Understanding VAEs in Fisher-Shannon Plane

Huangjie Zheng,[1] **Jiangchao Yao,**[1,2] **Ya Zhang,**[1]✉ **Ivor W. Tsang,**[2] **Jia Wang**[1]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, [2]University of Technology Sydney
{zhj865265, sunarker, ya_zhang, jiawang}@sjtu.edu.cn, ivor.tsang@uts.edu.au

Abstract

In information theory, Fisher information and Shannon information (entropy) are respectively used to quantify the uncertainty associated with the distribution modeling and the uncertainty in specifying the outcome of given variables. These two quantities are complementary and are jointly applied to information behavior analysis in most cases. The uncertainty property in information asserts a fundamental trade-off between Fisher information and Shannon information, which enlightens us the relationship between the encoder and the decoder in *variational auto-encoders* (VAEs). In this paper, we investigate VAEs in the Fisher-Shannon plane, and demonstrate that the representation learning and the log-likelihood estimation are intrinsically related to these two information quantities. Through extensive qualitative and quantitative experiments, we provide with a better comprehension of VAEs in tasks such as high-resolution reconstruction, and representation learning in the perspective of Fisher information and Shannon information. We further propose a variant of VAEs, termed as Fisher auto-encoder (FAE), for practical needs to balance Fisher information and Shannon information. Our experimental results have demonstrated its promise in improving the reconstruction accuracy and avoiding the non-informative latent code as occurred in previous works.

Introduction

The common latent variable models fit $p_\theta(x,z) = p_\theta(z)p_\theta(x|z)$ in order to model the data x with a latent variable z as representation. Variational Autoencoders (VAEs) (Kingma and Welling 2014), recently as one of the most popular latent variable models, maximize the evidence lower bound (ELBO) in an encoding/decoding mechanism.

$$\mathcal{L} = \mathop{\mathbf{E}}_{x\sim data}\left[\mathop{\mathbf{E}}_{z\sim q_\phi}\left[\log p_\theta(x|z)\right] - \mathcal{D}_{\mathrm{KL}}\left(q_\phi(z|x)||p(z)\right)\right]$$

where $p_\theta(x|z)$ and $q_\phi(z|x)$ are encoder and decoder implemented with neural networks parameterized by θ, ϕ; $\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence. The learning targets of VAEs may be interpreted in the following two perspectives. In the perspective of variational optimization (Cremer, Li, and Duvenaud 2018), VAEs aim to learn a proper model by maximum likelihood with $p_\theta(x|z)$. In the perspective of representation learning (Bengio, Courville, and Vincent 2013), VAEs target to learn latent codes that are sufficiently encoded with information about the input by $q_\phi(z|x)$. Several variants of VAEs (Van Oord, Kalchbrenner, and Kavukcuoglu 2016; Van den Oord et al. 2016; Chen et al. 2017; Zhao, Song, and Ermon 2017) have been developed based on above two perspectives in recent years.

In recent studies, it has been reported that VAEs are difficult to balance between the representation learning and likelihood maximizing. For instance, the latent variable generated by the encoder is approximately ignored when the decoder is too expressive (Bowman et al. 2015). As discussed in (Alemi et al. 2018), the evidence lower bound (ELBO) lacks a measure in the quality of the representation, since the KL divergence $\mathcal{D}_{\mathrm{KL}}\left(q_\phi(z|x)||p(z)\right)$ only controls the way VAE encodes a representation. Several studies in VAEs (Chen et al. 2017; Zhao, Song, and Ermon 2017) have attempted to improve the balance between the learned representation quality and ELBO maximization on the basis of information theory. A Shannon entropy-based constraint is introduced to assess the quality of representation learning when optimizing the ELBO, so as to guarantee that sufficient information of the observation flows into the latent code. For example, (Phuong et al. 2018) optimized ELBO plus a mutual information regularizer to explicitly control the information stored in latent codes. Although it is useful to consider mutual information (a member of Shannon family) in VAE encoding for representation learning, how these information quantities affect VAEs has not yet been theoretically analyzed so far. Besides, previous works mainly leverage the Shannon information, which usually suffers from the intractability in computing, yielding an approximation surrogate (Phuong et al. 2018).

In information theory, the uncertainty property (Dembo, Cover, and Thomas 1991; Vignat and Bercher 2003) has revealed a trade-off between Fisher information and Shannon information, which quantify the uncertainty associated with distribution modeling and the entropy in the predicted values of variables respectively (Rosso, Olivares, and Plastino 2015). Fisher-Shannon (FS) information plane (Vignat and Bercher 2003) is proposed to analyze the complementarity between Fisher information and Shannon information. In this paper, based on the uncertainty property, we attempt to investigate VAEs in FS information plane, which illustrates a novel insight to perform a theoretical analysis with

VAEs and show that representation learning with latent variable models via Maximum likelihood estimation is intrinsically related to the trade-off between Fisher information and Shannon entropy. Based on the above findings, we propose a family of VAEs regularized by Fisher information, named *Fisher auto-encoder* (FAE), to control the information quality during encoding/decoding. Finally, we perform a range of experiments to analyze a variety of VAE models in the FS information plane to empirically validate our findings. In addition, FAE is shown to not only provide a novel insight in the information trade-off, but can also improve the reconstruction accuracy and the representation learning.

