PRELIMINARY THEORETICAL EXPLORATION IN VARIATIONAL AUTOENCODER AND FITTING DATA NOISE IN IT

Anonymous authors
Paper under double-blind review

ABSTRACT

Why does variational autoencoder(VAE) suffer from bad reconstruction, and what influence the disentanglement of VAE? This paper tries to preliminarily address those issues theoretically in the idealistic situation and practically through noise modeling perspective in the realistic case. On one fold, information conservation, factor equivalence, separation of mutual information properties in the idealistic VAE case are revealed and several performance indicators regarding the disentanglement and generating influence are subsequently raised to evaluate the performance of VAE model and to supervise the used factors. On other fold, the paper proposes the adaptive noise learning algorithms of Gaussian noise and mixture Gaussian noise assumption which empirically contributes to a better reconstruction than original VAE noise assumptions. Theoretical analysis is reflexed in the experiment results.

1 INTRODUCTION

Variational AutoEncoder(VAE)s (Kingma & Welling (2013), Rezende et al. (2014)) have shown their powerful human-like abilities: modelling causal relationship, unsupervisedly extracting disentangled factors/representation (Bengio et al. (2013)) and generating signals with abundant diversities in a "latent-factor-controllable" way. Those capabilities enable the knowledge transferring through shared causes/factors among different tasks/experiences, emphasized as the important human advantages against the current machine by Lake et al. (2016) and compling with the ideal mental imagery mechanism in memory and thinking. Benefitted from those capabilities, VAEs have been widely applied to various applications, including disentangled representations learning of images and time series (Higgins et al. (2016), Kulkarni et al. (2015), Mathieu et al. (2016), (Fabius & van Amersfoort (2014)), few-shot and transfer learning (Rezende et al. (2016), Higgins et al. (2017b), Higgins et al. (2017a)), causal relationships modeling (Louizos et al. (2017)), pixel trajectory predicting (Walker et al. (2016)), joint multi-modal inference learning (Suzuki et al. (2016)), increasing diversity in imitation learning (Wang et al. (2017)), generation with memory (Li et al. (2016)) and etc.

However, the lack of public theoretical study regarding the generating and inference procedure induced by VAEs is stumbling the research process: we wouldn't exactly know whether the VAE could learn the real factors or just some fantasies that model itself makes up; we couldn't exactly clarify the need and range induced by the word "disentanglement"; we actually couldn't exactly guarantee the effectiveness of the current disentanglement metric. All those lacks of distillation would make researchers hard to effectively compute on their knowledge from the experiments and incline to make some avoidable arguments and considerations.

Further, even though some of the aforemention issues would be addressed in some idealistic situation, the real situation would be further complex and the side effect of ignorance of noise consideration in VAEs would be prominent. In VAE assumption, on the one hand, the major factors are learnt and inferred and the noise enables the optimization; on the other hand, more importantly, the prior for major factors and noise modeling together form the core inductive bias of VAE. In real implementation, different datasets may have drastically different noise and improper modeling for noise

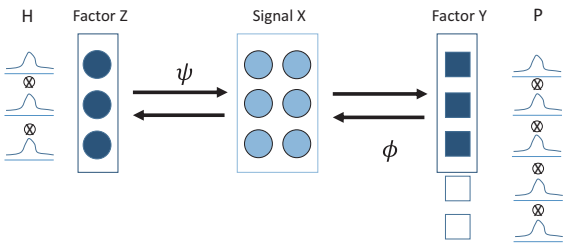


Figure 1: The illustration of the information conservation theorem. Suppose that the oracle data, denoted by random variable x, is generated by y (with P independent unit Gaussian random variables) with a homeomorphism mapping $x = \phi(y)$. Idealistic VAE will be forced to learn the factor z (with H independent unit Gaussian random variables) that generates the x with a homeomorphism mapping $x = \psi(z)$. It yields $z = \psi^{-1} \circ \phi(y)$ and $y = \phi^{-1} \circ \psi(z)$. Then according to the information conservation theorem, it must hold that $H = P$.

inclines to lead to incorrect model hypothesis and influence learning consequence. Therefore, the noise specification/modeling could significantly influence reconstruction and representation learning ability. Besides, the applications of VAE has suffered from the blurred reconstructions/generations and the unstable disentangled quality (Higgins et al. (2016)). In particular, numbers of applications assume a fixed noise that disables the model to flexible adapt to the real noise (Makhzani et al. (2015), Walker et al. (2016), Higgins et al. (2017a), Doersch (2016)). These issues, as results, sometimes could make VAE have to be an auxiliary part to only alleviate the generation shortcomings of generative adversarial network(GAN) (Larsen et al. (2015), Wang et al. (2017)).

