Student-t Variational Autoencoder for Robust Density Estimation

Hiroshi Takahashi[1], **Tomoharu Iwata**[2], **Yuki Yamanaka**[3], **Masanori Yamada**[3], **Satoshi Yagi**[1]

[1]NTT Software Innovation Center
[2]NTT Communication Science Laboratories
[3]NTT Secure Platform Laboratories

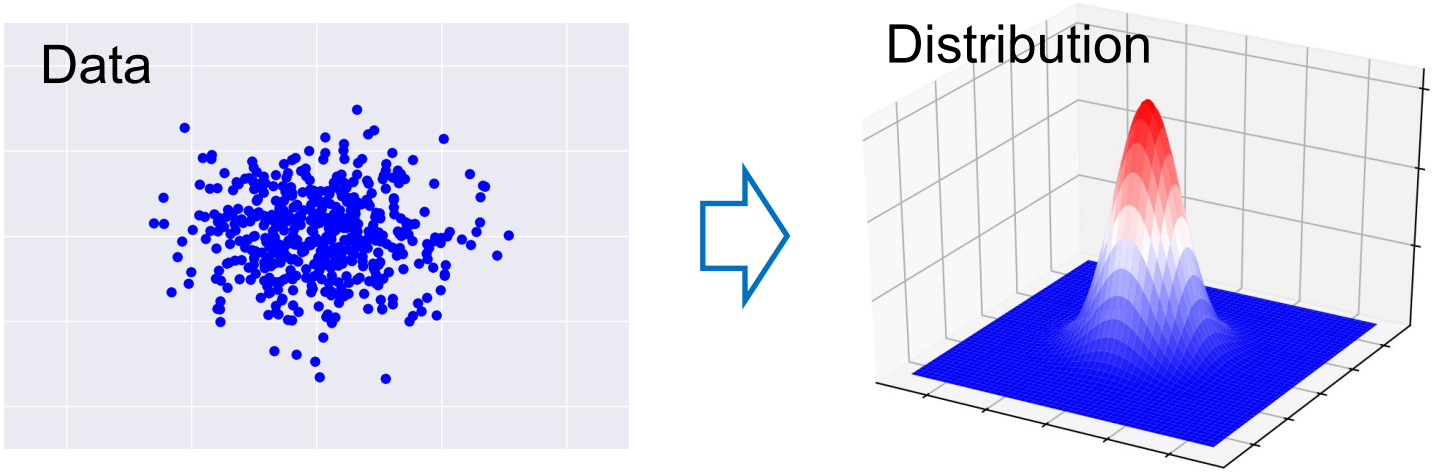
In short

If you use the VAE for continuous data,

we recommend using the Student-t distribution as the decoder!

[Introduction]
Multivariate density estimation

- Estimating data distributions is important for AI
  - especially for image, audio, video, and detection tasks



- The VAE is widely used since it can learn the high-dimensional complicated distributions in these tasks

- We focus on estimating distributions of **continuous data** with the VAE

[Preliminary]
Variational Autoencoders (VAE) (1/3)

Innovative R&D by NTT

- The VAE estimates the probability of a continuous data point x by using latent variable $\mathbf{z}$:

$$p_\theta(\mathbf{x}) = \int \underbrace{p_\theta(\mathbf{x} \mid \mathbf{z})}_{\text{decoder}} \underbrace{p(\mathbf{z})}_{\text{prior}} \mathrm{d}\mathbf{z}$$

- The log marginal likelihood of VAE is bounded below by the evidence lower bound (ELBO):

$$\ln p_\theta(\mathbf{x}) \geq \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}}_{\text{encoder}} [\ln p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{KL}(\underbrace{q_\phi(\mathbf{z} \mid \mathbf{x})}_{\text{encoder}} \| p(\mathbf{z}))$$