Related work

Information uncertainty: Fisher information and Shannon information are considered important tools to describe the informational behavior in information systems respectively in the distribution modeling view and in the variable view (Brunel and Nadal 1998). The generalization of Heisenberg Uncertainty Principle (Schrödinger 1930) into information system demonstrates that Fisher information and Shannon information are intrinsically linked, with the uncertainty property, where higher Fisher information will result in lower Shannon information and vice versa (Dembo, Cover, and Thomas 1991). With this property, Fisher information and Shannon information are considered complementary aspects and be widely used in solving dual problem when one aspect is intractable (Martin, Pennini, and Plastino 1999). To better take advantage of this property, (Vignat and Bercher 2003) construct the Fisher-Shannon information plane for signal analysis in joint view of Fisher-Shannon information.

VAEs in information perspective: Variational autoencoders (Kingma and Welling 2014), with a auto-encoding form, can be regarded as an information system serves two goals. On one hand, we expect a proper distribution estimation to maximize the marginal likelihood; on the other hand, we hope the latent code can provide sufficient information of data point so as to serve downstream tasks. To improve the log-likelihood estimation, several works, such as PixelCNN and PixelRNN (Van Oord, Kalchbrenner, and Kavukcuoglu 2016; Van den Oord et al. 2016) model the dependencies among pixels with autoregressive structure to achieve an expressive density estimator. As for the latent code, plenty of works address it in the perspective of Shannon information (Chen et al. 2017). Mutual information, an member of Shannon family, is applied to measure the mutual dependence between datapoint x and latent code z (Zhao, Song, and Ermon 2017; Alemi et al. 2018). The leverage of mutual information is achieved with Maximum-Mean Discrepancy (Zhao, Song, and Ermon 2017) and Wasserstein distance (Tolstikhin et al. 2018). More generally, (Phuong et al. 2018) regularize the mutual information in VAE's objective to control the information in latent code.

Effects of Fisher information and Shannon information: Fisher information and Shannon information (typically we call entropy), as complementary aspects, possess their properties. In (Rosso, Olivares, and Plastino 2015), the entropy is explained as a measure of "global character" that is invariant to strong changes in the distribution, while the Fisher information is interpreted as a measure of the ability to model a distribution, which corresponds to the "local" characteristics. The characteristics of these two sides have been taken into advantages of several existing works, *e.g.*, entropy has been introduced in improving deep neural networks on tasks like classification (Silva, de Sá, and Alexandre 2005; Saxe et al. 2018); FI has been introduced to to evaluate neural network's parameter estimation (Tu et al. 2016; Desjardins et al. 2015). To better understand how these two perspectives affect the mechanism of VAEs, we study VAEs in the Fisher-Shannon information plane to provide a complete understanding of VAEs in a joint view of Fisher information and Shannon information.

Fisher-Shannon Information Plane

In this section, we first present the information uncertainty property to link the Fisher information and the Shannon information (Dembo, Cover, and Thomas 1991) of the random variable. After that, a Fisher-Shannon information plane (Vignat and Bercher 2003) is constructed to jointly analyze the characteristics of the random variable on its distribution. This provides the simple basics to understand VAEs with the Fisher-Shannon information plane.

Fisher-Shannon Information Uncertainty property

In information theory, considering a random variable X, whose probability density function is denoted as $f(x)$, the Fisher information[1] for X and its Shannon entropy can be formulated as:

$$\begin{aligned} &\text{Fisher Information:} \quad \mathcal{J}(X) = \int_x \left(\frac{\partial}{\partial x} f(x)\right)^2 \frac{dx}{f(x)} \\ &\text{Shannon Entropy:} \quad \mathcal{H}(X) = -\int_x f(x) \log f(x) dx \end{aligned} \tag{1}$$

Above two information quantities are respectively related to the precision that the model fits in observed data and the uncertainty in the outcome of the random variable. For convenience to use in deduction, Shannon entropy is usually transformed to the following quantity which is called the entropy power (Stam 1959):

$$\mathcal{N}(X) = \frac{\exp(2\mathcal{H}(X))}{2\pi \exp(1)} \tag{2}$$

The measure of $\mathcal{N}(X)$ and $\mathcal{J}(X)$ verifies a set of resembling inequalities in information theory (Stam 1959). Specifically, one of the inequalities connecting the two quantities and being tightly related to the phenomena studied in VAEs, is the uncertainty inequality (Dembo, Cover, and Thomas 1991), which is formulated as:

$$\mathcal{N}(X)\mathcal{J}(X) \geq 1 \tag{3}$$

[1]Note that, we follow the non-parametric Fisher information definition that differentiates on random variables, which can be transformed with a translation of parameter from parametric version (Stam 1959).

where the equality holds when the random variable X is a Gaussian variable. Note that this inequality possesses several versions; in the case of a random vector $\boldsymbol{\mathcal{X}} = (X_1, X_2, ..., X_n)$, the corresponding Fisher information turns into a $n \times n$ dimensional Fisher information matrix and we need to compute the trace of this matrix $tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}}))$ (see (Dembo 1990)) and we have $\mathcal{N}(\boldsymbol{\mathcal{X}}) \cdot tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}})) \geq n$.

When the distribution of given variable is fixed, the product of the Fisher information $tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}}))$ and the entropy power $\mathcal{N}(\boldsymbol{\mathcal{X}})$ is a constant that is greater or equal to 1, which depends on the distribution form, the dimension of the random vector, *etc.* We can further formulate this property as follow:

$$\mathcal{N}(\boldsymbol{\mathcal{X}}) \cdot tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}})) = K \tag{4}$$

where K is a constant number and $K \geq 1$. Eq. (3) and (4) indicate the measure of Fisher information and Shannon information exists a trade-off between these two quantities.