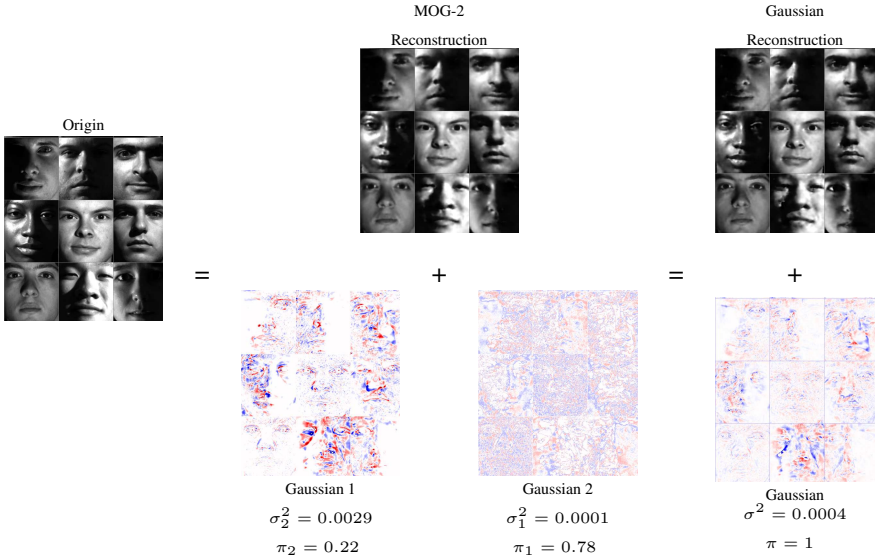


Figure 2: MoG-2/Gaussian VAE Reconstruction and Residual Gaussian Components Membership

This paper will first discuss the properties of the idealistic VAE, that is, roughly the VAE whose decoding procedure and encoding procedure are both deterministic. The information conservation theorem, the clarification on disentanglement, factor equivalence class, separation of mutual information property and several indicators will be introduced. After that we relax the discussion to the case that decoding procedure are not deterministic. We introduce Gaussian, Mixture of Gaussian noise modeling regime and discuss network parameterized Gaussian for fitting noise to data. The experiments are design to testify whether the knowledge derived form the idealistic case would be applied to the realistic sampling case and to study the behavior of different noise assumptions as well as our indictors.

2 GAUSSIAN-VAE MODEL

Gaussian-VAE (Kingma & Welling (2013), Rezende et al. (2014)) is an scalable unsupervised representation learning model (Higgins et al. (2016)), and since Gaussian distribution can be continuously and reversibly mapping to many other distributions, the theoretical analysis on it is also instructive for other continuous latent factors VAE.

Gaussian-VAE assume that input x is generated by several independent Gaussian factors z, that is $p_\theta(z) = \mathcal{N}(z|0, I_H)$. The generating/decoding process is modeled as $p_\theta(x|z)$ and the inference/encoding process $q_\phi(z|x)$ is treated as the approximate posterior distribution. Both of them are parameterized by the neural network with parameter θ and ϕ.

In VAE setting, the approximate inference[1] method is applied to maximizing the variational lower bound of $p_\theta(x) = \int p_\theta(x|z)p_\theta(z)dz$,

$$\mathcal{L}(q) = \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z)) \leq \log p_\theta(x). \tag{1}$$

3 GENERATING FACTOR PROPERTIES AND PERFORMANCE INDICTORS

In order to get assess to the question, whether the VAE could learn the real factors or just some fantasies that model itself makes up, we start the analysis from the idealistic situation[2].

3.1 DETERMINISTIC ASSUMPTION ON $q_\phi(x|z)$ AND $p_\theta(z|x)$

If the factors of x are well understood, the generation process should be deterministic, that is $p_\theta(x|z) = \delta(x = \psi(z))$. We limit the consideration that $q_\phi(z|x) = \delta(z = \psi^{-1}(x))$ is also deterministic as well for simplicity of analysis in this paper. For more complex situation, this consideration could be basic and instructive.

3.2 INFORMATION CONSERVATION

We try to address this issue by disregarding the training procedure and direct considering the idealistic VAE's behavior through the following theorem.

Theorem 1 (Information Conservation). *Suppose that $z = (z_1, \cdots, z_H)$ and $y = (y_1, \cdots, y_P)$ are sets of H and P ($H \neq P$) independent unit Gaussian random variables, respectively, then these two sets of random variables can not be the generating factor of each other. That is, there are no continuous functions $f : \mathbb{R}^H \to \mathbb{R}^P$ and $g : \mathbb{R}^P \to \mathbb{R}^H$ such that*

$$z = g(y) \quad and \quad y = f(z).$$

Proof in Appendix 8.3. The principle of the theorem is visually illustrated in Fig. 1.

