Quantitative Understanding of VAE as a Non-linearly Scaled Isometric Embedding

Akira Nakagawa [1] **Keizo Kato** [1] **Taiji Suzuki** [2 3]

Abstract

Variational autoencoder (VAE) estimates the posterior parameters (mean and variance) of latent variables corresponding to each input data. While it is used for many tasks, the transparency of the model is still an underlying issue. This paper provides a quantitative understanding of VAE property through the differential geometric and information-theoretic interpretations of VAE. According to the Rate-distortion theory, the optimal transform coding is achieved by using an orthonormal transform with PCA basis where the transform space is isometric to the input. Considering the analogy of transform coding to VAE, we clarify theoretically and experimentally that VAE can be mapped to an implicit isometric embedding with a scale factor derived from the posterior parameter. As a result, we can estimate the data probabilities in the input space from the prior, loss metrics, and corresponding posterior parameters, and further, the quantitative importance of each latent variable can be evaluated like the eigenvalue of PCA.

1. Introduction

Variational autoencoder (VAE) (Kingma & Welling, 2014) is one of the most successful generative models, estimating posterior parameters of latent variables for each input data. In VAE, the latent representation is obtained by maximizing an evidence lower bound (ELBO). A number of studies have attempted to characterize the theoretical property of VAE. Indeed, there still are unsolved questions, e.g., what is the meaning of the latent variable VAE obtained, what β represents in β-VAE (Higgins et al., 2017), whether ELBO converges to an appropriate value, and so on. Alemi et al. (2018) introduced the RD trade-off based on the information-theoretic framework to analyse β-VAE. However, they did not clarify what VAE captures after optimization. Dai et al. (2018) showed VAE restricted as a linear transform can be considered as a robust principal component analysis (PCA). But, their model has a limitation for the analysis on each latent variable basis because of the linearity assumption. Rolínek et al. (2019) showed the Jacobian matrix of VAE is orthogonal, which seems to make latent variables disentangled implicitly. However, they do not uncover the impact of each latent variable on the input data quantitatively because they simplify KL divergence as a constant. Locatello et al. (2019) also showed the unsupervised learning of disentangled representations fundamentally requires inductive biases on both the metric and data. Yet, they also do not uncover the quantitative property of disentangled representations which is obtained under the given inductive biases. Kumar & Poole (2020) connected the VAE objective with the Riemannian metric and proposed new deterministic regularized objectives. However, they still did not uncover the quantitative property of VAE after optimizing their objectives.

These problems are essentially due to the lack of a clear formulation of the quantitative relationship between the input data and the latent variables. To overcome this point, Kato et al. (2020) propose an isometric autoencoder as a non-VAE model. In the *isometric embedding* (Han & Hong, 2006), the distance between arbitrary two input points is retained in the embedding space. With isometric embedding, the quantitative relationship between the input data and the latent variables is tractable. Our intuition is that if we could also map VAE to an isometric autoencoder, the behavior of VAE latent variables will become clear. Thus, the challenge of this paper is to resolve these essential problems by utilizing the view point of isometric embedding.

1. First of all, we show that VAE obtains an implicit isometric embedding of the support of the input distribution as its latent space. That is, the input variable is embedded through the encoder in a low dimensional latent space in which the distance in the given metric between two points in the input space is preserved. Surprisingly, this characterization resolves most unsolved problems of VAE such as what we

[1] Fujitsu Limited, Kanagawa, Japan [2] Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan [3] Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan. Correspondence to: Akira Nakagawa <anaka@fujitsu.com>.

Proceedings of the 38

have enumerated above. This implicit isometric embedding is derived as a non-linear scaling of VAE embedding.

2. More concretely, we will show the following issues via the isometric embedding characterization theoretically:
(a) Role of β in β-VAE: β controls each dimensional posterior variance of isometric embedding as a constant $\beta/2$.
(b) Estimation of input distribution: the input distribution can be quantitatively estimated from the distribution of implicit isometric embedding because of the constant Jacobian determinant between the input and implicit isometric spaces.
(c) Disentanglement: If the manifold has a separate latent variable in the given metric by nature, the implicit isometric embedding captures such disentangled features as a result of minimizing the entropy.
(d) Rate-distortion (RD) optimal: VAE can be considered as a rate-distortion optimal encoder formulated by RD theory (Berger, 1971).

3. Finally, we justify our theoretical findings through several numerical experiments. We observe the estimated distribution is proportional to the input distribution in the toy dataset. By utilizing this property, the performance of the anomaly detection for real data is comparable to state-of-the-art studies. We also observe that the variance of each dimensional component in the isometric embedding shows the importance of each disentangled property like PCA.

2. Variational autoencoder

In VAE, ELBO is maximized instead of maximizing the log-likelihood directly. Let $\boldsymbol{x} \in \mathbb{R}^m$ be a point in a dataset. The original VAE model consists of a latent variable with fixed prior $\boldsymbol{z} \sim p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I}_n) \in \mathbb{R}^n$, a parametric encoder $\mathrm{Enc}_\phi : \boldsymbol{x} \Rightarrow \boldsymbol{z}$, and a parametric decoder $\mathrm{Dec}_\theta : \boldsymbol{z} \Rightarrow \hat{\boldsymbol{x}}$. In the encoder, $q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_{(\boldsymbol{x})}, \boldsymbol{\sigma}_{(\boldsymbol{x})})$ is provided by estimating parameters $\boldsymbol{\mu}_{(\boldsymbol{x})}$ and $\boldsymbol{\sigma}_{(\boldsymbol{x})}$. Let $L_{\boldsymbol{x}}$ be a local cost at data $\boldsymbol{x}$. Then, ELBO is described by

$$E_{p(\boldsymbol{x})}\left[E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \| p(\boldsymbol{z}))\right]. \quad (1)$$

In $E_{p(\boldsymbol{x})}[\,\cdot\,]$, the second term $D_{\mathrm{KL}}(\cdot)$ is a Kullback–Leibler (KL) divergence. Let $\mu_{j(\boldsymbol{x})}$, $\sigma_{j(\boldsymbol{x})}$, and $D_{\mathrm{KL}j(x)}$ be j-th dimensional values of $\boldsymbol{\mu}_{(\boldsymbol{x})}$, $\boldsymbol{\sigma}_{(\boldsymbol{x})}$, and KL divergence, respectively. Then $D_{\mathrm{KL}}(\cdot)$ is derived as:

$$D_{\mathrm{KL}}(\cdot) = \sum_{j=1}^{n} D_{\mathrm{KL}j(x)}, \quad \text{where}$$
$$D_{\mathrm{KL}j(x)} = \frac{1}{2}\left({\mu_{j(\boldsymbol{x})}}^2 + {\sigma_{j(\boldsymbol{x})}}^2 - \log {\sigma_{j(\boldsymbol{x})}}^2 - 1\right). \quad (2)$$

The first term $E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$ is called the reconstruction loss. Instead directly estimate $\log p_\theta(\boldsymbol{x}|\boldsymbol{z})$ in training, $\hat{\boldsymbol{x}} = \mathrm{Dec}_\theta(\boldsymbol{z})$ is derived and $D(\boldsymbol{x}, \hat{\boldsymbol{x}}) = -\log p_{\mathbb{R}p}(\boldsymbol{x}|\hat{\boldsymbol{x}})$ is evaluated as reconstruction loss, where $p_{\mathbb{R}p}(\boldsymbol{x}|\hat{\boldsymbol{x}})$ denotes the predetermined conditional distribution. In the case Gaussian and Bernoulli distributions are used as $p_{\mathbb{R}p}(\boldsymbol{x}|\hat{\boldsymbol{x}})$, $D(\boldsymbol{x}, \hat{\boldsymbol{x}})$ becomes the sum square error (SSE) and binary cross-entropy (BCE), respectively. In training β-VAE (Higgins et al., 2017), the next objective is used instead of Eq. 1, where β is a parameter to control the trade-off.

$$L_{\boldsymbol{x}} = E_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[D(\boldsymbol{x}, \hat{\boldsymbol{x}})] + \beta D_{\mathrm{KL}}(\cdot). \quad (3)$$