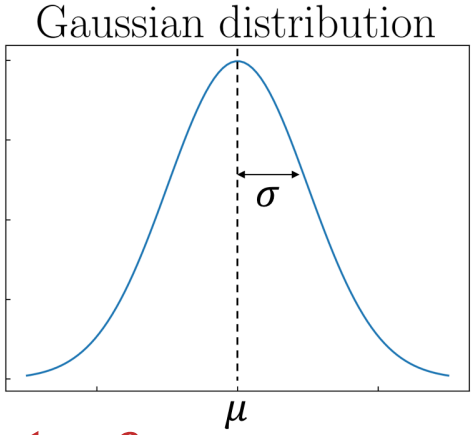
- The VAE is trained to maximize the sum of ELBO

NTT

[Preliminary]
Variational Autoencoders (VAE) (2/3)

- For continuous data, the encoder, decoder, and prior distributions are usually modeled by a Gaussian:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$$: standard Gaussian



$$p_\theta(\mathbf{x}|\mathbf{z}) = \mathcal{N}\left(\mathbf{x} \mid \mu_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})\right)$$

estimated by neural networks with parameter θ

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}\left(\mathbf{z} \mid \mu_\phi(\mathbf{x}), \sigma^2_\phi(\mathbf{x})\right)$$

estimated by neural networks with parameter ϕ

[Preliminary]
Variational Autoencoders (VAE) (3/3)