3.2.1 EFFICIENT REPRESENTATION AND CLARIFICATION ON DISENTANGLEMENT

According to the information conservation theorem, the independent unit Gaussian factor assumption forms a strong inductive bias and facilitates the model incline to achieve most efficient coding. Under this assumption, the number of the learnt "used" factors of idealistic VAE should be the same as the true factors number under some assumptions such as the learnt $q_\phi(z)$ should equal $p_\theta(z)$ and

[1]The approximate posterior distribution $q_\phi(z|x)$ follows $\mathcal{N}(En(x); \Sigma_{z|x}(x))$ where En denotes the Encoder and $\Sigma_{z|x}$ represents the covariance matrix parameterized by the neural network. The reparameterization trick, $z|x = En(x) + \Sigma_{z|x}^{1/2}(x)e$ where $e \sim \mathcal{N}(0, I_H)$, enables the calculation of coefficient gradient of the first part of VAE through sampling, that is

$$\mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) = \mathop{\mathbb{E}}_{e\sim\mathcal{N}(0,1)} \log p_\theta(x|En(x) + \Sigma_{z|x}^{1/2}(x)e) \approx \frac{1}{M}\sum_{m=1}^{M} \log p_\theta(x|En(x) + \Sigma_{z|x}^{1/2}(x)e^m).$$

[2]An idealistic VAE model means that it can perfectly encode the signal into "used" factors and perfectly decode the "used" factors to original input signal and the factors follows i.i.d unit Gaussian distribution.

decode/encode process is continuous and reversible. Empirically, sometimes only a small amount of unit Gaussian variables regarding the factors of VAE have been used and the $q_\phi(z|x)$ is close to deterministic though the number latent of factors sometimes is pre-specified larger. This theorem helps provide an interpretation to explain this phenomenon.

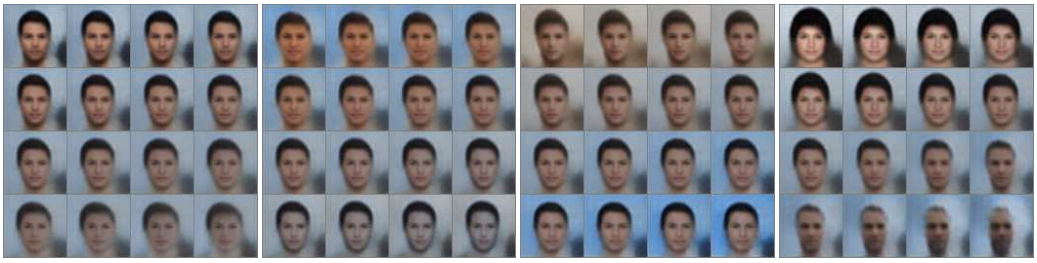
Here, in order to avoid the ambiguity of the terminology of disentanglement in this paper, based on the information conservation theorem, we make the following clarification.[3]

- The disentanglement of the learnt representation/factors in this literature refers to two parts depicted in Theorem 1:
  - the factors are closer to be independent with each other,
  - the factors incline to be able to generate the oracle signal and to be inferred perfectly from the oracle signal through a continuous procedure/mapping.
- The "disentanglement" refers to the closeness of the learnt factors to the pre-specified independent factors/concpets that can generate the oracle signal and be perfect inferred through a continuous procedure/mapping such as the independent semantic/visual factors.

Therefore, the estimation for $D_{KL}(q_\phi(z)||p_\theta(z))$ that reflects the divergence of the learnt factor distribution and the i.i.d. unit Gaussian prior can be good a indicator to supervise the independence of the factors and served to quantitatively assess the disentanglement of each extracted factor, while the similarity regarding the original signal and reconstruction place another part of the disentanglement.

The "disentanglement" will be shown hard to be obtained in an unsupervised manner. Concretely, even in the idealistic cases, the extracted factors tend to possess the intrinsic number of latent factors of the model, while there are still possibly large variations of these factors due to it can be obtained only in the equivalent class induced by the pre-specified factors as proved in the next subsection.

3.3 FACTORS EQUIVALENCE



(a) (b) (c) (d)

Figure 3: One-shot traversal & generating factor equivalence class demonstration. The images are generated by MoG-2 β(=40) VAE trained on CelebA. The seed image is obtained out from the datasets. Each block represents the traversal of the generating factor from $[-3+z_{seedh}, +3+z_{seedh}]$. (a) corresponds to face color white-yellow & female-male change. (b) corresponds to face color white to yellow change. (c) corresponds to background yellow to blue change. (d) corresponds to hair color white to black & face width change. It can be seen that changing one factor results in multiple semantic factor changes in a comprehensible manner rather than the "one-to-one" correspondence which reflexed analysis regarding generating factor equivalence.