3. Isometric embedding

Isometric embedding (Han & Hong, 2006) is a smooth embedding from $\boldsymbol{x}$ to $\boldsymbol{z}$ ($\boldsymbol{x}, \boldsymbol{z} \in \mathbb{R}^m$) on a Riemannian manifold where the distances between arbitrary two points are equivalent in both the input and embedding spaces. Assume that $\boldsymbol{x}$ and $\boldsymbol{z}$ belong to a Riemannian metric space with a metric tensor $\boldsymbol{G}_{\boldsymbol{x}}$ and a Euclidean space, respectively. Then, the isometric embedding from $\boldsymbol{x}$ to $\boldsymbol{z}$ satisfies the following condition for all inputs and dimensions as shown in Kato et al. (2020), where δ_{jk} denotes Kronecker delta:

$${}^t\partial \boldsymbol{x}/\partial z_j \, \boldsymbol{G}_{\boldsymbol{x}} \, \partial \boldsymbol{x}/\partial z_k = \delta_{jk}. \quad (4)$$

The isometric embedding has several preferable properties. First of all, the probability density of input data at the given metric is preserved in the isometric embedding space. Let $p(\boldsymbol{x})$ and $p(\boldsymbol{z})$ be distributions in their respective metric spaces. $J_{\det}$ denotes $|\det(\partial \boldsymbol{x}/\partial \boldsymbol{z})|$, i.e., an absolute value of the Jacobian determinant. Since $J_{\det}$ is 1 from orthonormality, the following equation holds:

$$p(\boldsymbol{z}) = J_{\det}\, p(\boldsymbol{x}) = p(\boldsymbol{x}). \quad (5)$$

Secondly, the entropies in both spaces are also equivalent. Let X and Z be sets of $\boldsymbol{x}$ and $\boldsymbol{z}$, respectively. $H(X)$ and $H(Z)$ denotes the entropies of X and Z in each metric spaces. Then $H(X)$ and $H(Z)$ are equivalent as follows:

$$\begin{aligned} H(Z) &= -\int p(\boldsymbol{z}) \log p(\boldsymbol{z}) \,\mathrm{d}\boldsymbol{z} \\ &= -\int J_{\det}\, p(\boldsymbol{x}) \log \left(J_{\det}\, p(\boldsymbol{x})\right) J_{\det}^{-1} \,\mathrm{d}\boldsymbol{x} \\ &= -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} \\ &= H(X). \end{aligned} \quad (6)$$

Thus, the isometric embedding is a powerful tool to analyse input data. Note that Eqs. 5-6 do not hold in general if the embedding is not isometric.

Recently, Kato et al. (2020) proposed an isometric autoencoder RaDOGAGA (Rate-distortion optimization guided autoencoder for generative analysis), inspired by deep image compression (Ballé et al., 2018). In the conventional image compression using orthonormal transform coding, Rate-distortion optimization (RDO) objective has been widely used (Sullivan & Wiegand, 1998). Let R and D be a rate and distortion after encoding, respectively. Then RDO finds the best encoding parameters that minimizes $L = D + \lambda R$ at given Lagrange multiplier λ. In the deep image compression (Ballé et al., 2018), the model is composed of a parametric prior and posterior with constant variance, then trained using the RDO objective. Kato et al. (2020) proved that such a model achieves an isometric embedding in Euclidean space, and they proposed an isometric autoencoder RaDOGAGA for quantitative analysis. By contrast, VAE uses a fixed prior with a variable posterior. Here, we have an intuition that VAE can be mapped to an isometric embedding such as RaDOGAGA by introducing a non-linear scaling of latent space. If our intuition is correct, the behavior of VAE will be quantitatively explained.

4. Understanding of VAE as a scaled isometric embedding

This section shows the quantitative property of VAE by introducing an implicit isometric embedding. First, we present the hypothesis of mapping VAE to an implicit isometric embedding. Second, we theoretically formulate the derivation of implicit isometric embedding as the minimum condition of the VAE objective. Lastly, we explain the quantitative properties of VAE to provide a practical data analysis.

4.1. Mapping β-VAE to implicit isometric embedding

In this section, we explain our motivations for introducing an implicit isometric embedding to analyse β-VAE. Rolínek et al. (2019) showed that each pair of column vectors in the Jacobian matrix $\partial \boldsymbol{x}/\partial \boldsymbol{\mu}_{(\boldsymbol{x})}$ is orthogonal such that ${}^t\partial \boldsymbol{x}/\partial \mu_{j(\boldsymbol{x})} \cdot \partial \boldsymbol{x}/\partial \mu_{k(\boldsymbol{x})} = 0$ for $j \neq k$ when $D(\boldsymbol{x}, \hat{\boldsymbol{x}})$ is SSE. From this property, we can introduce the implicit isometric embedding by scaling the VAE latent space appropriately as follows: $\boldsymbol{x}_{\mu_j}$ denotes $\partial \boldsymbol{x}/\partial \mu_{j(\boldsymbol{x})}$. Let $\boldsymbol{y}$ and y_j be an implicit variable and its j-th dimensional component which satisfies $\mathrm{d}y_j/\mathrm{d}\mu_{j(\boldsymbol{x})} = |\boldsymbol{x}_{\mu_j}|_2$. Then $\partial \boldsymbol{x}/\partial y_j$ forms the isometric embedding in Euclidean space:

$$ {}^t\partial \boldsymbol{x}/\partial y_j \cdot \partial \boldsymbol{x}/\partial y_k = \delta_{jk}. \tag{7}$$

If the L2 norm of $\boldsymbol{x}_{\mu_j}$ is derived mathematically, we can formulate the mapping VAE to an implicit isometric embedding as in Eq. 7. Then, this mapping will strongly help to understand the quantitative behavior of VAE as explained in section 3. Thus, our motivation in this paper is to formulate the implicit isometric embedding theoretically and analyse VAE properties quantitatively.

Figure 1 shows how β-VAE is mapped to an implicit isometric embedding. In VAE encoder, $\boldsymbol{\mu}_{(\boldsymbol{x})}$ is calculated from an input $\boldsymbol{x} \in X$. Then, the posterior $\boldsymbol{z}$ is derived by adding a stochastic noise $\mathcal{N}(0, \boldsymbol{\sigma}_{(\boldsymbol{x})})$ to $\boldsymbol{\mu}_{(\boldsymbol{x})}$. Finally, the reconstruction data $\hat{\boldsymbol{x}} \in \hat{X}$ is decoded from $\boldsymbol{z}$.

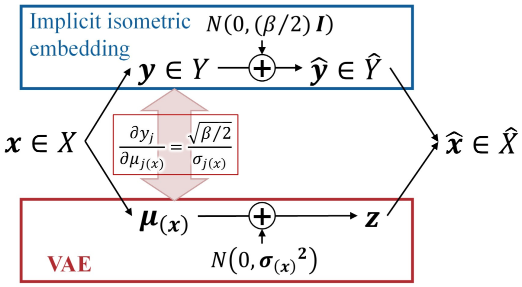


Figure 1. Mapping of β-VAE to implicit isometric embedding.

Our theoretical analysis in section 4.2 reveals that implicit isometric embedding $\boldsymbol{y} \in Y$ can be introduced by mapping $\boldsymbol{\mu}_{(\boldsymbol{x})}$ to $\boldsymbol{y}$ with a scaling $\mathrm{d}y_j/\mathrm{d}\mu_{j(\boldsymbol{x})} = |\boldsymbol{x}_{\mu_j}|_2 = \sqrt{\beta/2}/\sigma_{j(\boldsymbol{x})}$ in each dimension. Then, the posterior $\hat{\boldsymbol{y}} \in \hat{Y}$ is derived by adding a stochastic noise $\mathcal{N}(0, (\beta/2)I_n)$ to $\boldsymbol{y}$. Note that the noise variances, i.e., the posterior variances, are a constant $\beta/2$ for all inputs and dimensions, which is analogous to RaDOGAGA. Then, the mutual information $H(X;\hat{X})$ in β-VAE can be estimated as:

$$\begin{aligned} I(X;\hat{X}) &= I(Y;\hat{Y}) \\ &\simeq H(Y) - H\left(\mathcal{N}(0,(\beta/2)I_n)\right) \\ &= H(Y) - \frac{n}{2}\log(\pi e \beta). \end{aligned} \tag{8}$$

This implies that the posterior entropy $\frac{n}{2}\log(\pi e \beta)$ should be smaller enough than $H(X)$ to give the model a sufficient expressive ability. Thus, the posterior variance $\beta/2$ should be also sufficiently smaller than the variance of input data. Note that Eq. 8 is consistent with the Rate-distortion (RD) optimal condition in the RD theory as shown in section 6.

