d = 30,\ r = 6$

Extend to Conditional VAE

Add a **conditioning variable** c with t effective dimensions

Such c can help to reconstruct at most t dimensions





$$\frac{1}{\gamma}\mathbb{E}_{z \sim q_\phi(z|x,c)}[||x-\mu_x(z,c)||^2]$$

$$\mathbb{KL}(q_\phi(z|x,c)||p(z|c))$$

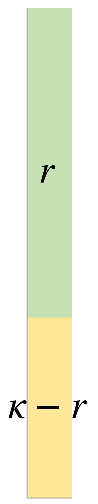
$$q_\phi(z|x,c) = N(\mu_z(x,c;\phi), \sigma_z^2(x,c;\phi))$$

How does the CVAE model use *c*?

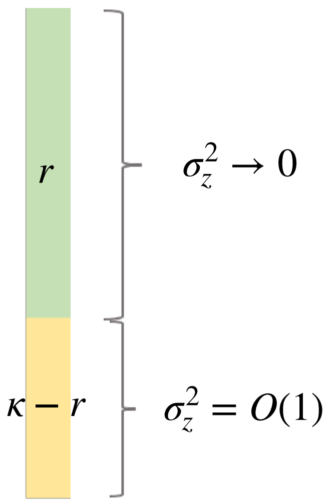
$$\underbrace{-\mathbb{E}_{z\sim q_\phi(z|x,c)}[\log p_\theta(x|z,c)]}_{\downarrow} + \underbrace{\mathbb{KL}(q_\phi(z|x,c)||p(z|c))}_{\downarrow}$$

$$= \frac{1}{2}d\log(2\pi\gamma) + \frac{1}{2\gamma}\mathbb{E}_{q_\phi(z|x,c)}||x-\mu_x(z,c)||^2 \qquad ? = -\frac{1}{2}(r-t)\log\gamma + constant$$

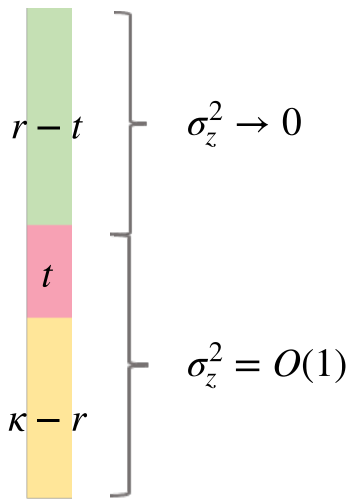
If *c* only shows in the encoder



If *c* only shows in the prior



If *c* only shows in the decoder



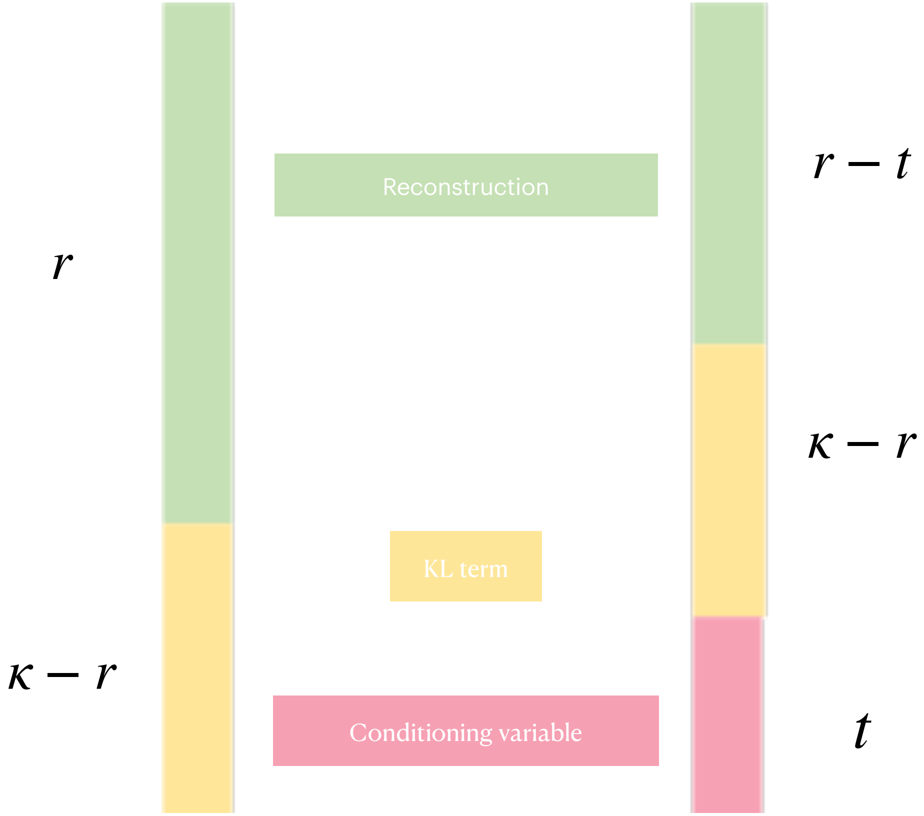
The encoder and prior will not use *c* when the model reaches its optimum