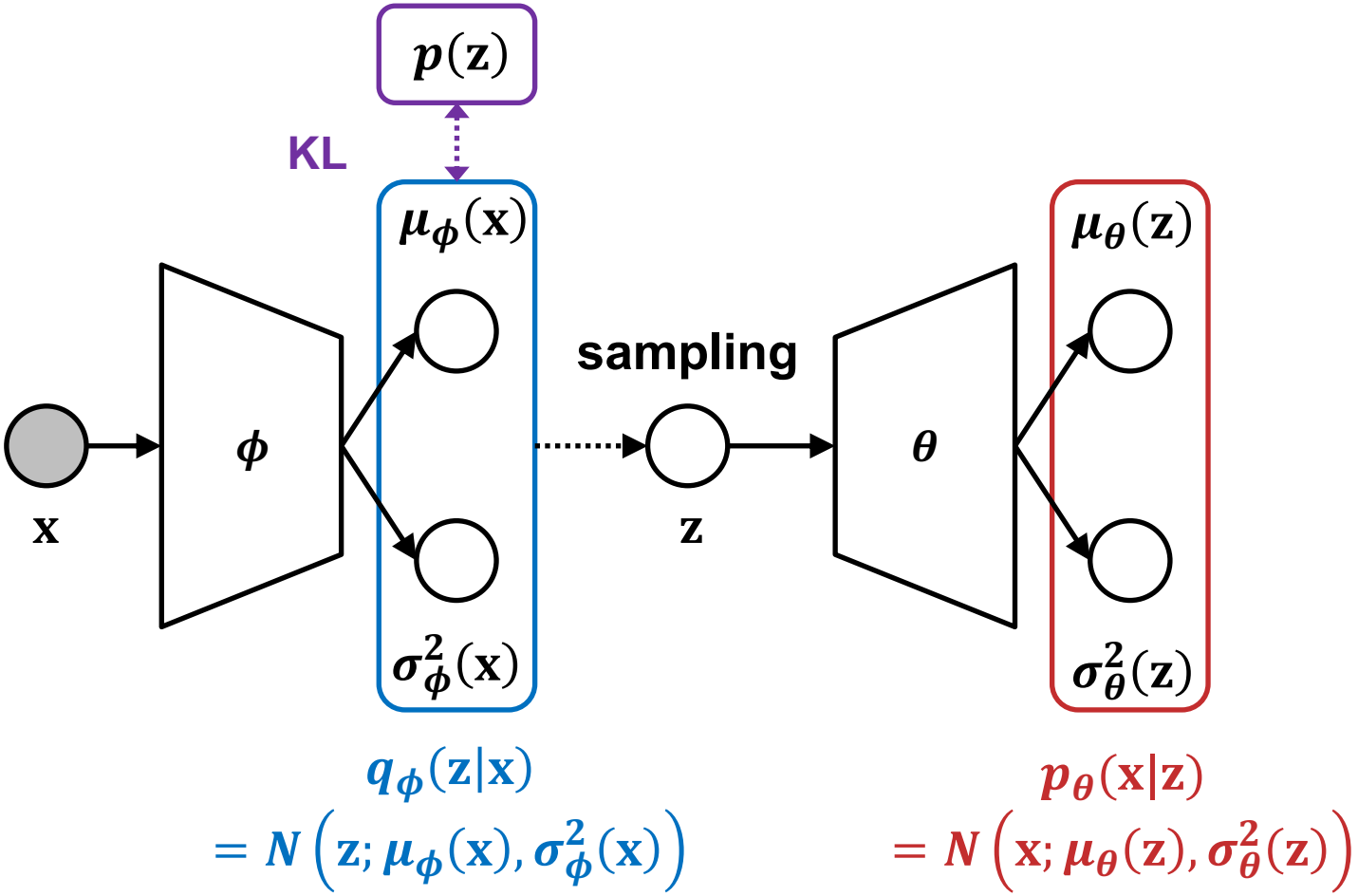
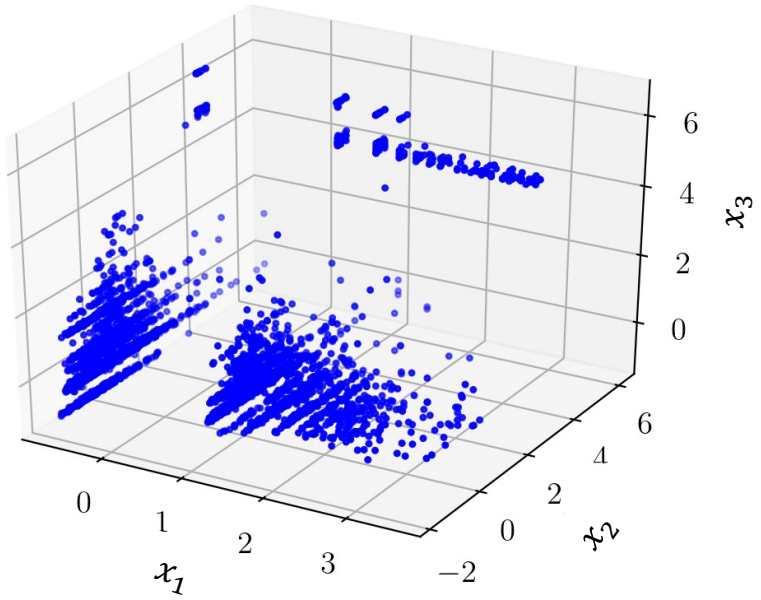


Diagram of VAE for continuous data