Theorem 2 (Gaussian Factor Equivalence)**.** *Suppose that $z = (z_1, \cdots, z_H)$ is a set of H independent unit Gaussian random variables. Let $Q \in \mathbb{R}^{H\times H}$ be an orthogonal matrix and then $y = Qz$ is also a set of H independent unit Gaussian random variables. Besides, z and y can generate each other through a linear homeomorphism mapping.*

[3]Notice this clarification is based on the assumption that $q_\phi(z|x)$ is deterministic. If not, then continuous and reversible mapping of encoder constrain should be loosen and also the reversibility of the decoder should be loosen. If we further demand the enhancement of the pattern separation and completion ability, that is, to make the hyperspheres induced by the observation points in the factor domain fully fill up the whole compact factor ball, then some auxiliary constrains including the restriction on mutual information $\mathcal{I}_{encoder}(x; z)$ (defined in Section 3.4) induced by the encoder network need to be introduced.

Proof in Appendix 8.3. This theorem implies that there are a class of unit Gaussian random variables which can generate each other and have equivalent conservation information, as indicated by the following theorem.

Theorem 3 (Linear Gaussian Factor Equivalence Class)**.**

$$[z] = \{y | y = Qz, \quad Q \in \mathbb{R}^{H \times H} \textit{be the orthogonal mapping.}\}$$

Then $\forall y \in [z]$, y is a set of H independent unit Gaussian random variables and can generate z through an linear homeomorphism mapping.

The theorem clarifies that if Gaussian-VAEs have an linear matrix multiplication freedom degree of learning the factors, then the factors in the equivalence class can all be possibly learnt.

The empirically results tally with the above analysis(see Fig. 3). Suppose the visual semantic concepts can be viewed as a set of independent Gaussian variables ($z = (z_{rotation}, z_{gender}, z_{with-glass}, \cdots)^T$) which are desired to be captured and learnt by VAEs, while the model is also possible to learn the independent factor set $y = (y_1, \cdots)^T = Qz$ in the equivalence class $[z]$. This explains why changing one factor like y_1 sometimes empirically results in change in multiple visual concepts.

This perspective suggests that it's actually hard to obtain the "disentangled representation" that exactly "one-to-one" corresponds to the "independent semantic representation" even though they are in the same equivalence class. As a result, the idealistic VAE model just tend to learn the "entangle representation" if we do preset "oracle generating factors" belonging to the equivalent class.

However, though those conclusions might be upsetting, it seems not be biology impossible. A neuron in hippocampus of animals was suggested to combinatorially possess several representation capabilities. E.g., Aronov et al. (2017) found that some neurons in rat's hippocampus involved in spatial representation also were involved in representing sound frequencies after training rats by a tasks that required them to use a joystick to manipulate sound in frequency continuously.

If we assume that the most of the visual concept/factors follow the assumption regarding the "disentanglement", it is rational to qualitatively measure the interpretability of extracted latent factors by human perception to infer the disentanglement. Besides, previous empirical evidences of VAE applications (Higgins et al. (2016), Higgins et al. (2017b), Larsen et al. (2015), Mathieu et al. (2016)) suggest it an effective way.

3.3.1 Deficiency of the Existing Disentanglement Metic

Higgins et al. (2016) proposed a "simulated factor" based "disentanglement" metric on the simulation datasets. However, this metric could be hardly calculated in the real datasets and provide direct feedback of the "disentanglement". The reason is that it must pre-know the generating factors expected to be learnt.

According to Gaussian factor equivalence theorem 2 that even idealistic VAE will still learn the factors in the equivalence class, their metric could be effective sometimes for disentanglement but might suffer instability when evaluating the VAE in different trials (detailed in Appendix 8.4).

3.4 Separation of the Mutual Information

The mutual information[4] regarding the factors learned by the inference/encoder network and the signal x can be a good quantity for evaluating the generating influence. That is,

$$\mathcal{I}_{encoder}(x; z) = \mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x) || q_\phi(z)). \tag{2}$$

In order to understand and estimate which factor of the VAE was learnt and influenced the generating process, $\mathcal{I}_{encoder}(x; z_h)$ can be taken as a rational indicator[5]. If we assume that $z_1, z_2, \cdots, z_H$ is conditional independent given x[6], it can yield a useful result as the following.

[4]This term can be seen as a lower bound of the channel capacity (defined in Cover & Thomas (2012)) of the inference/encoder network.

[5]If $\mathcal{I}_{encoder}(x; z_h) = 0$, it yields x and z_h are independent with each other. The bigger $\mathcal{I}_{encoder}(x; z_h)$, the more information z_h conveys regarding x.

[6]It follows the real implementation assumption that $\Sigma_{z|x}(x) = diag(\sigma_{z1}(x), \cdots, \sigma_{zH}(x))$.

Theorem 4 (Separation of the Mutual Information)**.** *Let $z_1, \cdots, z_H$ be independent unit Gaussian distribution, and $z_1, z_2, \cdots, z_H$ be conditional independent given x. Then*

$$\mathcal{I}(z_1, \cdots, z_H; x) = \sum_{h=1}^{H} \mathcal{I}(z_h; x). \tag{3}$$

Proof in Appendix 8.3. This theorem suggests that if the learnt $q_\phi(z)$ can factorize and the $q_\phi(z|x)$ can factorize, then the consideration of each $\mathcal{I}_{encoder}(z_h; x)$ won't be excess or lose information. Besides, when those term comes in the optimization objective, then it can the negotiation between the information preservation and dimension reduction. While if the separation of mutual information doesn't hold, the negotiation could suffer from the exponential selection complexity in dimension.