4.2. Theoretical derivation of implicit isometric embedding

In this section, we derive the implicit isometric embedding theoretically. First, we reformulate $D(\boldsymbol{x}, \hat{\boldsymbol{x}})$ and $D_{\mathrm{KL}}(\,\cdot\,)$ in β-VAE objective $L_{\boldsymbol{x}}$ in Eq. 3 for mathematical analysis. Then we derive the implicit isometric embedding as a minimum condition of $L_{\boldsymbol{x}}$. Here, we set the prior $p(\boldsymbol{z})$ to $\mathcal{N}(\boldsymbol{z}; 0, I_n)$ for easy analysis. The condition where the approximation in this section is valid is that $\beta/2$ is smaller enough than the variance of the input dataset, which is important to achieve a sufficient expressive ability. We also assume the data manifold is smooth and differentiable.

Firstly, we introduce a metric tensor to treat arbitrary kinds of metrics for the reconstruction loss in the same framework. $D(\boldsymbol{x}, \hat{\boldsymbol{x}}) = -\log p_{\mathbb{R}p}(\boldsymbol{x}|\hat{\boldsymbol{x}})$ denotes a metric between two points $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$. Let $\delta\boldsymbol{x}$ be $\hat{\boldsymbol{x}} - \boldsymbol{x}$. If $\delta\boldsymbol{x}$ is small, $D(\boldsymbol{x}, \hat{\boldsymbol{x}}) = D(\boldsymbol{x}, \boldsymbol{x} + \delta\boldsymbol{x})$ can be approximated by ${}^t\delta\boldsymbol{x}\, \boldsymbol{G}_{\boldsymbol{x}} \delta\boldsymbol{x}$ using the second order Taylor expansion, where $\boldsymbol{G}_{\boldsymbol{x}}$ is an $\boldsymbol{x}$ dependent positive definite metric tensor. Appendix G.2 shows the derivations of $\boldsymbol{G}_{\boldsymbol{x}}$ for SSE, BCE, and structural similarity (SSIM) (Wang et al., 2001). Especially for SSE, $\boldsymbol{G}_{\boldsymbol{x}}$ is an identity matrix $\boldsymbol{I}$, i.e., a metric tensor in Euclidean space.

Next, we formulate the approximation of L_x via the following three lemmas, to examine the Jacobian matrix easily.

Lemma 1. Approximation of reconstruction loss: Let $\breve{\boldsymbol{x}}$ be $\mathrm{Dec}_\theta(\boldsymbol{\mu}_{(\boldsymbol{x})})$. $\boldsymbol{x}_{\mu_j}$ denotes $\partial \boldsymbol{x}/\partial \mu_{j(\boldsymbol{x})}$. Then the reconstruction loss in L_x can be approximated as:

$$E_{\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[D(\boldsymbol{x},\hat{\boldsymbol{x}})\right] \simeq D(\boldsymbol{x},\breve{\boldsymbol{x}})+\sum_{j=1}^{n} {\sigma_{j(\boldsymbol{x})}}^2\, {}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_j}. \tag{9}$$

Proof: Appendix A.1 describes the proof. The outline is as follows: Rolínek et al. (2019) show $D(\boldsymbol{x},\hat{\boldsymbol{x}})$ can be decomposed to $D(\boldsymbol{x},\breve{\boldsymbol{x}})+D(\breve{\boldsymbol{x}},\hat{\boldsymbol{x}})$. We call the first term $D(\boldsymbol{x},\breve{\boldsymbol{x}})$ a transform loss. Obviously, the average of transform loss over $\boldsymbol{z}\sim q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is still $D(\boldsymbol{x},\breve{\boldsymbol{x}})$. We call the second term $D(\breve{\boldsymbol{x}},\hat{\boldsymbol{x}})$ a coding loss. The average of coding loss can be further approximated as the second term of Eq. 9.

Lemma 2. Approximation of KL divergence: Let $p(\boldsymbol{\mu}_{(\boldsymbol{x})})=\mathcal{N}(\boldsymbol{\mu}_{(\boldsymbol{x})};0,I_n)$ be a prior probability density where $\boldsymbol{z}=\boldsymbol{\mu}_{(\boldsymbol{x})}$. Then the KL divergence in L_x can be approximated as:

$$\begin{aligned}&D_{\mathrm{KL}}(q_\phi(\boldsymbol{z}|\boldsymbol{x})\|p(\boldsymbol{z}))\\ &\simeq -\log\Big(p(\boldsymbol{\mu}_{(\boldsymbol{x})})\prod_{j=1}^{n}\sigma_{j(\boldsymbol{x})}\Big)-\frac{n\log 2\pi e}{2}\\ &\simeq -\log\Big(p(\boldsymbol{x})\left|\det\Big(\frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\mu}_{(\boldsymbol{x})}}\Big)\right|\prod_{j=1}^{n}\sigma_{j(\boldsymbol{x})}\Big)-\frac{n\log 2\pi e}{2}.\end{aligned} \tag{10}$$

Proof: The detail is described in Appendix A.2. The outlines is as follows: First, ${\sigma_{j(\boldsymbol{x})}}^2 \ll 1$ will be observed in meaningful dimensions. For example, ${\sigma_{j(\boldsymbol{x})}}^2 < 0.1$ will almost hold if the dimensional component has information that exceeds only 1.2 nat. Furthermore, when ${\sigma_{j(\boldsymbol{x})}}^2<0.1$, we have $-({\sigma_{j(\boldsymbol{x})}}^2/\log {\sigma_{j(\boldsymbol{x})}}^2)<0.05$. Thus, by ignoring the ${\sigma_{j(\boldsymbol{x})}}^2$ in Eq. 2, $D_{\mathrm{KL}j(x)}$ can be approximated as:

$$\begin{aligned}D_{\mathrm{KL}j(x)} &\simeq \frac{1}{2}\left({\mu_{j(\boldsymbol{x})}}^2-\log{\sigma_{j(\boldsymbol{x})}}^2-1\right)\\ &= -\log\left(\sigma_{j(\boldsymbol{x})}\,\mathcal{N}(\mu_{j(\boldsymbol{x})};0,1)\right)\; - \frac{\log 2\pi e}{2}.\end{aligned} \tag{11}$$

As a result, the second equation of the proposition Eq. 10 is derived by summing the last equation of Eq. 11. Then, using $p(\boldsymbol{\mu}_{(\boldsymbol{x})})=p(\boldsymbol{x})\,|\det(\partial\boldsymbol{x}/\partial\boldsymbol{\mu}_{(\boldsymbol{x})})|$, the last equation of Eq. 10 is derived. Appendix A.2 shows that the approximation of the second line in Eq. 10 can be also derived for arbitrary priors, which suggests that the theoretical derivations that follow in this section also hold for arbitrary priors.

Lemma 3. Estimation of transform loss: Let $x\sim\mathcal{N}(x;0,{\sigma_x}^2)$ be a 1-dimensional dataset. When β-VAE is trained for x, the ratio between the transform loss $D(x,\breve{x})$ and the coding loss $D(\breve{x},\hat{x})$ is estimated as:

$$\frac{D(x,\breve{x})}{D(\breve{x},\hat{x})}\simeq\frac{\beta/2}{{\sigma_x}^2}. \tag{12}$$

Proof: Appendix A.3 describes the proof. As explained there, this is analogous to the Wiener filter (Wiener, 1964), one of the most basic theories for signal restoration.