Fisher-Shannon Information Plane

To facilitate the analysis of above two information quantities together, an information plane based on the Fisher information and the Shannon entropy power is proposed in (Vignat and Bercher 2003) and we generalize it as follows,

$$\begin{aligned}\mathcal{D} = \{ (\mathcal{N}(\boldsymbol{\mathcal{X}}), tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}}))) \,|\, \mathcal{N}(\boldsymbol{\mathcal{X}}) \geq 0, tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}})) \geq 0 \text{ and} \\ \mathcal{N}(\boldsymbol{\mathcal{X}}) \cdot tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}})) \geq 1\}.\end{aligned} \tag{5}$$

where $\mathcal{D}$ denotes a region $\subset \mathbb{R}^2$, which is limited by the Gaussian case. This plane consists of several Fisher-Shannon (FS) curves $\mathcal{N}(\boldsymbol{\mathcal{X}}) \cdot tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{X}})) = K$, which characterizes the random variable with different distributions.

As discussed in (Alemi et al. 2018), the quality of latent variable is hard to measure in maximizing ELBO, and various VAEs, like (Chen et al. 2017; Higgins et al. 2017) have been proposed to balance the trade-off between representation learning and optimization. In the FS plane, different VAEs can be analyzed jointly with Fisher information and Shannon information. By observing their location in FS plane, we can identify the characteristic of this VAE model.

In addition, from the uncertainty property between Fisher information and Shannon information, these two quantities are shown tightly connected. As shown in Eq. (4), when the distribution of given random variable is fixed, the Fisher information and Shannon entropy power's product is a constant, where the trade-off exists. We can take advantage of this trade-off to avoid the intractability in Shannon information computing. In this paper, we propose a family of VAEs that control the Fisher information, named Fisher Auto-Encoder (FAE), which allows a more accurate description in situations where the Shannon information shows limited dynamics (Martin, Pennini, and Plastino 1999) in VAEs. The details of FAE will be discussed in the next section.

The Fisher Auto-Encoder

As shown in the previous section, one can apply either Fisher information or Shannon entropy power to control the trade-off between the likelihood estimation $p(x)$ and the dependence between data x and the latent code z. In this section, we come up with a family of VAEs that takes advantage of the Fisher information, named Fisher Autoencoder (FAE), and analyze its characteristics in this section.

Fisher Information Control in VAE

The Fisher AutoEncoder aims to to control the Fisher information quantity in the objective. Thus, the objective becomes to maximize the evidence lower bound (ELBO) with constraint of Fisher information and we reformulate the VAE's objective as follows:

$$\begin{aligned}\max_{\theta,\phi} \quad & \mathop{\mathbf{E}}_{x\sim p}\left[\mathop{\mathbf{E}}_{z\sim q_\phi}\left[\log p(x|z,\theta)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right)\right] \\ \text{s.t.} \quad & tr(\boldsymbol{\mathcal{J}})(x) = F_x, \quad tr(\boldsymbol{\mathcal{J}})(z) = F_z\end{aligned} \tag{6}$$

where $\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence (Joyce 2011); F_x and F_z are positive constant that denote the desired Fisher information value. A large value of F_x (*resp.* F_z) implies we favor a precise distribution estimation in the model parameterized by θ (*resp.* ϕ); while a low value of F_x (*resp.* F_z) indicates we weaken the distribution modeling to increase the Shannon entropy power.

To solve the scenario described in Eq. (6), we transfer this optimization problem into a Lagrangian objective, formulated as:

$$\begin{aligned}\mathcal{L}_F(\theta,\phi) = & \underbrace{\mathop{\mathbf{E}}_{x\sim p}\left[\mathop{\mathbf{E}}_{z\sim q_\phi}\left[\log p(x|z,\theta)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right)\right]}_{\text{ELBO}} \\ & \underbrace{- \lambda_z\left|tr(\boldsymbol{\mathcal{J}})(z) - F_z\right|}_{\text{FI control in encoder}} \underbrace{- \lambda_x\left|tr(\boldsymbol{\mathcal{J}})(x) - F_x\right|}_{\text{FI control in decoder}}\end{aligned} \tag{7}$$

Now the objective consists of three parts, the ELBO to maximize, and two Fisher information regularizers in encoder and decoder, where λ_z and λ_x are positive constant that control the regularizers. With this objective, we can control the Fisher information in encoder/decoder with an expected desired value $F_{z/x}$. In the most cases, the calculation of Fisher information is not difficult. We can estimate the Fisher information directly by its definition.