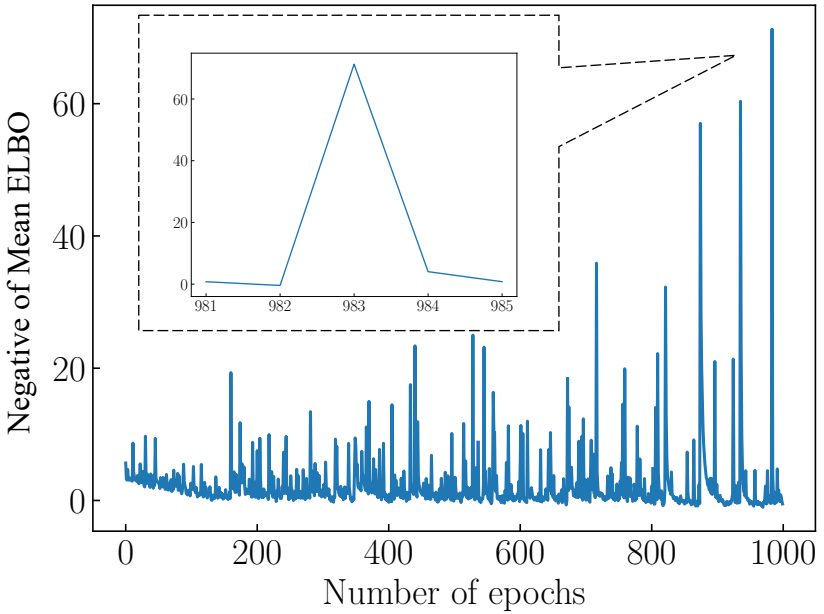
[Our problem]
Instability of training VAE with Gaussian decoder