Lemma 3 is also validated experimentally in the multi-dimensional non-Gaussian toy dataset. Fig. 24 in Appendix D.2 shows that the experimental results match the theory well. Thus, we ignore the transform loss $D(x,\breve{x})$ in the discussion that follows, since we assume $\beta/2$ is smaller enough than the variance of the input data. Using Lemma 1-3, we can derive the approximate expansion of L_x as follows:

Theorem 1. Approximate expansion of VAE objective: Assume $\beta/2$ is smaller enough than the variance of input dataset. The objective L_x can be approximated as:

$$\begin{aligned}L_{\boldsymbol{x}} &\simeq \sum_{j=1}^{n}{\sigma_{j(\boldsymbol{x})}}^2\,{}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_j}\\ &-\beta\log\Big(p(\boldsymbol{x})\left|\det\Big(\frac{\partial\boldsymbol{x}}{\partial\boldsymbol{\mu}_{(\boldsymbol{x})}}\Big)\right|\prod_{j=1}^{n}\sigma_{j(\boldsymbol{x})}\Big)-\frac{n\beta\log 2\pi e}{2}.\end{aligned} \tag{13}$$

Proof: Apply Lemma 1-3 to $L_{\boldsymbol{x}}$ in Eq. 3.

Then, we can finally derive the *implicit isometric embedding* as a minimum condition of Eq. 13 via Lemma 4-5.

Lemma 4. Orthogonality of Jacobian matrix in VAE: At the minimum condition of Eq. 13, each pair $\boldsymbol{x}_{\mu_j}$ and $\boldsymbol{x}_{\mu_k}$ of column vectors in the Jacobian matrix $\partial\boldsymbol{x}/\partial\boldsymbol{\mu}_{(\boldsymbol{x})}$ show the orthogonality in the Riemannian metric space, i.e., the inner product space with the metric tensor $\boldsymbol{G}_{\boldsymbol{x}}$ as:

$$({2\sigma_{j(\boldsymbol{x})}}^2/\beta)\;{}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_k}=\delta_{jk}. \tag{14}$$

Proof: Eq. 14 is derived by examining the derivative $\mathrm{d}L_{\boldsymbol{x}}/\mathrm{d}\boldsymbol{x}_{\mu_j}=0$. The proof is described in Appendix A.4. A diagonal posterior covariance is the key for orthogonality.

Eq. 14 is consistent with Rolínek et al. (2019) who show the orthogonality for SSE metric. In addition, we quantify the Jacobian matrix for arbitrary metric spaces.

Lemma 5. L2 norm of $\boldsymbol{x}_{\mu_j}$: the L2 norm of $\boldsymbol{x}_{\mu_j}$ in the metric space of $\boldsymbol{G}_{\boldsymbol{x}}$ is derived as:

$$|\boldsymbol{x}_{\mu_j}|_2=\sqrt{{}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_j}}=\sqrt{\beta/2}/\sigma_{j(\boldsymbol{x})}. \tag{15}$$

Proof: Apply $k=j$ to Eq. 14 and arrange it.

Theorem 2. Implicit isometric embedding: An implicit isometric embedding $\boldsymbol{y}$ is introduced by mapping j-th component $\mu_{j(\boldsymbol{x})}$ of VAE latent variable to y_j with the following scaling factor:

$$\mathrm{d}y_j/\mathrm{d}\mu_{j(\boldsymbol{x})} = |\boldsymbol{x}_{\mu_j}|_2 = \sqrt{\beta/2}/\sigma_{j(\boldsymbol{x})}. \tag{16}$$

$\boldsymbol{x}_{y_j}$ denotes $\partial\boldsymbol{x}/\partial y_j$. Then $\boldsymbol{x}_{y_j}$ satisfies the next equation:

$${}^t\boldsymbol{x}_{y_j}\boldsymbol{G_x}\boldsymbol{x}_{y_k} = \delta_{jk}. \tag{17}$$

This shows the isometric embedding from the inner product space of $\boldsymbol{x}$ with metric $\boldsymbol{G_x}$ to the Euclidean space of $\boldsymbol{y}$.

Proof: Apply $\boldsymbol{x}_{\mu_j} = \mathrm{d}y_j/\mathrm{d}\mu_{j(\boldsymbol{x})}\ \boldsymbol{x}_{y_j}$ to Eq. 14.

Remark 1: Isometricity in Eq. 17 is on the decoder side. Since the transform loss $D(\boldsymbol{x}, \breve{\boldsymbol{x}})$ is close to 0, $\mathrm{Dec}_\theta(\boldsymbol{\mu}_{j(\boldsymbol{x})}) \simeq \mathrm{Enc}_\phi^{-1}(\boldsymbol{\mu}_{j(\boldsymbol{x})})$ holds. As a result, the isometricity on the encoder side is also almost achieved. If $D(\boldsymbol{x}, \breve{\boldsymbol{x}})$ is explicitly reduced by using a decomposed loss, the isometricity will be further promoted.

Theorem 3. Posterior variance in isometric embedding: The posterior variance of implicit isometric embedding is a constant $\beta/2$ for all inputs and dimensional components.

Proof: Let $\sigma_{y_j(\boldsymbol{x})}{}^2$ be a posterior variance of the implicit isometric component y_j. By scaling $\sigma_{j(\boldsymbol{x})}$ for the original VAE latent variable with Eq. 16, $\sigma_{y_j(\boldsymbol{x})}$ is derived as:

$$\sigma_{y_j(\boldsymbol{x})} \simeq \sigma_{j(\boldsymbol{x})}\frac{\mathrm{d}y_j}{\mathrm{d}\mu_{j(\boldsymbol{x})}} = \sqrt{\beta/2}. \tag{18}$$

Thus, the posterior variance $\sigma_{y_j(\boldsymbol{x})}{}^2$ is a constant $\beta/2$ for all dimensions j at any inputs $\boldsymbol{x}$ as in Section 4.1.

4.3. Quantitative data analysis method using implicit isometric embedding in VAE

This section describes three quantitative data analysis methods by utilizing the property of isometric embedding.

4.3.1. Estimation of the data probability distribution:

Estimation of data distribution is one of the key targets in machine learning. We show VAE can estimate the distribution in both metric space and input space quantitatively.

Proposition 1. Probability estimation in metric space: Let $p_{\boldsymbol{G_x}}(\boldsymbol{x})$ be a probability distribution in the inner product space of $\boldsymbol{G_x}$. $p_{\boldsymbol{G_x}}(\boldsymbol{x})$ can be quantitatively estimated as:

$$\begin{aligned} p_{\boldsymbol{G_x}}(\boldsymbol{x}) \ \simeq\ p(\boldsymbol{y}) \quad &\propto \quad p(\boldsymbol{\mu}_{(\boldsymbol{x})})\prod_{j=1}^{m}\sigma_{j(\boldsymbol{x})} \\ &\propto \quad \exp(-L_{\boldsymbol{x}}/\beta). \end{aligned} \tag{19}$$

Proof: Appendix A.5 explains the detail. The outline is as follows: The third equation is derived by applying Eq. 16 to $p(\boldsymbol{y}) = \prod_j p(y_j) = \prod_j (\mathrm{d}y_j/\mathrm{d}\mu_{j(\boldsymbol{x})})^{-1}p(\mu_j)$, showing that $\sigma_{j(\boldsymbol{x})}$ bridges between the distributions of input data and prior. The fourth equation is derived by applying Eq. 16 to Eq. 13. The last equation implies that the VAE objective converges to the log-likelihood of the input $\boldsymbol{x}$ as expected. When the metric is SSE, Eq. 19 show the probability distribution in the input space since $\boldsymbol{G_x}$ is an identity matrix.