3.5 INDICATORS

In order to quantify the disentanglement performance as well as the $\mathcal{I}_{encoder}(x; z)$, we assume that $q^*(z)$ is a factorized zero mean Gaussian estimation for $q_\phi(z)$. We first propose the following relevant theorem and then provide the indicators.

Theorem 5. *The terminology follows the aforemention definitions and if the involved KL-divergence and mutual information is well defined, then*

$$\mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||p_\theta(z)) = \mathcal{I}_{encoder}(x; z) + D_{KL}(q_\phi(z)||p_\theta(z)). \tag{4}$$

Proof in Appendix 8.3. The theorem demonstrates that the second term in variation lower bound in Eq. (14) is capable of controlling both the mutual information of x and z induced by the encoder network as well as the similarity of the learnt $q_\phi(z)$ and the prior $p_\theta(z)$ (one part of disentanglement performance). We can then list the indicators for assessing latent factor disentanglement:

Definition 1 (Estimation for $\mathbb{E}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||p_\theta(z))$)**.**

$$\tilde{D}_{KL}(q_\phi(z|x)||p_\theta(z)) = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z|x_m)||p_\theta(z)). \tag{5}$$

Definition 2 (Estimation for $I_{encoder}(x; z)$)**.**

$$\tilde{I}_{encoder}(x; z) = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z|x_m)||q^*(z)). \tag{6}$$

Definition 3 (Estimation for $I_{encoder}(x; z_h)$ which quantifies the influence of each factor)**.**

$$\tilde{I}_{encoder}(x; z_h) = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z_h|x_m)||q^*(z_h)). \tag{7}$$

Definition 4 (Estimation for $D_{KL}(q_\phi(z)||p_\theta(z))$)**.**

$$\tilde{D}_{KL}(q_\phi(z)||p_\theta(z)) = \tilde{D}_{KL}(q_\phi(z|x)||p_\theta(z)) - \tilde{I}_{encoder}(x; z). \tag{8}$$

Note that the above indicators 2-4 need the value of $q^*(z)$, we now introduce how to calculate this term based on Theorem 5. Through the minimization equivalence, we know that

$$\min_Q \mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||Q(z)) \Leftrightarrow \min_Q D_{KL}(q_\phi(z)||Q(z))dz, \tag{9}$$

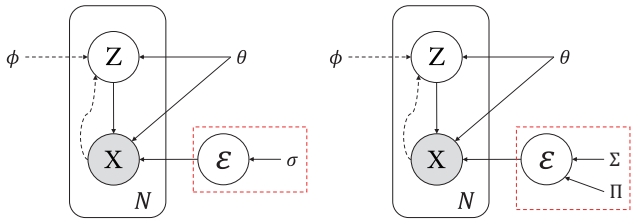
the $q^*(z)$ can then be obtained from solving the following optimization problem which can be calculated by gradient method.

$$q^*(z) = \arg\min_Q \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z|x_m)||Q(z)). \tag{10}$$

4 VAE WITH NOISE MODELING

When the consideration meets the real situation, the numbers of the factors underlaying x can be tremendous. Its not necessary and even possible to learn them all, especially for those less contribute to the generating as well as the latter potential tasks. Let we roughly call those factor as noise. More concretely, this situation correspond to that $p_\theta(x|z)$ is not deterministic since the major factors z forms a subset of the whole factors. Since VAEs learn in a purely unsupervised manner, the specification of the noise distribution $p_\theta(x|z)$ would limit the freedom of the learned major factors and define the major factor parts to some extent.[7] The superiority brought by proper consideration on noise modeling has been verified in various applications like denoising (Chen et al. (2017)), background substraction (Yong et al. (2017)), derain (Wei et al. (2017)), medical image reconstruction (Xie et al. (2017)), detecting differentially expressed genes (Lopez et al. (2017)) and ect.

In this section, we take the Gaussian variance parameter as an optimization variable and integrate it into the VAE model to make the noise fitted by data, and furthermore, we ameliorate the noise as mixture of Gaussian[8] to further enhance its noise modeling capability. The β-VAE is employed to integrate such noise modeling mechanism and also place as the comparison to the noise specification/modelling.



(a) (b)

Figure 4: The directed graph model of noise modeling VAE. (a) is under Gaussian noise assumption. (b) is under mixture of Gaussian noise assumption. Solid lines denote the generative model $p_\theta(z)p_\theta(x|z)$ (σ is treated as a variable in θ) in (a) and $p_\theta(z)p_\theta(x|z)$ (Σ and Π are treated as two variables in θ) in (b). Dashed black lines denote the variational approximation $q_\phi(z|x)$ to the intractable posterior $p_\theta(z|x)$ in (a) or $p_\theta(z|x)$ in (b).