Characteristic of FAE: an example of FI regularization in Gaussian encoder

Here we give a FAE exemplar that only controls the Fisher information in encoder, which means we set λ_x in Eq. (7) as zero. In this model, we assume that all random variables are of dimension 1 (*i.e.* in the scalar case) for simplicity in presentation. The FAE objective is formulated as:

$$\begin{aligned}\mathcal{L} = & \mathop{\mathbf{E}}_{x\sim p}\left[\mathop{\mathbf{E}}_{z\sim q_\phi}\left[\log p(x|z,\theta)\right]\right] \\ & \underbrace{- \mathop{\mathbf{E}}_{x\sim p}\left[\mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right)\right] - \lambda_z\left|\mathcal{J}(z) - F_z\right|}_{\mathcal{R}(\phi,\theta)}\end{aligned} \tag{8}$$

This objective consists of a reconstruction term and a generalized regularizor $\mathcal{R}(\phi, \theta)$ that considering the Fisher information other than KL divergence. Same to the VAE (Kingma and Welling 2014), both prior distribution $p_\theta(z) = \mathcal{N}(0, 1)$ and posterior approximation $q_\phi(z|x)$ are Gaussian, thus the KL-divergence can be analytically computed as:

$$-\mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right) = \frac{1}{2}\left(1 + \log((\sigma)^2)\right) - (\mu)^2 - (\sigma)^2 \quad (9)$$

where μ and σ respectively correspond to the mean and standard derivation of a Gaussian distribution. The Fisher information can be easily computed by definition:

$$\mathcal{J}(z|x) = \int_z \left(\frac{\partial}{\partial z} q(z|x)\right)^2 \frac{dz}{q(z|x)} = \frac{1}{\sigma^2(x)} = \frac{1}{\sigma^2}. \quad (10)$$

Finally, putting Eq. (9) and Eq. (10) together, we have the following regularizer $\mathcal{R}(\phi, \theta)$,

$$\begin{aligned}\mathcal{R}(\phi,\theta) &= -\mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right) - \lambda_z\left|\mathcal{J}(z) - F_z\right| \\ &= \frac{1}{2}\left(\left(1 + \log((\sigma)^2)\right) - (\mu)^2 - (\sigma)^2\right) - \lambda_z\left|\frac{1}{\sigma^2} - F_z\right|\end{aligned} \quad (11)$$

Considering the KL-divergence term in the original VAEs (Kingma and Welling 2014), the optimal is reached at $\sigma^2 = 1$, which aligns the posterior $q_\phi(z|x)$ to a normal distribution $\mathcal{N}(0, 1)$. However, in Eq. (11), we can observe that the variance is also penalized by the desired Fisher information value F_z, which will push the variance to approach zero when F_z is large or make the variance larger than 1 when F_z is picked as a small value.

In the above discussion, we analyze the characteristics of FAE in variance control. This property corresponds to the inequality of Cramer-Rao, from which the uncertainty principle shown in Eq. (3) can be derive (Dembo, Cover, and Thomas 1991). Given a stochastic variable X of mean μ and variance σ^2, the Fisher information is the lower bound of the variance in a non-biased estimation:

$$\sigma_X^2 \geq \frac{1}{\mathcal{J}(X)}, \quad (12)$$

the equality holds if and only if X is Gaussian. This inequality gives us the first impression of the characteristic of Fisher information: When FI is in a low value, the variance of the estimation is forced to be high, causing larger uncertainty of the model estimation. Thus, we need to enlarge the FI to make the variance more controllable.

Connection to the Mutual Auto-Encoder

In this section, we demonstrate the connection between the Fisher Auto-Encoder and the Mutual Auto-Encoder (MAE) (Phuong et al. 2018), which is representative in the family of VAEs that leverage the Shannon information.

As discussed in the previous section, the product of Fisher information and entropy power is a constant when the distribution of variable is fixed, as shown in Eq. (4). We can derive:

$$\begin{aligned} & log(\mathcal{N}(Z|X)) = log(K) - log(\boldsymbol{\mathcal{J}}(Z|X)) \\ \Longleftrightarrow\ & log(\boldsymbol{\mathcal{J}}(Z|X)) = -2\mathcal{H}(Z|X) + constant. \end{aligned} \quad (13)$$

where the FI regularizor in FAE is equivalent to a regularizor of the conditional entropy $\mathcal{H}(Z|X)$.

Looking back into the MAE proposed in (Phuong et al. 2018), this model controls the mutual information between latent variable z and data x as follows:

$$\mathcal{L} = \underbrace{\mathop{\mathbf{E}}_{x\sim p}\left[\mathop{\mathbf{E}}_{z\sim q_\phi}\left[\log p(x|z,\theta)\right] - \mathcal{D}_{\mathrm{KL}}\left(q(z|x,\phi)||p(z)\right)\right]}_{\textbf{ELBO}} - \underbrace{C\left|\mathcal{I}(x,z) - M\right|}_{\text{MI regularizor}}$$

where C and M are positive constants that respectively control the information regularization and the desired mutual information quantity. Since the mutual information is difficult to compute directly, the mutual information $\mathcal{I}(x, z)$ is inferred using Gibbs inequality (Barber and Agakov 2003):

$$\begin{aligned}\widehat{\mathcal{I}}(x,z) &= \mathcal{H}(z) - \mathcal{H}(z|x) \\ &\geq \mathcal{H}(z) + \mathop{\mathbf{E}}_{x,z\sim p}\left[\log r_\omega(z|x)\right]\end{aligned} \quad (14)$$

where $r_\omega(z|x)$ is a parametric distribution that can be modeled by a network. The objective is to maximizing $\mathbf{E}_{x,z}\left[\log r_\omega(z|x)\right]$ in Eq. (14) with the constraint M. Thus, MAE intrinsically controls the mutual information by controlling the conditional entropy:

$$\begin{aligned}\mathcal{I}(x,z) &= \mathcal{H}(x) - \mathcal{H}(x|z) \\ &= \mathcal{H}(z) - \mathcal{H}(z|x).\end{aligned} \quad (15)$$

with M of large value, the conditional entropy $\mathcal{H}(Z|X)$ can minimized more to obtain a larger mutual information, and vice versa. FAE can also set constraint F to control the conditional entropy $\mathcal{H}(Z|X)$ (or $\mathcal{H}(X|Z)$). As Eq. (13) shows, using Fisher information, the FI regularizers are equivalent to the regularizers of the conditional entropy; thus the mutual information between X and Z can also be assessed without derive approximative upper or lower bounds. FAE can thus implicitly control the mutual information $\mathcal{I}(X, Z)$ by setting proper Fisher information constraint F.