Proposition 2. Probability estimation in the input space: In the the case $m = n$, the probability distribution $p(\boldsymbol{x})$ in the input space can be estimated as:

$$\begin{aligned} p(\boldsymbol{x}) = |\det(\boldsymbol{G}_x)|^{\frac{1}{2}}\ p_{\boldsymbol{G_x}}(\boldsymbol{x}) &\simeq |\det(\boldsymbol{G}_x)|^{\frac{1}{2}}\ p(\boldsymbol{y}) \\ &\propto |\det(\boldsymbol{G}_x)|^{\frac{1}{2}}\ p(\boldsymbol{\mu}_{(\boldsymbol{x})})\prod_{j=1}^{m}\sigma_{j(\boldsymbol{x})} \\ &\propto |\det(\boldsymbol{G}_x)|^{\frac{1}{2}}\exp(-L_{\boldsymbol{x}}/\beta). \end{aligned} \tag{20}$$

In the case $m > n$ and $\boldsymbol{G}_x = a_{\boldsymbol{x}}\boldsymbol{I}_m$ holds where $a_{\boldsymbol{x}}$ is an $\boldsymbol{x}$-dependent scalar factor, $p(\boldsymbol{x})$ can be estimated as:

$$p(\boldsymbol{x}) \propto a_{\boldsymbol{x}}{}^{\frac{n}{2}}\ p(\boldsymbol{\mu}_{(\boldsymbol{x})})\prod_{j=1}^{n}\sigma_{j(\boldsymbol{x})} \propto a_{\boldsymbol{x}}{}^{\frac{n}{2}}\exp(-L_{\boldsymbol{x}}/\beta). \tag{21}$$

Proof: The absolute value of Jacobian determinant between the input and metric spaces gives the the PDF ratio. In the case $m = n$, this is derived as $|\det(\boldsymbol{G}_x)|^{\frac{1}{2}}$. In the case $m > n$ and $\boldsymbol{G}_x = a_{\boldsymbol{x}}\boldsymbol{I}_m$, the Jacobian determinant is proportional to $a_x{}^{n/2}$. Appendix A.6 explains the detail.

4.3.2. Quantitative analysis of disentanglement

Assume the data manifold has a disentangled property with independent latent variable by nature. Then each y_j will capture each disentangled latent variable like to PCA. This subsection explains how to derive the importance of each dimension in the given metrics for data analysis.

Proposition 3. Meaningful dimension: The dimensional components y_j with $D_{\mathrm{KL}j(x)} > 0$ have meaningful information for representation, where the entropy of y_j is larger than $H(\mathcal{N}(0, \beta/2)) = \log(\beta\pi e)/2$. In contrast, the dimension with $D_{\mathrm{KL}j(x)} = 0$ has no information, where $\mu_{j(\boldsymbol{x})} = 0$ and $\sigma_{j(\boldsymbol{x})} = 1$ will be observed.

Proof: Appendix A.7 shows the detail in view of RD theory. This appendix also explains that the entropy of $\boldsymbol{y}$ becomes minimum after optimization.

Proposition 4. Importance of each dimension: Assume that the prior $p(\boldsymbol{z})$ is a Gaussian distribution $\mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I}_n)$. Let $\mathrm{Var}(y_j)$ be the variance of the j-th implicit isometric component y_j, indicating the quantitative importance of each dimension. $\mathrm{Var}(y_j)$ in the meaningful dimension ($D_{\mathrm{KL}j(x)} > 0$) can be roughly estimated as:

$$\mathrm{Var}(y_j) \simeq (\beta/2)\ E_{\boldsymbol{x}\sim p(\boldsymbol{x})}[\sigma_{j(\boldsymbol{x})}{}^{-2}]. \tag{22}$$

Proof: Appendix A.8 shows the derivation from Eq. 16. The case other than Gaussian prior is also explained there.

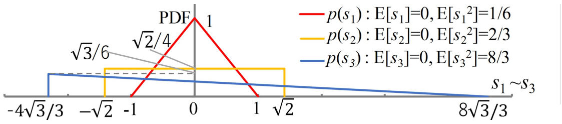


Figure 2. PDFs of three variables to generate a toy dataset.

4.3.3. CHECK THE ISOMETRICITY AFTER TRAINING

This subsection explains how to determine if the model acquires isometric embedding by evaluating the norm of $\boldsymbol{x}_{y_j}$. Let $\boldsymbol{e}_{(j)}$ be a vector $(0, \cdots, 1, \cdots, 0)$ where the j-th dimension is 1 and others are 0. Let $D'_j(\boldsymbol{z})$ be $D(\mathrm{Dec}_\theta(\boldsymbol{z}), \mathrm{Dec}_\theta(\boldsymbol{z}+\epsilon\boldsymbol{e}_{(j)}))/\epsilon^2$, where ϵ denotes a minute value for the numerical differential. Then the squared L2 norm of y_j can be evaluated as the last equation:

$$
\begin{aligned}
{}^t\boldsymbol{x}_{y_j}\boldsymbol{G}_x\boldsymbol{x}_{y_j} &\simeq (2/\beta)\,\left(\sigma_{j(\boldsymbol{x})}{}^2\ {}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_j}\right) \\
&\simeq (2/\beta)\,\sigma_{j(\boldsymbol{x})}{}^2 D'_j(\boldsymbol{z}). \qquad (23)
\end{aligned}
$$

Observing a value close to 1 means a unit norm and indicates that an implicit isometric embedding is captured.

Remark 2: Eq. 23 will not hold and the norm will be 0 in such a dimension where $D_{\mathrm{KL}j(x)} = 0$, since the reconstruction loss, i.e., $\beta/2$ times squared L2 norm of y_j, and $D_{\mathrm{KL}j(x)}$ do not have to be balanced in Eq. 13.

5. Experiment

This section describes three experimental results. First, the results of the toy dataset are examined to validate our theory. Next, the disentanglement analysis for the CelebA dataset is presented. Finally, an anomaly detection task is evaluated to show the usefulness of data distribution estimation.

5.1. Quantitative evaluation in the toy dataset

The toy dataset is generated as follows. First, three dimensional variables s_1, s_2, and s_3 are sampled in accordance with the three different shapes of distributions $p(s_1)$, $p(s_2)$, and $p(s_3)$, as shown in Fig. 2. The variances of s_1, s_2, and s_3 are $1/6$, $2/3$, and $8/3$, respectively, such that the ratio of the variances is 1:4:16. Second, three 16-dimensional uncorrelated vectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, and $\boldsymbol{v}_3$ with L2 norm 1 are provided. Finally, $50,000$ toy data with 16 dimensions are generated by $\boldsymbol{x} = \sum_{i=1}^{3} s_i\boldsymbol{v}_i$. The data distribution $p(\boldsymbol{x})$ is also set to $p(s_1)p(s_2)p(s_3)$. If our hypothesis is correct, $p(y_j)$ will be close to $p(s_j)$. Then, $\sigma_{j(\boldsymbol{x})} \propto \mathrm{d}z_j/\mathrm{d}y_j = p(y_j)/p(z_j)$ will also vary a lot with these varieties of PDFs. Because the properties in Section 4.3 are derived from $\sigma_{j(\boldsymbol{x})}$, our theory can be easily validated by evaluating those properties.

Then, the VAE model is trained using Eq. 3. We use two kinds of the reconstruction loss $D(\cdot,\cdot)$ to analyze the effect of the loss metrics. The first is the square error loss equivalent to SSE. The second is the downward-convex loss which we design as Eq. 24, such that the shape becomes similar to the BCE loss as in Appendix G.2:

$$
\begin{aligned}
&D(\boldsymbol{x}, \hat{\boldsymbol{x}}) = a_{\boldsymbol{x}}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2^2, \\
&\quad \text{where } a_{\boldsymbol{x}} = (2/3 + 2\ \|\boldsymbol{x}\|_2^2/21) \text{ and } \boldsymbol{G}_x = a_{\boldsymbol{x}}\boldsymbol{I}_m. \quad (24)
\end{aligned}
$$

Here, $a_{\boldsymbol{x}}$ is chosen such that the mean of $a_{\boldsymbol{x}}$ for the toy dataset is 1.0 since the variance of $\boldsymbol{x}$ is 1/6+2/3+8/3=7/2. The details of the networks and training conditions are written in Appendix C.1.