- When we use the Gaussian as the decoder, the training of VAE often becomes unstable
  - For example, when we train KDD99 SMTP with VAE, Negative of Mean ELBO sharply jumped up during training



KDD99 SMTP Dataset

$$\mathbf{x} = (x_1, x_2, x_3)^T$$



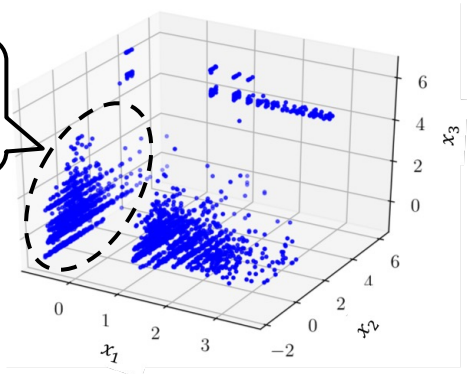
Negative of Mean ELBO

[Our problem]

Cause of this instability: too small variance

- The cause is division by too small variance in ELBO

the variance of these data points is too small along x_1 direction



First term of ELBO

$$\ln p_\theta(\mathbf{x} \mid \mathbf{z}) = \ln \mathcal{N}\left(\mathbf{x} \mid \mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})\right)$$

$$= \sum_d \left[-\frac{\left(\mathbf{x}_d - \mu_{\theta,d}(\mathbf{z})\right)^2}{2\sigma_{\theta,d}^2(\mathbf{z})} - \frac{1}{2} \ln 2\pi\sigma_{\theta,d}^2(\mathbf{z}) \right]$$

d : dimension index

When the decoded varianve $\sigma_\theta^2(\mathbf{z})$ is almost zero, this term is sensitive to the error between $\mathbf{x}$ and its decoded mean $\mu_\theta(\mathbf{z})$

[Idea]
Introducing a prior for the variance

- We can avoid this instability problem by preventing the decoded variance $\sigma_\theta^2(\mathbf{z})$ from being too small

- To penalize small variance, we introduce a Gamma distribution as the prior for the decoded variance $\sigma_\theta^2(\mathbf{z})$

$$\mathrm{Gam}\left(\tau \mid a, b\right) = \frac{b^a \tau^{a-1} \exp\left(-b\tau\right)}{\Gamma\left(a\right)}$$

τ : the inverse of the variance $(1/\sigma^2)$

[Proposed method]

As a simple way: MAP estimation

- First, we present the MAP estimation for the VAE
  - To simplify the calculation, we use $\mathrm{Gam}(\tau|1, b)$ as the prior
  - Then, the objective function of MAP estimation is:

$$\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\ln p_\theta(\mathbf{x}\mid\mathbf{z}) - \frac{b}{\sigma_\theta^2(\mathbf{z})}\right] - D_{KL}\left(q_\phi(\mathbf{z}\mid\mathbf{x})\|p(\mathbf{z})\right)$$

small variance is penalized with regularization parameter b

- However, there are two drawbacks in MAP estimation
  1. Tuning b is difficult
  2. The constant b lacks flexibility in density estimation
     - b should depend on a data point