Experiments

In this section, we perform a range of experiments to investigate the Fisher-Shannon impacts in VAEs. Meanwhile, we expose how the Fisher Auto-Encoder can improve VAEs in encoding/decoding with the Fisher information constraint.

Experiment Goals and Experimental Settings

As discussed, the entropy power and Fisher Information corresponds to different characteristics. Thus, we aim to explore these characteristics and give corresponding analysis in order to give a better understanding of existing VAE variants. Some specific goals are summarized as:

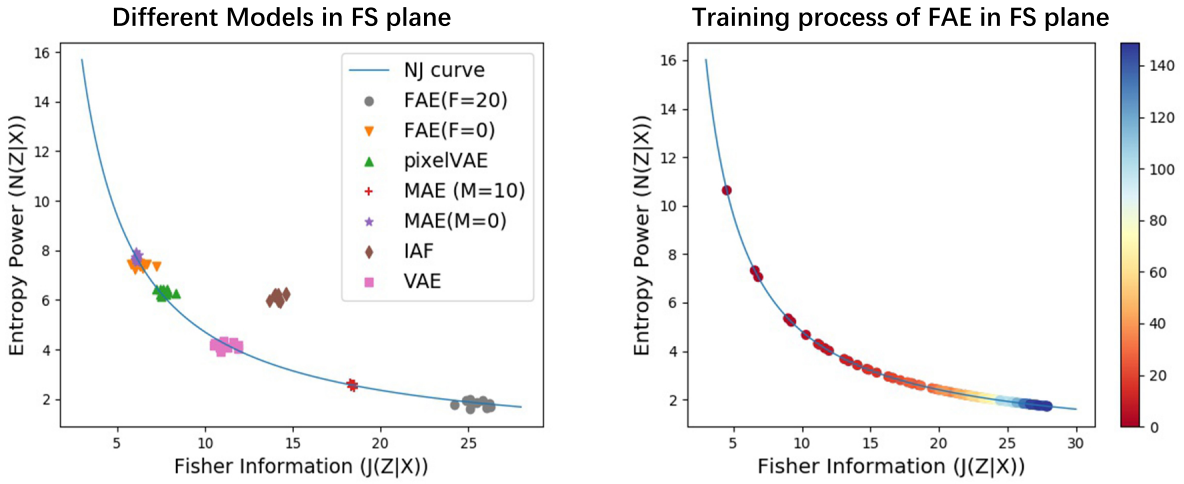


Figure 1: (Best view in color) Visualization of VAE models in the Fisher-Shannon Plane (**left**) and visualization of FAE's training process in the plane. (**right**, the color bar indicates the training) epoch.

- Explore several variants of VAEs models' characteristics in the FS plane.
- Explore the characteristics of latent code *w.r.t.* the Fisher Information and entropy power.
- Discuss the effect of different FI constraint in FAE.

The experiments are conducted on the MNIST dataset (Lecun et al. 1998) and the SVHN dataset (Netzer et al. 2011). The first dataset consists of ten categories of 28×28 hand-written digits and is binarized as in (Larochelle and Murray 2011). We follow the original partition to split the data as 50,000/10,000/10,000 for the training, validation and test. For SVHN dataset , it is MNIST-like but consists of 32×32 color images. We apply 73257 digits for training, 26032 digits for testing, and 20000 samples from the additional set (Netzer et al. 2011) for validation.

In FAE and its baselines, all random variables are supposed to be Gaussian variables. Here we only concern the hyper-parameters λ_z to adjust the constraint of Fisher information in encoding. In practice, we observe this value can be effective when set from 0.01 to 10 (depends on dataset and tasks). For the architecture of inference network and generative network, we both deploy a 5-layers network. Since the impacts of fully-connected and convolution architecture do not differ much in the experiments, we here present results using the architecture as 5 full-connected layers of dimension 300. The latent code is of dimension 40.

Quantitative Results

Fisher-Shannon Plane Analysis In this part, we conduct a series of experiments on different models to evaluate them in Fisher-Shannon plane to present different characteristics of using Fisher information and entropy power.

We first evaluate the test log-likelihood. To compute the test marginal negative log-likelihood (NLL), we apply the Importance Sampling (Burda, Grosse, and Salakhutdinov 2015) with 5,000 data points for the previously mentioned models. We select the most representative average results from extensive hyper-parameter configuration and expose them in Table 1: when the Fisher information constraint of $q_{z|x}$ in FAE ($F_z = 20$) (or the mutual information constraint between data x and latent variable z in MAE $M = 10$) is large, the models can achieve a competitive results of state-of-the-arts like pixelVAE (Van den Oord et al. 2016) and Inverse Autoregressive Flow(IAF) (Kingma et al. 2016). When set the information constraint F or M to zero, we can observe that the results are comparable to the plain VAE, but less competitive than the former models.