4.1 GAUSSIAN CASE

When modelling the real-valued generation process, VAEs often assume the conditional distribution to be

$$p_\theta(x|z) = \mathcal{N}(x|G(z), \sigma^2 I_d),$$

G is the generating/decoder function parameterized by neural networks and σ^2 is the variance parameter of the Gaussian distribution. The data are formulated by the ideal generation $G(z)$ (oracle part of data) corrupted by an additional element-wise Gaussian noise ε with variance σ^2.[9] That is,

$$x = G(z) + \varepsilon. \tag{11}$$

Since the objective of the log-likelihood needs to negotiate with the KL-divergence term which controls one perspective of the disentanglement depicted in Theorem 5 in the optimization process

[7]Notice the indeterministic $p_\theta(x|z)$ could lead to the indeterministic $p_\theta(z|x)$, but since the minor factors are supposed to have less influence on x, it could not bother the deduction using the knowledge we derived in section 3 too much.

[8]The mixture of Gaussian is of a strong universal approximation capability to general probability distributions (Maz'Ya & Schmidt (1996)).

[9]This setting is quite common used in many other fields, and normally leads to L^2 loss during the optimization if there is no other term have to be negotiated with negative-log likelihood,

$$\arg\inf_G -\log p_\theta(x|z) = \frac{\|x - G(z)\|_2^2}{2\sigma^2} + \frac{1}{2}\log(2\pi)^m|\sigma^2 I_D| \longleftrightarrow \arg\inf_G \|x - G(z)\|_2^2.$$

for VAE, if the noise setting cannot guarantee to comply with the goal and prior, which naturally tends to lead to more unpredictable performance of the method in practice.

We can easily integrate the noise parameters into the VAE objective to make the noise as a learnable part in VAE to compensate a better performance on both representation and reconstruction of VAE and if necessary, more prior knowledge of noise could also be introduced to the objective.

4.2 Mixture of Gaussian Case

The noise ε in Eq. (11) in real situation might be more complex than a simple Gaussian, like that existed in real photographs (Plotz & Roth (2017)). We thus try to further ameliorate the noise setting as a mixture of Gaussian(MoG) noise. Such noise modeling strategy has been widely verified to be effective in applications, like matrix factorization (Meng & Torre (2014)) and robust principal component analysis (Zhao et al. (2014)). That is, we assume that

$$\varepsilon \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma_k^2). \tag{12}$$

In order to optimize under this noise assumption, we implement empirical EM-like SGVB algorithm detailed in appendix 8.2.

4.3 Network Parameterized Gaussian Case

Here we want to give a short discussion on the difference of the proposed VAE model with noise modeling with that raised by Kingma & Welling (2013), in which they model the noise as the following parameterized network structure:

$$\begin{aligned} x|z &\sim \mathcal{N}(G(z), \sigma^2(z)) \\ G(z) &= W_u h(z) + b_u \\ \log \sigma^2(z) &= W_\sigma h(z) + b_\sigma, \end{aligned} \tag{13}$$

where h represents the mapping induced by the previous network layers; $\sigma^2(z)$ represents the diagonal of diagonal covariance matrix; $\{W_u, W_\sigma, b_u, b_\sigma\}$ are the weights and biases of the last layer of network.

Figure 5: Reconstruction Visual Comparison

This noise assumption has very strong expressibility. The model assumes that each pixel noise indexed by d has its own level determined by $b_{\sigma d}$ and is also influenced by the deterministic part $W_\sigma h(z)$. When the noise level is shared among all data points and not influenced by the deterministic part, then it degenerates to the Gaussian assumption; if the noise has several discrete level, it then tends to degenerate to MoG assumption on noise. However, on the one hand, the assumption used in the model inclines to make the optimization for the model difficult due to the numerical instability caused even by one zero-variance residual pixel. This is possibly why most applications on VAE

have not employed this assumption while prefer to fix and manually set a noise level before VAE training. On the other hand, the over-parameterized noise could significantly increase the difficulty to find a better deterministic part since the model might be inclined to fit the noise hypothesis rather than learn a good G. These limitations have been empirically verified by all our experiments and can be observed in the experimental results as listed in Section 5.