How about optimal loss?

VAE $(d - r)\log\gamma + O(1)$

CVAE $(d - r + t)\log\gamma + O(1)$

κ is not in the loss formula because the “redundant” part can be cancelled by matching prior!



Experiment results

t	-ELBO	Recon	$\mathbb{KL}$	γ	AD
1	-31.41	4.61×10^{-2}	33.26	2.4×10^{-3}	9
3	-36.67	4.66×10^{-2}	27.78	2.4×10^{-3}	7
5	-42.78	4.86×10^{-2}	20.81	2.6×10^{-3}	5
7	-52.39	4.29×10^{-2}	13.72	2.2×10^{-3}	3
9	-62.25	3.84×10^{-2}	6.07	2×10^{-3}	1

$r = 10$

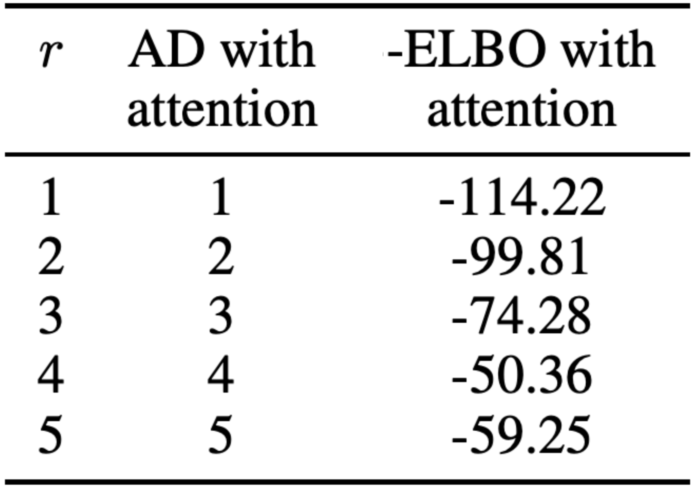
3.6159e-03	9.6320e-01	7.6566e-04	3.5173e-04
9.8518e-01	9.6739e-01	9.6077e-01	8.1020e-04
9.8065e-01	9.7336e-01	3.7781e-03	7.1394e-04
9.6985e-01	6.1294e-03	9.7449e-01	9.8012e-01
7.8233e-04	9.7318e-01	9.8596e-01	2.4359e-04
9.7785e-01	9.7737e-01	9.7315e-01	9.8431e-01
9.2616e-01	9.8335e-01	9.6775e-01	1.2756e-03
1.0324e-03	9.6723e-01	9.6046e-01	2.1289e-03

$\sigma_z^2(x, c)$ on MNIST dataset. $\kappa = 32$ and the number of active dimensions is 12

When data lies on a union of manifolds

Each manifold is with a locally-defined value of r

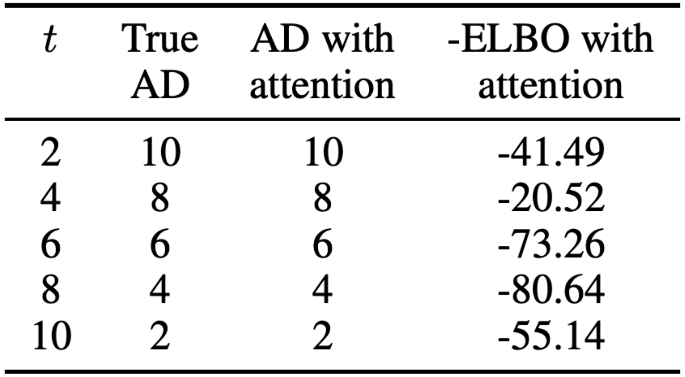
Case 1: c is a discrete variable indicating different manifolds, then the manifold dimension itself may vary conditioned on the value of c in a single model.