Table 1: Test negative log-likelihood (NLL) estimates for different models on MNIST

Model	Test NLL	Model	Test NLL
VAE	85.56	FAE(F=0)	83.2
PixelVAE	79.21	FAE(F=20)	79.30
IAF	79.85	MAE(M=0)	81.58
		MAE(M=10)	80.86

We put the former models in the FS information plane and draw the "NJ curve" for the Gaussian variable (where $\mathcal{N}(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}}) \cdot tr(\boldsymbol{\mathcal{J}}(\boldsymbol{\mathcal{Z}}|\boldsymbol{\mathcal{X}})) = K$) in the left subfigure of Figure 1. According to the illustration, we can observe the trade-off between the Fisher information and entropy power in VAE. When the Fisher information elevates, the corresponding entropy power abases and vice versa. When the dependence between data and latent code is higher, where we set larger information constraint F or M in $q_{z|x}$, the corresponding models appear in the bottom-right corner in the FS plane. In the contrary, the models that contains less information in latent code appear in the upper-left corner, for instance, pixelVAE, which was reported to ignore the latent code (Chen et al. 2017) appears nearby FAE and MAE with

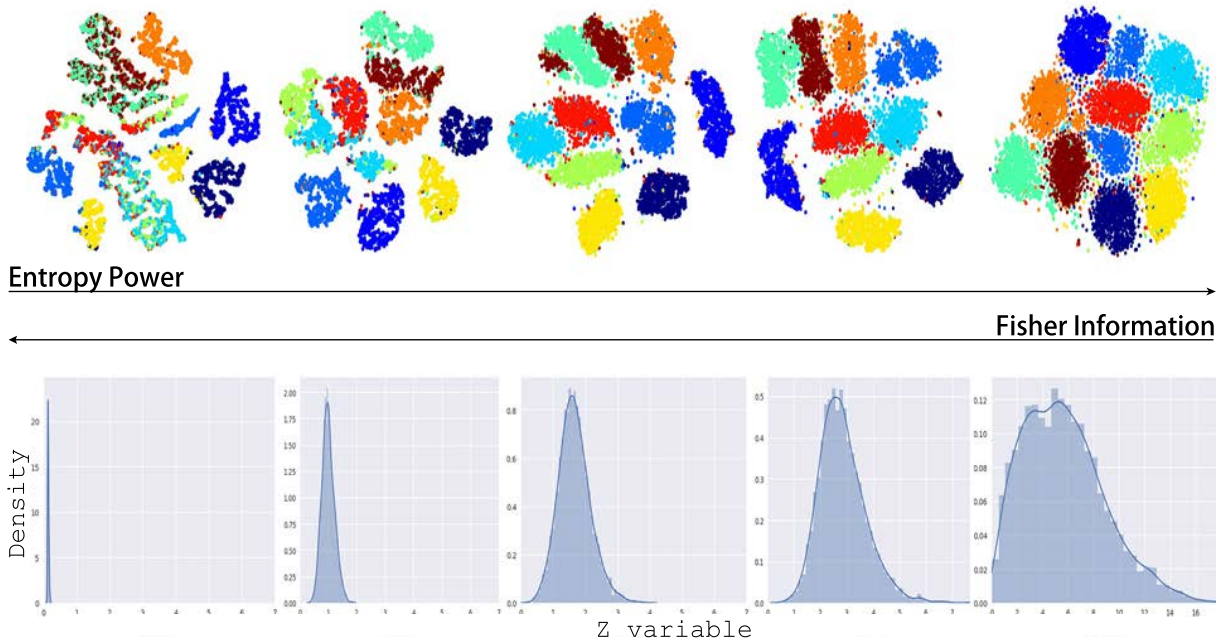


Figure 2: **From left to right:** Visualization of latent code embedding (**upper**) and distribution of $q_\phi(z)$ (**lower**) as entropy power increases (FI decreases).

$F = M = 0$. It is also interesting to notice that the inverse autoregressive flow VAE (Kingma et al. 2016) is beyond the curve. This is due to the IAF transforms the posterior into a more complex distribution from the simple Gaussian distribution. This phenomenon gives us the inspiration to improve the Fisher information and entropy power at the same time. That is to apply a more complex and a more proper distribution assumption for the modeling.

From this plane, it is not hard to learn that we can vary the Fisher information constraint F and "move" on this curve. The upper-left corner indicates a less informative code, while the bottom-right indicates a more informative code. In the right subfigure of Figure 1, the training process of a FAE is also visualized in the FS plane. We plot the location of different epochs in FS plane. It is obvious that in FS plane the training process is intrinsically moving along the "NJ curve" from upper-left side to the bottom-right side. In fact, for most models, the goal is to move further in the bottom-right corner, thus we get better knowledge about data. Setting constraint of the Fisher information means to tell the model how much information we can transfer from data to the latent code, which can affect how far we can move to the right side along this curve.