[Proposed method]
Student-t decoder

- We propose a more flexible approach by introducing a Gamma prior that depends on latent variables:

$$\mathrm{Gam}\left(\tau \mid a\left(\mathbf{z}\right), b\left(\mathbf{z}\right)\right)$$

- By analytically integrating out the τ, we can obtain a Student-t decoder:

$$p_\theta\left(\mathbf{x} \mid \mathbf{z}\right) = \int_0^\infty \mathcal{N}\left(\mathbf{x} \mid \mu_\theta\left(\mathbf{z}\right), \tau^{-1}\right) \mathrm{Gam}\left(\tau \mid a\left(\mathbf{z}\right), b\left(\mathbf{z}\right)\right) \mathrm{d}\tau$$

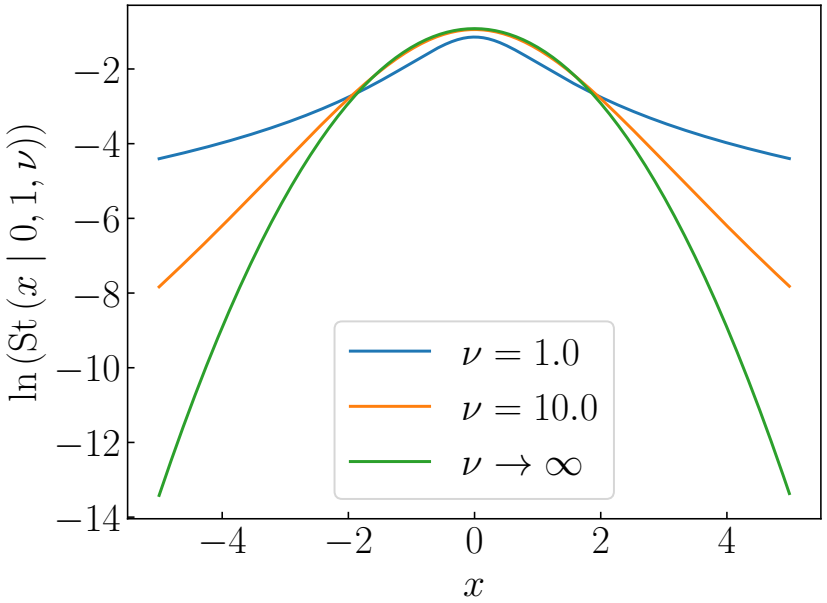
$$= \mathrm{St}\left(\mathbf{x} \mid \mu_\theta\left(\mathbf{z}\right), \lambda_\theta\left(\mathbf{z}\right), \nu_\theta\left(\mathbf{z}\right)\right)$$

where

$$\lambda_\theta\left(\mathbf{z}\right) = a(\mathbf{z})/b\left(\mathbf{z}\right), \nu_\theta\left(\mathbf{z}\right) = 2a\left(\mathbf{z}\right)$$

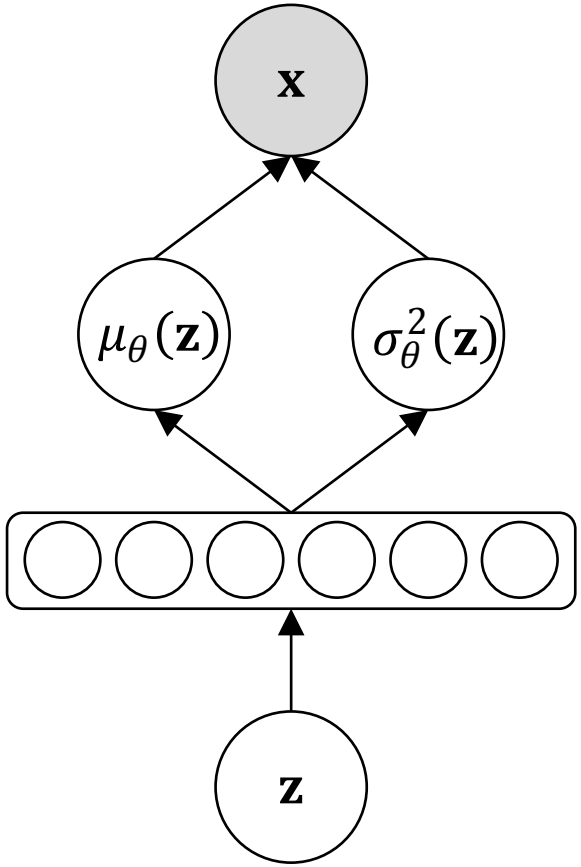
[Proposed method]
Robustness of Student-t decoder