r	AD with attention	-ELBO with attention
1	1	-114.22
2	2	-99.81
3	3	-74.28
4	4	-50.36
5	5	-59.25

A union of 5 manifolds with
$r = \{1,2,3,4,5\}$, $d = 20$, $\kappa = 40$.
A discrete c labels indicates each manifold/class

Case 2: c is a continuous variable. t varies for different values of c, i.e. different value can help to reconstruct t dimensions of the manifold.



t	True AD	AD with attention	-ELBO with attention
2	10	10	-41.49
4	8	8	-20.52
6	6	6	-73.26
8	4	4	-80.64
10	2	2	-55.14

A continuous c associated with $t \in \{2,4,6,8,10\}$

$r = 12,\ d = 20,\ \kappa = 90.$

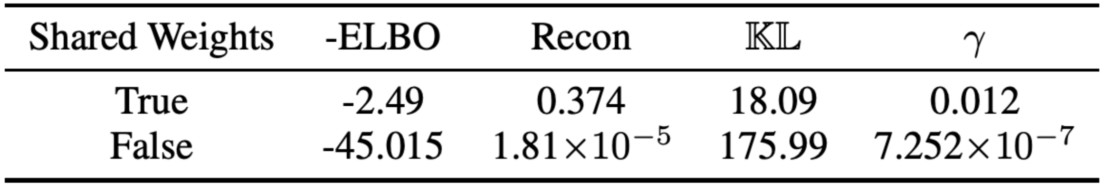
Some diagnoses of CVAE models

1. Encoder/prior model weights sharing in sequence models

$$p_\theta(z_l|x_{<l}) = q_\phi(z_l|x_{<l}) = q_\phi(z_l|x_{\le l-1})$$

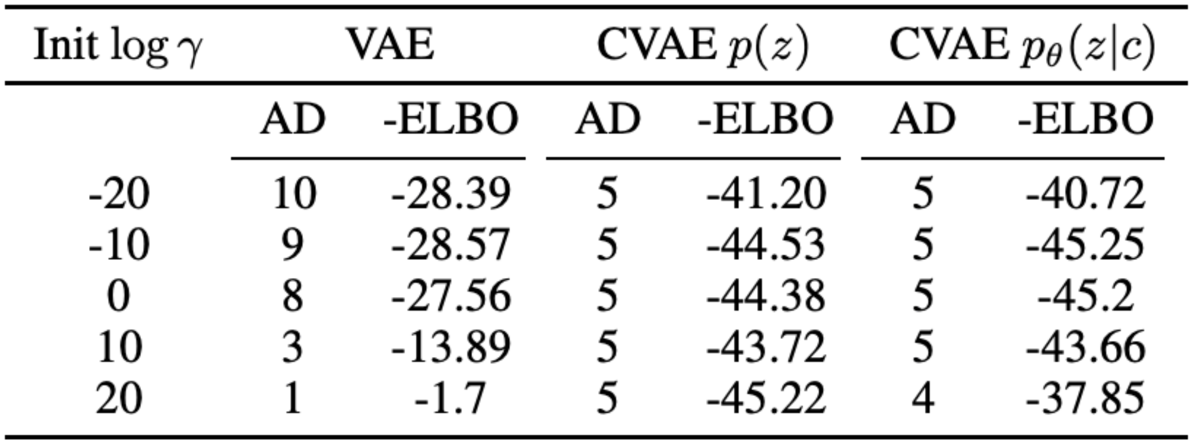
↓ The encoder at time $l-1$

↓ The prior at time l



Shared Weights	-ELBO	Recon	$\mathbb{KL}$	γ
True	-2.49	0.374	18.09	0.012
False	-45.015	1.81×10^{-5}	175.99	7.252×10^{-7}