4.4 Noise Modelling With Auxiliary Constraint

The entangled representation could be caused by the over-large of searching space of $q_\phi(z|x)$. If the learned $q_\phi(z) = \int q_\phi(z|x)p_{data}(x)dx$ has a big divergence to $p_\theta(z)$, then the VAE model in the perspective of inference network tends to learn the entangle representation as it violates the one part of the disentanglement (clarified in section 3.2.1). Actually, in the VAE model, what we want is to search in the space that $q_\phi(z)$ is possibly similar to $p_\theta(z)$ [10]. By implementing this ideal, we add auxiliary upper bound $\mathbb{E}_{x\sim p_{data}(x)} D_{KL}(q_\phi(z|x)||p_\theta(z))$ (detailed in Theorem 5) of $D_{KL}(q_\phi(z)||p_\theta(z))$ to the original objective. This equivalently leads to the approach of β-VAE raised by Higgins et al. (2016)[11],

$$\begin{aligned}
&\sup_{q_\phi(z|x)} \mathop{\mathbb{E}}_{x\sim p_{data}(x)} \mathcal{L}(q_\phi(z|x)) - (\beta-1)D_{KL}(q_\phi(z|x)||p_\theta(z)) \\
&= \mathop{\mathbb{E}}_{x\sim p_{data}(x)} \mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \log p_\theta(x|z) - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)) \qquad (14)
\end{aligned}$$

where $\beta > 1$.

5 Experiment

5.1 Questions

The experiments will try to give the answer to the following question:

- **Would noise modeling/specification influence the disentanglement?**
- **What's the difference characteristic of disentanglement induced by auxiliary constraints?**
- **How do different noise assumptions including the Gaussian, Mixture of Gaussian and Network Parameterized Gaussian act?**
- **Would the information conservation theorem be instructive in real implementation?**
- **Would the generating factor equivalence class properties exist in real implementation?**
- **How do our indictors act and whether they can determine the used the factors or not?**

5.2 Dataset

MNIST is a database of handwritten digits (Lcun et al. (1998)). CelebA (Liu et al. (2015)) is a large-scale celebfaces attributes datasets and only its images are used in our experiments. Extended Yale Face B contains (Georghiades et al. (2001), Lee et al. (2005)) images of several human subjects under different poses and different illumination conditions. More details are in Appendix 8.5.

The extensive comparison of Gaussian-noise modeling β-VAE and Gaussian-noise with specified variance (β)-VAE is made on MNIST based on our indicators to exploring the disentanglement as well as to testify the theorem and analysis from idealistic case to realistic sampling case.

[10] Jensen Shannon Divergence and other integral probability metric which can be good choice and directly be optimized through a adversarial format (Makhzani et al. (2015)) as well. However, in practice, we were defeated by the unstability of training GAN-like model.

[11] This constraint also has some richness interpretation angles and function abilities (Alemi et al. (2016)) other than the perspective we present here.

The experiments on Extended Yale Face B will focus on the comparison of behavior regarding Gaussian, Mixture of Gaussian and Network Parameterized Gaussian.

The experiments on CelebA will be auxiliary to further support the generating equivalence class theorem.

5.3 EXPERIMENTS ON MNIST

5.3.1 RELATION OF VAE AND β-VAE UNDER GAUSSIAN NOISE ASSUMPTION WHEN σ^2 IS PRE-SPECIFIED

Equivalent objective of σ^2 pre-specified Gaussian noise VAE:

$$\mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \|x - G(z)\|_2^2 - 2\sigma^2 D_{KL}(q_\phi(z|x)||p_\theta(z)).$$

Equivalent objective of σ^2 pre-specified as σ^2_{pre} Gaussian noise β-VAE:

$$\mathop{\mathbb{E}}_{z\sim q_\phi(z|x)} \|x - G(z)\|_2^2 - 2\beta\sigma^2_{pre} D_{KL}(q_\phi(z|x)||p_\theta(z)),$$

where we call $\beta\sigma^2_{pre}$ the normalized variance.

It's shown that when the σ^2 is pre-specified, manually tuning it is the same as manually tuning β with a fixed σ^2_{pre} (for example $\sigma^2_{pre} = 1$) under Gaussian noise assumption. This equivalence saves our time for extra experiment studying the behaviors of this two cases and we call those two case noise specified β-VAE.

5.3.2 NOISE MODELING/SPECIFICATION INFLUENCE THE DISENTANGLEMENT

By setting β as different values, we compare the performance of β-VAE with and without pre-specified noise on this dataset. We specifically listed the result of β(=1)-VAE in all cases. More details can be found in Appendix 8.5.2.