- Since the Student-t distribution is **heavy-tailed** (has large variance), the Student-t decoder is **robust** to the error between the data point and its decoded mean
  - The appropriate robustness is set by $\nu_\theta(\mathbf{z})$, which makes the training of VAE stable!

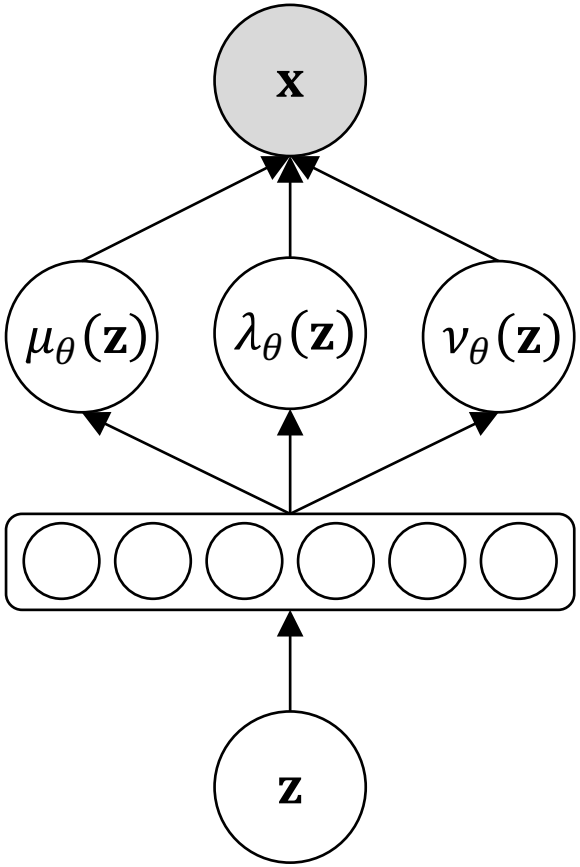


Plot of $\mathrm{St}(x|0,1,\nu)$ in log scale

[Proposed method]
Diagram of Student-t decoder



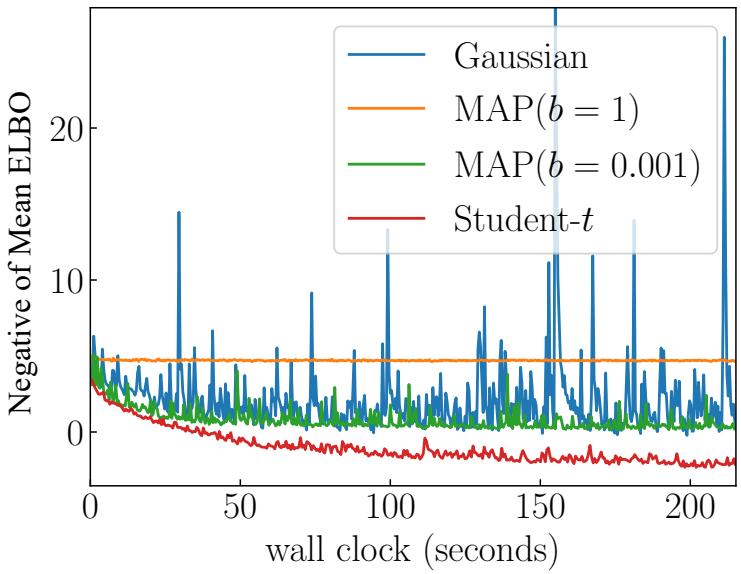
Gaussian decoder



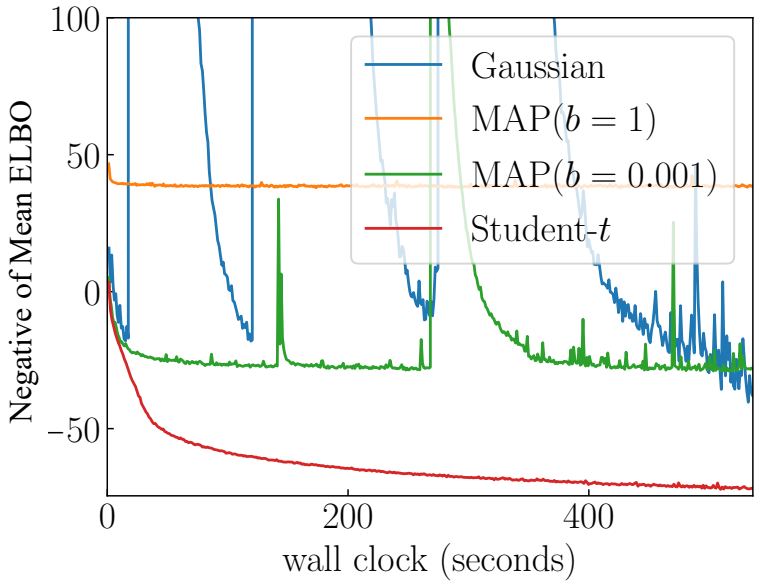
Student-t decoder

[Experiments]
Stability of training

- Our approach reduced the negative ELBO equal to or more stably than other approaches



SMTP



Aloi

Negative of Mean ELBO for each dataset

[Experiments]
Test log-likelihoods

- Our approach obtained the equal to or better density estimation performance than that of other approaches.

	Gaussian	MAP($b = 1$)	MAP($b = 0.001$)	Student-t
SMTP	-1.248 ± 0.404	-4.864 ± 0.020	-1.932 ± 0.404	**0.827 ± 0.105**
Aloi	45.418 ± 5.457	-38.210 ± 0.156	30.406 ± 0.383	**77.022 ± 0.539**
Thyroid	15.519 ± 4.422	-31.266 ± 0.159	18.037 ± 1.318	**69.543 ± 0.634**
Cancer	**-18.668 ± 3.448**	-45.895 ± 0.843	**-19.017 ± 3.273**	**-18.253 ± 2.629**
Satellite	**-1.852 ± 0.370**	-50.895 ± 0.238	**-1.899 ± 0.372**	**-1.811 ± 0.289**

Comparison of test log-likelihoods[1]

[1]We highlighted the best result in bold, and we also highlighted the results in bold which are not statistically different from the best result according to a pair-wise t-test.

In conclusion

- We proposed the Student-t VAE for robust multivariate density estimation

- We experimentally showed that the stability of the training and the high density estimation performance of the Student-t VAE

- We recommend using the Student-t distribution as the decoder If you use the VAE for continuous data!

Thank you!

Innovative R&D by NTT

Thank you for your attention!

If you have any questions,
email me: takahashi.hiroshi@lab.ntt.co.jp

NTT

FAQ

- **Q1: Did you compare this model with GAN?**
  - A1: With SMTP dataset, we compared Student-t VAE with Wasserstein GAN, and confirmed that the test log likelihood of the Student-t VAE was better than that of Wasserstein GAN.

- **Q2: What is the limitation of this approach?**
  - A2: This approach requires heavier computational cost than Gaussian decoder. (about 1.5 times)

- **Q3: Is this approach useful when the dataset is discrete?**
  - A3: If the dataset is binary, we recommend using the Bernoulli distribution as the decoder. Other than that, our approach may be useful.