2. Initial γ is significant to model convergence



Init $\log \gamma$	VAE		CVAE $p(z)$		CVAE $p_\theta(z\|c)$	
	AD	-ELBO	AD	-ELBO	AD	-ELBO
-20	10	-28.39	5	-41.20	5	-40.72
-10	9	-28.57	5	-44.53	5	-45.25
0	8	-27.56	5	-44.38	5	-45.2
10	3	-13.89	5	-43.72	5	-43.66
20	1	-1.7	5	-45.22	4	-37.85

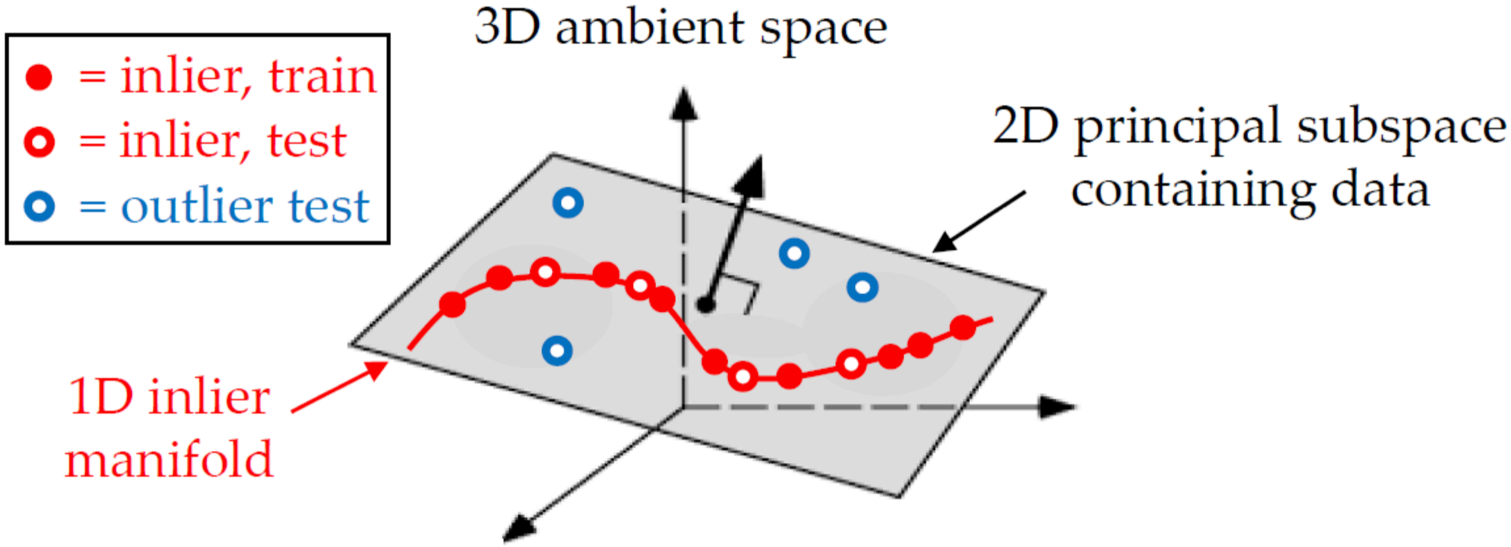
$d = 20,\ r = 10,\ t = 5,\ \kappa = 20$

3. Equivalence of conditional and unconditional priors

Prior: $p(z\,|\,c) = N(\mu_p(c), \sigma_p^2(c))$ $\longrightarrow$ $p'(z) = N(0,I)$

Decoder: $p_\theta(x\,|\,z, c)$ $\longrightarrow$ $p'(x\,|\,z', c) = p_\theta(x\,|\,z' * \sigma_p(c) + \mu_p(c), c)$

Application: outlier screening



Some take-home messages

- A trainable γ as decoder variance is preferred
- At global optimality, the encoder variance has some dimensions goes to zero. These dimensions show the number of manifold dimensions.
- Given a trainable γ, a near zero KL term is not a signal for good convergence
- Conditional VAE can learn a union of manifold dimensions
- A good initial γ can help the start of the training process
- Weight sharing between the prior and posterior compromises performance of sequential modeling
- A conditioned prior is not necessary