After training with two types of reconstruction losses, the loss ratio $D(x,\breve{x})/D(\breve{x},\hat{x})$ for the square error loss is 0.023, and that for the downward-convex loss is 0.024. As expected in Lemma 3, the transform losses are negligibly small.

First, an implicit isometric property is examined. Tables 1 and 2 show the measurements of $\frac{2}{\beta}\sigma_{j(\boldsymbol{x})}{}^2 D'_j(\boldsymbol{z})$ (shown as $\frac{2}{\beta}\sigma_j{}^2 D'_j$), $D'_j(\boldsymbol{z})$, and $\sigma_{j(\boldsymbol{x})}{}^{-2}$ described in Section 4.3. In these tables, z_1, z_2, and z_3 show acquired latent variables. "Av." and "SD" are the average and standard deviation, respectively. In both tables, the values of $\frac{2}{\beta}\sigma_{(\boldsymbol{x})j}{}^2 D'_j(\boldsymbol{z})$ are close to 1.0 in each dimension, showing isometricity as in Eq. 21. By contrast, the average of $D'_j(\boldsymbol{z})$, which corresponds to ${}^t\boldsymbol{x}_{\mu_j}\boldsymbol{G}_x\boldsymbol{x}_{\mu_j}$, is different in each dimension. Thus, $\boldsymbol{x}_{\mu_k}$ for the original VAE latent variable is not isometric.

Next, the disentanglement analysis is examined. The average of $\sigma_{j(\boldsymbol{x})}{}^{-2}$ in Eq.22 and its ratio are shown in Tables 1 and 2. Although the average of $\sigma_{j(\boldsymbol{x})}{}^{-2}$ is a rough estimation of variance, the ratio is close to 1:4:16, i.e., the variance ratio of generation parameters s_1, s_2, and s_3. When comparing both losses, the ratio of s_2 and s_3 for the downward-convex loss is somewhat smaller than that for the square error. This is explained as follows. In the downward-convex loss, $|\boldsymbol{x}_{y_j}|_2^2$ tends to be $1/a_{\boldsymbol{x}}$ from Eq. 17, i.e. ${}^t\boldsymbol{x}_{y_j}\,(a_x\boldsymbol{I}_m)\,\boldsymbol{x}_{y_k} = \delta_{jk}$. Therefore, the region in the metric space with a larger norm is shrunk, and the estimated variances corresponding to s_2 and s_3 become smaller.

Finally, we examine the probability estimation. Figure 3 shows the scattering plots of the data distribution $p(\boldsymbol{x})$ and estimated probabilities for the downward-convex loss. Figure 3a shows the plots of $p(\boldsymbol{x})$ and the prior probabilities $p(\boldsymbol{\mu}_{(\boldsymbol{x})})$. This graph implies that it is difficult to estimate $p(\boldsymbol{x})$ only from the prior. The correlation coefficient shown as "R" (0.434) is also low. Figure 3b shows the plots of $p(\boldsymbol{x})$ and $\exp(-L_{\boldsymbol{x}}/\beta)$ in in Eq. 19. The correlation coefficient (0.771) becomes better, but is still not high. Lastry, Figures 3c-3d are the plots of $a_{\boldsymbol{x}}{}^{3/2}\ p(\boldsymbol{\mu}_{(\boldsymbol{x})})\prod_j \sigma_{j(\boldsymbol{x})}$ and $a_{\boldsymbol{x}}{}^{3/2}\exp(-L_{\boldsymbol{x}}/\beta)$ in Eq. 21, showing high correlations around 0.91. This strongly supports our theoretical probability estimation which considers the metric space.

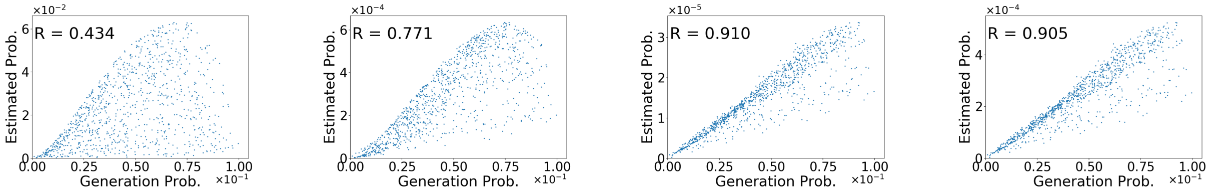
Appendix D also shows results using square error loss. The correlation coefficient for $\exp(-L_{\boldsymbol{x}}/\beta)$ also gives a high

Table 1. Property measurements of the toy dataset trained with the square error loss.

variable		z_1	z_2	z_3
$\frac{2}{\beta}{\sigma_j}^2 D'_j$	Av.	0.965	0.925	0.972
	SD	0.054	0.164	0.098
$D'_j(\boldsymbol{z})$	Av.	0.162	0.726	2.922
	SD	0.040	0.466	1.738
${\sigma_{j(\boldsymbol{x})}}^{-2}$	Av.	3.33e1	1.46e2	5.89e2
(Ratio)	Av.	1.000	4.39	17.69

Table 2. Property measurements of the toy dataset trained with the downward-convex loss.

variable		z_1	z_2	z_3
$\frac{2}{\beta}{\sigma_j}^2 D'_j$	Av.	0.964	0.928	0.978
	SD	0.060	0.160	0.088
$D'_j(\boldsymbol{z})$	Av.	0.161	0.696	2.695
	SD	0.063	0.483	1.573
${\sigma_{j(\boldsymbol{x})}}^{-2}$	Av.	3.30e1	1.40e2	5.43e2
(Ratio)	Av.	1.000	4.25	16.22



(a) $p(\boldsymbol{\mu}_{(\boldsymbol{x})})$ (b) $\exp(-L_x/\beta)$ (c) $a_{\boldsymbol{x}}^{3/2} p(\boldsymbol{\mu}_{(\boldsymbol{x})}) \prod_j \sigma_{j(\boldsymbol{x})}$ (d) $a_{\boldsymbol{x}}^{3/2} \exp(-L_x/\beta)$

Figure 3. Scattering plots of the data distribution (x-axis) versus four estimated probabilities (y-axes) for the downward-convex loss. y-axes are (a) $p(\boldsymbol{\mu}_{(\boldsymbol{x})})$, (b) $\exp(-L_x/\beta)$, (c) $a_{\boldsymbol{x}}^{3/2} p(\boldsymbol{\mu}_{(\boldsymbol{x})}) \prod_j \sigma_{j(\boldsymbol{x})}$, and (d) $a_{\boldsymbol{x}}^{3/2} \exp(-L_x/\beta)$.

score 0.904, since the input and metric spaces are equivalent.

Appendix D shows the exhaustive ablation study with different PDFs, losses, and β, which further supports our theory.

5.2. Evaluations in CelebA dataset

This section presents the disentanglement analysis using VAE for the CelebA dataset [1] (Liu et al., 2015). This dataset is composed of 202,599 celebrity facial images. In use, the images are center-cropped to form 64×64 sized images. As a reconstruction loss, we use SSIM which is close to subjective quality evaluation. The details of networks and training conditions are written in Appendix C.2.

Figure 4 shows the averages of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ in Eq.22 as the estimated variances, as well as the average and the standard deviation of $\frac{2}{\beta}{\sigma_{j(\boldsymbol{x})}}^2 D'_j(\boldsymbol{z})$ in Eq.23 as the estimated square norm of implicit transform. The latent variables z_i are numbered in descending order by the estimated variance. In the dimensions greater than the 27th, the averages of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ are close to 1 and that of $\frac{2}{\beta}{\sigma_{j(\boldsymbol{x})}}^2 D'_j(\boldsymbol{z})$ is close to 0, implying $D_{\mathrm{KL}}(\cdot) = 0$. Between the 1st and 26th dimensions, the mean and standard deviation of $\frac{2}{\beta}\sigma_{j(\boldsymbol{x})}{}^2 D'_j(\boldsymbol{z})$ averages are 1.83 and 0.13, respectively. This also implies the variance ${\sigma_{y_j(\boldsymbol{x})}}^2$ is around $1.83(\beta/2)$. These values seem almost constant with a small standard deviation; however, the mean is somewhat larger than the expected value 1. This suggests that the implicit embedding $\boldsymbol{y}'$ which satisfies $\mathrm{d}{y_j}'/\mathrm{d}\mu_{j(\boldsymbol{x})} = \sqrt{1.83(\beta/2)}/\sigma_{j(\boldsymbol{x})}$ can be considered as almost isometric. Thus, ${\sigma_{j(\boldsymbol{x})}}^{-2}$ averages still can determine the quantitative importance of each dimension.