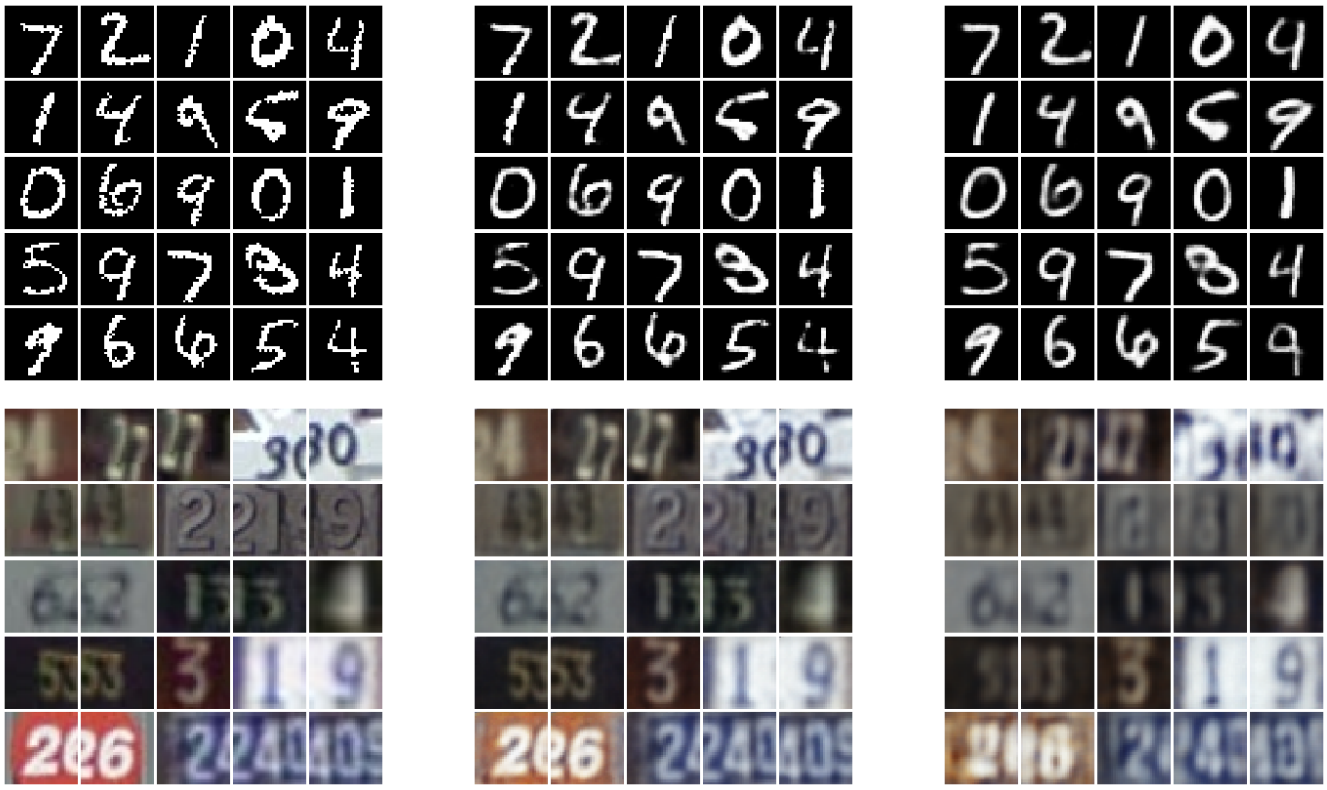
Effects of Fisher Information and Entropy Power The former part discusses the characteristics of different models and the corresponding performance. We are still curious about the effects of Fisher information and entropy power in VAEs: when should keep larger Fisher information and when should keep larger entropy power? This part shows how Fisher information and entropy power affect VAEs.

We set the latent code size to 10, and gradually increase the value of F in FAE (from 0 to 20). The embedding of latent variable z is visualized with T-SNE (Maaten and Hinton 2008); the distribution of $q_{z|x}$ is visualized by sampling z_i from $q_{z|x}$ and count the norm of normalized z_i, *i.e.* $\|z_i - \overline{z}\|$. The results are presented in Figure 2.

In Figure 2, from left to right, the entropy power increases while the Fisher information decreases. As the entropy power increases, the latent variable embedding becomes more and more expanded in the latent space, where we can observe the clusters become more and more identifiable; while as the Fisher information increases, the embedding becomes more constrained to a smaller space. When observing the distribution $q_{z|x}$, it is obvious that the distribution is more centered when Fisher information is larger, while the distribution swells with larger variance when the Fisher information is abased.

As discussed, the Fisher information will control the variance of the encoding distribution. We can easily find out in Figure 2, the variance of the distribution is getting smaller as FI increases. Intuitively, when VAEs encode the data, if we assign a large Fisher information constraint, the encoding variance is compressed to be smaller, thus the hashing cost is smaller and facilitates the model in distribution fitting. In the contrary, we can set a larger entropy power (or a smaller Fisher information) leaves more uncertainty to the encoding space, thus the latent code grabs the most common information from data points. This helps assemble data points in tasks like classification.

In brief, we conclude the characteristics of large Fisher information and entropy power:



(a) Real images (b) Reconstructed with $F = 20$ (c) Reconstructed with $F = 0$

Figure 3: **From left to right**: **Real** images from test sets of MNIST (**upper**) and SVHN (**lower**, best viewed in color); images reconstructed by FAE with **large** ($F = 20$) and **small** ($F = 0$) Fisher information.

- Large Fisher information provides with a more refined encoding mechanism, which ensures the latent code contains more detailed information.
- Large entropy power provides with a more coarse encoding mechanism, which helps in global information extraction.

Larger Fisher information is thus helpful in learning of detailed features, high quality reconstruction, *etc.*; while larger entropy power is helpful in classification, generalization, *etc.*

Qualitative Evaluation

In this section, we present some qualitative results to provide an intuitive visualization. This will help us better understand the characteristics of Fisher information.