We also train VAE using the decomposed loss explicitly, i.e., $L_{\boldsymbol{x}} = D(\boldsymbol{x}, \breve{\boldsymbol{x}}) + D(\breve{\boldsymbol{x}}, \hat{\boldsymbol{x}}) + \beta D_{\mathrm{KL}}(\cdot)$. Figure 5 shows the result. Here, the mean and standard deviation of $\frac{2}{\beta}{\sigma_{j(\boldsymbol{x})}}^2 D'_j(\boldsymbol{z})$ averages are 0.92 and 0.04, respectively, which suggests almost a unit norm. This result implies that the explicit use of decomposed loss promotes isometricity and allows for better analysis, as explained in Remark 1.

Figure 6 shows decoder outputs where the selected latent variables are traversed from -2 to 2 while setting the rest to 0. The average of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ is also shown there. The components are grouped by ${\sigma_{j(\boldsymbol{x})}}^{-2}$ averages, such that z_1, z_2, z_3 to the large, z_{16}, z_{17} to the medium, and z_{32} to the small, respectively. In the large group, significant changes of background brightness, face direction, and hair color are observed. In the medium group, we can see minor changes such as facial expressions. However, in the small group, there are almost no changes. In addition, Appendix E.1 shows the traversed outputs of all dimensional components in descending order of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ averages, where the degree of image changes clearly depends on ${\sigma_{j(\boldsymbol{x})}}^{-2}$ averages. Thus, it is strongly supported that the average of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ indicates the importance of each dimensional component like PCA.

5.3. Anomaly detection with realistic data

Using a vanilla VAE model with a single Gaussian prior, we finally examine the performance in anomaly detection

[1] (http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html)

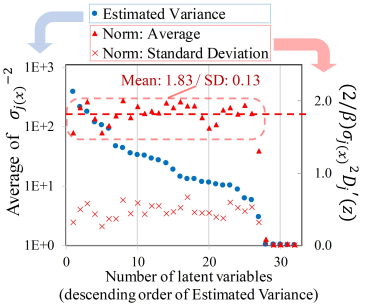


Figure 4. Graph of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ average and $\frac{2}{\beta}{\sigma_{j(\boldsymbol{x})}}^{2}D_j'(\boldsymbol{z})$ in VAE for CelebA dataset.

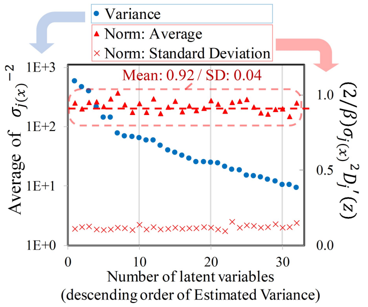


Figure 5. Graph of ${\sigma_{j(\boldsymbol{x})}}^{-2}$ average and $\frac{2}{\beta}{\sigma_{j(\boldsymbol{x})}}^{2}D_j'(\boldsymbol{z})$ in VAE for CelebA dataset with explicit decomposed loss.

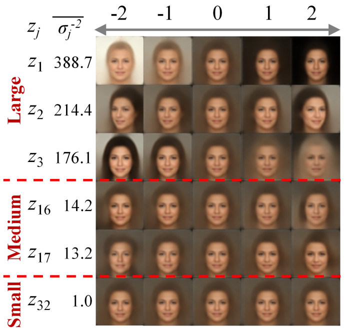


Figure 6. Dependency of decoded image changes with $z_j = -2$ to 2 on the average of ${\sigma_{j(\boldsymbol{x})}}^{-2}$.

Table 3. Average and standard deviations (in brackets) of F1

Dataset	Methods	F1
KDDCup	GMVAE*	0.9326
	DAGMM*	0.9500 (0.0052)
	RaDOGAGA(d)*	0.9624 (0.0038)
	RaDOGAGA(log(d))*	0.9638 (0.0042)
	vanilla VAE	**0.9642 (0.0007)**
Thyroid	GMVAE*	0.6353
	DAGMM*	0.4755 (0.0491)
	RaDOGAGA(d)*	0.6447 (0.0486)
	RaDOGAGA(log(d))*	**0.6702 (0.0585)**
	vanilla VAE	0.6596 (0.0436)
Arrythmia	GMVAE*	0.4308
	DAGMM*	0.5060 (0.0395)
	RaDOGAGA(d)*	**0.5433 (0.0468)**
	RaDOGAGA(log(d))*	0.5373 (0.0411)
	vanilla VAE	0.4985 (0.0412)
KDDCup-rev	DAGMM*	0.9779 (0.0018)
	RaDOGAGA(d)*	0.9797 (0.0015)
	RaDOGAGA(log(d))*	0.9865 (0.0009)
	vanilla VAE	**0.9880 (0.0008)**

in which PDF estimation is the key issue. We use four public datasets[‡]: KDDCUP99, Thyroid, Arrhythmia, and KDDCUP-Rev. The details of the datasets and network configurations are given in Appendix H. For a fair comparison with previous works, we follow the setting in Zong et al. (2018). Randomly extracted 50% of the data were assigned to the training and the rest to the testing. Then the model is trained using normal data only. Here, we use the explicit decomposed loss to promote isometricity. The coding loss is set to SSE. For the test, the anomaly score for each sample is set to $L_{\boldsymbol{x}}$ in Eq. 3 after training since $-L_{\boldsymbol{x}}/\beta$ gives a log-likelihood of the input data from Proposition 1. Then, samples with anomaly scores above the threshold are identified as anomalies. The threshold is given by the ratio of the anomaly data in each data set. For instance, in KDDCup99, data with $L_{\boldsymbol{x}}$ in the top 20 % is detected as an anomaly. We run experiments 20 times for each dataset split by 20 different random seeds.

5.3.1. BASELINE METHODS

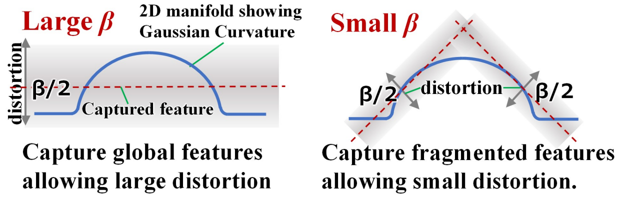
We compare previous methods such as GMVAE (Liao et al., 2018), DAGMM (Zong et al., 2018), and RaDOGAGA (Kato et al., 2020) that conducted the same experiments. All of them apply GMM as a prior because they believe GMM is more appropriate to capture the complex data distribution than VAE with a single Gaussian prior.

5.3.2. RESULTS

Table 3 reports the average F1 scores and standard deviations (in brackets). Recall and precision are shown in Appendix H. Liao et al. (2018) insisted that the vanilla VAE is not appropriate for PDF estimation. Contrary to their claim, by considering the quantitative property as proven in this paper, even a vanilla VAE achieves state-of-the-art performance in KDDCup99 and KDDCup-rev. In other data sets, the score of VAE is comparable with RaDOGAGA, which is the previous best method. Here, RaDOGAGA attempts to adapt the parametric distribution such as GMM to the input distribution in the isometric space. However, fitting sufficiency is strongly dependent on the capability of the parametric distribution. By contrast, VAE can flexibly adapt a simple prior distribution to the input distribution via trainable posterior variance $\sigma_{j(\boldsymbol{x})}$. As a result, VAE can provide a simpler tool for estimating the data distribution.

[‡] Datasets can be downloaded at `https://kdd.ics.uci.edu/` and `http://odds.cs.stonybrook.edu`.

[*] Scores are cited from Liao et al. (2018) (GMVAE) and Kato et al. (2020)(DAGMM, RaDOGAGA)


Large β
2D manifold showing Gaussian Curvature
distortion
β/2
Captured feature
Capture global features allowing large distortion
Small β
β/2
distortion
β/2
Capture fragmented features allowing small distortion.

(a) Features for large β

(b) Features for small β

Figure 7. Conceptual explanation of captured features in the implicit isometric space for 2D manifold with non-zero Gaussian curvature.

6. Relation with previous studies

First of all, we show VAE can be interpreted as a Rate-distortion (RD) optimal encoder based on RD theory (Berger, 1971), which has been successfully applied to image compression in the industry. The optimal transform coding (Goyal, 2001) for the Gaussian data with SSE metric is formulated as follows: First, the data are transformed deterministically using the orthonormal transform (orthogonal and unit norm) with a PCA basis. Note that the orthonormal transform is a part of the isometric embedding where the encoder is restricted as linear. Then, the transformed data is entropy-coded. Here, the key point for optimizing RD is to introduce equivalent stochastic distortion in all dimensions (or to use a uniform quantizer for image compression). Then the rate R_{opt} at the optimum condition is derived as follows: $\boldsymbol{z} \in Z$ denotes transformed data from inputs. Let z_j be the j-th dimensional component of $\boldsymbol{z}$. ${\sigma_{zj}}^2$ denotes a variance of z_j in a dataset. Note that ${\sigma_{zj}}^2$ is equivalent to the eigenvalue of PCA in each dimension. Let ${\sigma_d}^2$ be a distortion equally allowed in each dimensional channel. Assume the input dimension is m and ${\sigma_d}^2$ is smaller than ${\sigma_{zj}}^2$ for all j. Then, R_{opt} is derived as:

$$
\begin{aligned}
R_{\text{opt}} &= \sum_{j=1}^{m} \left(H(\mathcal{N}(z_j; 0, {\sigma_{zj}}^2) - H(\mathcal{N}(z_j; 0, {\sigma_d}^2)\right) \\
&= H(Z) - H(0, {\sigma_d}^2\, \boldsymbol{I}_m). \qquad (25)
\end{aligned}
$$

Here, if ${\sigma_d}^2$ is set to $\beta/2$, Eq. 8 is equivalent to Eq. 25. This suggests that VAE can be considered as a rate-distortion optimal encoder where RD theory is extended from linear orthonormal transform to general isometric embedding in the given metric. More details are described in Appendix B.6.

Next, our theory can intuitively explain how the captured features in β-VAE behave when varying β. Higgins et al. (2017) suggests that β-VAE with large β can capture a global features while degrading the reconstruction quality. Our intuitive explanation is as follows: Assume the case of 2D manifold in 3D space. According to Gauss's Theorema Egregium, the Gaussian curvature is an intrinsic invariant of a 2D surface and its value is unchanged after any isometric embeddings (Andrews, 2002). Figure 7 shows the conceptual explanation of captured features in the implicit isometric space for 2D manifold with non-zero Gaussian curvature. Our theory shows that $\beta/2$ is considered as the allowable distortion in each dimensional component of implicit isometric embedding. If β is large as shown in Fig. 7a, β-VAE can capture global features in the implicit isometric space allowing large distortion with lower rate. If β is small as shown in Fig. 7b, by contrast, β-VAE will capture only fragmented features allowing small distortion with higher rate. We believe similar behaviors occur in general higher-dimensional manifolds.

Finally, we correct the analysis in Alemi et al. (2018). They describe ”the ELBO objective alone cannot distinguish between models that make no use of the latent variable versus models that make large use of the latent variable and learn useful representations for reconstruction,” because the reconstruction loss and KL divergence have unstable values after training. From this reason, they introduce a new objective $D(\boldsymbol{x}, \hat{\boldsymbol{z}}) + |D_{\text{KL}}(\cdot) - \sigma|$ to fix this instability using a target rate σ. Correctly, the reconstruction loss and KL divergence are stably derived as a function of β as shown in Appendix B.1 and B.4.

Our theory can further explain the analysis results of related prior works such as Higgins et al. (2017); Alemi et al. (2018); Dai et al. (2018); Dai & Wipf (2019), and Tishby et al. (1999). The details are described in Appendix B.

7. Conclusion

This paper provides a quantitative understanding of VAE by non-linear mapping to an isometric embedding. According to the Rate-distortion theory, the optimal transform coding is achieved by using orthonormal transform with a PCA basis, where the transform space is isometric to the input. From this analogy, we show theoretically and experimentally that VAE can be mapped to an implicit isometric embedding with a scale factor derived from the posterior parameter. Based on this property, we also clarify that VAE can provide a practical quantitative analysis of input data such as the probability estimation in the input space and the PCA-like quantitative multivariate analysis. We believe the quantitative properties thoroughly uncovered in this paper will be a milestone to further advance the information theory-based generative models such as VAE in the right direction.

Acknowledgement

We express our gratitude to Tomotake Sasaki and Takashi Katoh for improving the clarity of the manuscript. Taiji Suzuki was partially supported by JSPS KAKENHI (18H03201, and 20H00576), and JST CREST.

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In *Proceedings of the 35th International Conference on Machine Learning(ICML)*, pp. 159–168. PMLR, July 2018.

Andrews, B. Notes on the isometric embedding problem and the nash-moser implicit function theorem. *Proceedings of the Centre for Mathematics and its Applications*, 20: 157–208, January 2002.

Ballé, J., Valero, L., and Eero P., S. Density modeling of images using a generalized normalization transformation. In *Proceedings of the 4t International Conference on Learning Representations (ICLR)*, May 2016.

Ballé, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. Variational image compression with a scale hyperprior. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, April 2018.

Berger, T. (ed.). *Rate Distortion Theory: A Mathematical Basis for Data Compression*. Prentice Hall, 1971. ISBN 0137531036.

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006. ISBN 978-0387-31073-2.

Dai, B. and Wipf, D. Diagnosing and enhancing vae models. In *International Conference on Learning Representations (ICLR)*, May 2019.

Dai, B., Wang, Y., Aston, J., Hua, G., and Wipf, D. Hidden talents of the variational autoencoder. *The Journal of Machine Learning Research*, 19:1573–1614, January 2018.

Dua, D. and Graff, C. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2019.

Goyal, V. K. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18:9–21, September 2001.

Han, Q. and Hong, J.-X. *Isometric Embedding of Riemannian Manifolds in Euclidean Spaces*. American Mathematical Society, 2006. ISBN 0821840711.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, April 2017.

Huang, S., Makhzani, A., Cao, Y., and Grosse, R. Evaluating lossy compression rates of deep generative models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, July 2020.

Kato, K., Zhou, Z., Sasaki, T., and Nakagawa, A. Rate-distortion optimization guided autoencoder for generative analysis. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, July 2020.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, Banff, Canada, April 2014.

Kumar, A. and Poole, B. On implicit regularization in β-VAEs. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, July 2020.

Liao, W., Guo, Y., Chen, X., and Li, P. A unified unsupervised gaussian mixture variational autoencoder for high dimensional outlier detection. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1208–1217. IEEE, 2018.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, pp. 4114–4124. PMLR, June 2019.

Pearlman, W. A. and Said, A. *Digital Signal Compression: Principles and Practice*. Cambridge University Press, 2011. ISBN 0521899826.

Rao, K. R. and Yip, P. (eds.). *The Transform and Data Compression Handbook*. CRC Press, Inc., Boca Raton, FL, USA, 2000. ISBN 0849336929.

Rolínek, M., Zietlow, D., and Martius, G. Variational autoencoders pursue pca directions (by accident). In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Sullivan, G. J. and Wiegand, T. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15:74–90, November 1998.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. In *The 37th annual Allerton Conference on Communication, Control, and Computing*, September 1999.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing*, 13:600–612, April 2001.

Wiener, N. *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. The MIT Press, 1964. ISBN 978-0-262-73005-1.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., and Chen, H. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *In International Conference on Learning Representations*, 2018.