We present some reconstruction samples of FAE with large and small Fisher information constraint F in Figure 3. As shown, the samples reconstructed with large F provide with more pixel details and are more similar to the real images. This is especially more obvious in the case of SVHN, where we can observe more clear texture compared to the one reconstructed with larger constraint F. In the contrary, we can find some blurry samples reconstructed with small F on MNIST. The blur is more obvious among reconstructed samples from SVHN.

Conclusion

Based on the uncertainty property between Fisher information and Shannon information, in this paper, we apply the Fisher-Shannon plane to study VAEs in a joint view of these two quantities. In our study of VAEs in Fisher-Shannon plane, these information quantities are demonstrated related to the representation learning and likelihood maximization; the trade-off between Fisher information and Shannon information is shown to result in different characteristics of VAEs. We further propose the Fisher Auto-Encoder for the information control by different Fisher information constraints. In our experiments, we demonstrate the complementary characteristics of Fisher information and Shannon information and provide with a novel understanding of VAEs; we also justify the effectiveness of FAE in information control, high-accuracy reconstruction and non-informative latent code resistance.

Acknowledgement

This work is supported by NSFC (61771305, 61521062), STCSM (18DZ2270700) and Australian Research Council grants (FT130100746, DP180100106 and LP150100671). We appreciate Jie Chang, Alain Chilles, Xu Chen, Kenan Cui, Maosen Li, Jialiang Lu, Olivier Rioul, Lingxi Xie and Ye Zhu for discussion.

References

Alemi, A.; Poole, B.; Fischer, I.; Dillon, J.; Saurous, R. A.; and Murphy, K. 2018. Fixing a broken elbo. In *International Conference on Machine Learning*, 159–168.

Barber, D., and Agakov, F. 2003. The im algorithm: a variational approach to information maximization. In *Proceedings of the 16th International Conference on Neural Information Processing Systems*, 201–208. MIT Press.

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35(8):1798–1828.

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Brunel, N., and Nadal, J.-P. 1998. Mutual information, fisher information, and population coding. *Neural computation*.

Burda, Y.; Grosse, R.; and Salakhutdinov, R. 2015. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Chen, X.; Kingma, D. P.; Salimans, T.; Duan, Y.; Dhariwal, P.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2017. Variational lossy autoencoder. *International Conference on Learning Representation*.

Cremer, C.; Li, X.; and Duvenaud, D. 2018. Inference suboptimality in variational autoencoders. *International Conference on Learning Representation*.

Dembo, A.; Cover, T. M.; and Thomas, J. A. 1991. Information theoretic inequalities. *IEEE Transactions on Information Theory* 37(6):1501–1518.

Dembo, A. 1990. Information Inequalities and Uncertainty Principles. *Technical Report No. 75*.

Desjardins, G.; Simonyan, K.; Pascanu, R.; and Kavukcuoglu, K. 2015. Natural neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2071–2079.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representation*.

Joyce, J. M. 2011. Kullback-leibler divergence. In *International encyclopedia of statistical science*. Springer. 720–722.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. *International Conference on Learning Representations*.

Kingma, D. P.; Salimans, T.; Jozefowicz, R.; Chen, X.; Sutskever, I.; and Welling, M. 2016. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, 4743–4751.

Larochelle, H., and Murray, I. 2011. The neural autoregressive distribution estimator. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 29–37.

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11):2278–2324.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(Nov):2579–2605.

Martin, M.; Pennini, F.; and Plastino, A. 1999. Fisher's information and the analysis of complex signals. *Physics Letters A* 256(2-3):173–180.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Phuong, M.; Welling, M.; Kushman, N.; Tomioka, R.; and Nowozin, S. 2018. The mutual autoencoder: Controlling information in latent code representations.

Rosso, O. A.; Olivares, F.; and Plastino, A. 2015. Noise versus chaos in a causal fisher-shannon plane. *Papers in physics* 7:070006.

Saxe, A. M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B. D.; and Cox, D. D. 2018. On the information bottleneck theory of deep learning. In *International Conference on Learning Representations*.

Schrödinger, E. 1930. About heisenberg uncertainty relation. *Proc. Prussian Acad. Sci. Phys. Math. XIX* 293.

Silva, L. M.; de Sá, J. M.; and Alexandre, L. A. 2005. Neural network classification using shannon's entropy. In *ESANN*, 217–222.

Stam, A. 1959. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control* 2(2):101 – 112.

Tolstikhin, I.; Bousquet, O.; Gelly, S.; and Schoelkopf, B. 2018. Wasserstein auto-encoders. In *International Conference on Learning Representations*.

Tu, M.; Berisha, V.; Woolf, M.; Seo, J.; and Cao, Y. 2016. Ranking the parameters of deep neural networks using the fisher information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2647–2651.

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems*, 4790–4798.

Van Oord, A.; Kalchbrenner, N.; and Kavukcuoglu, K. 2016. Pixel recurrent neural networks. In *International Conference on Machine Learning*, 1747–1756.

Vignat, C., and Bercher, J.-F. 2003. Analysis of signals in the fisher-shannon information plane. *Physics Letters A* 312(1):27 – 33.

Zhao, S.; Song, J.; and Ermon, S. 2017. InfoVAE: Information maximizing variational autoencoders. *arXiv:1706.02262*.