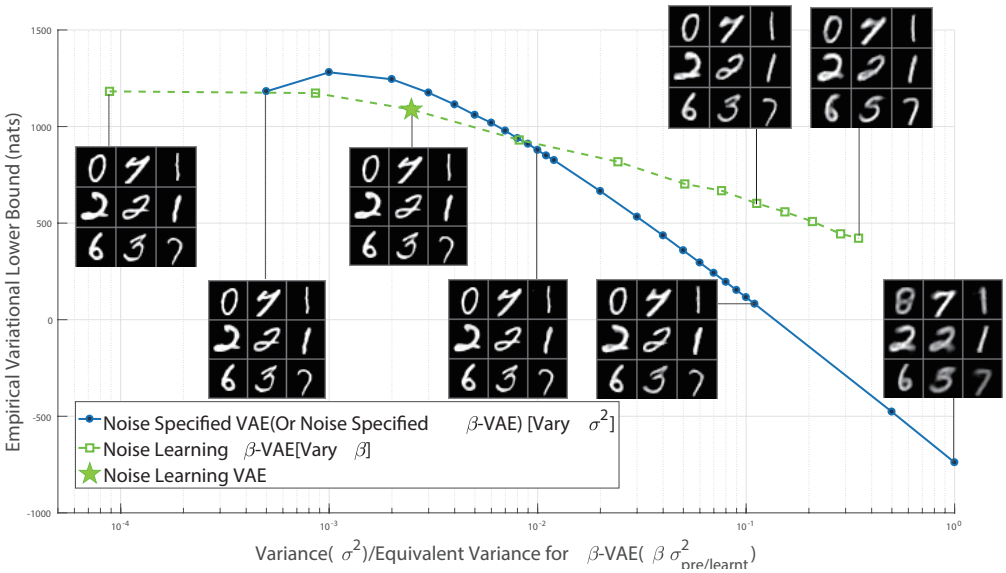
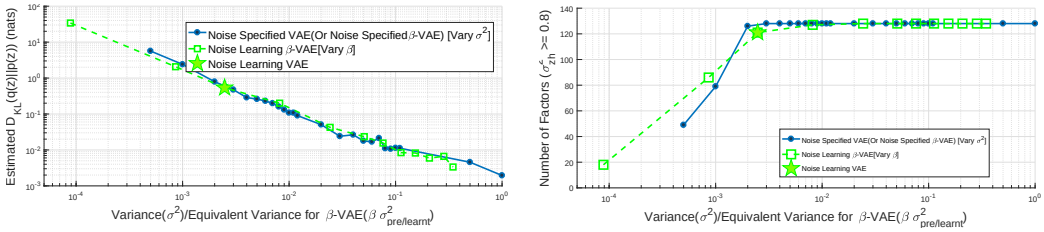


Figure 6: Noise specification/modeling influence the learned hypothesis and the reconstruction. Blue Line: the EVBL (defined in Appendix 8.1) of different specified σ^2 VAE [correspond to pre-specified σ^2_{pre} β-VAE illustrated in Section 5.3.1 with equivalent $\sigma^2 = \beta\sigma^2_{pre}$] . Green Line: the EVLB of noise learning β-VAE with different specified β [normalized to $\sigma^2 = \beta\sigma^2_{learnt}$ for convenient comparing] . Green Pentagram: the EVLB of Gaussian noise learning VAE. Other Figures: their reconstructions on the testing set. The bigger EVLB, the better hypothesis that model learnt.

The noise specifications and modeling significantly influence the model evidence quantitatively and reconstruction qualitatively, as clearly shown in Fig. (6). [12]



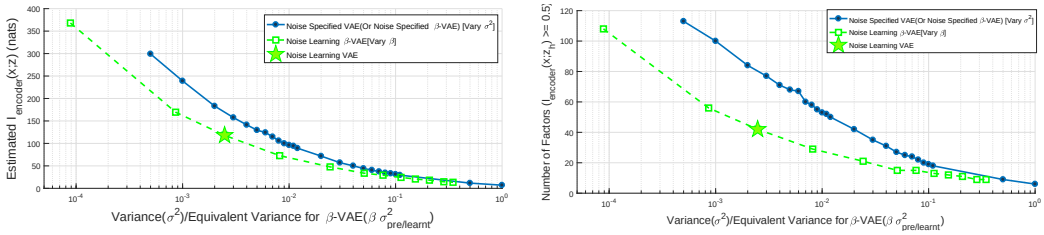
(a) (b)

Figure 7: (a) $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ of different VAE models & (b) Number of normal-variance factors of different VAE models (with 128 factors.)

The noise specifications significantly influence the divergence regarding $q_\phi(z)$ and $p_\theta(z)$ and noise specified and noise learning β-VAE achieve similar disentanglement quantitatively based on the similar indicator behaviours of $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ in Fig. (7a) and the number of normal variance factors Fig. (7b) in regard to the normalized variance.

5.3.3 Difference induced by auxiliary constraints

The prominent difference induced by auxiliary constraints would be its stronger suppression on $\tilde{I}_{encoder}(x;z)$ and the number of influential factors although changing the noise level also possess this capability indirectly. It's somewhat obvious to see and compare value of different indictors under the different β setting by sliding on the green/blue line. The bigger β, the lower $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ and roughly the better reconstruction and hypothesis learnt. However, it's more interesting that in regard to the normalized variance, noise learning β-VAE enhances the suppression on $\tilde{I}_{encoder}(x;z)$, as depicted in Fig. (8a) and that is comprehensible since β-VAE is minimizing the auxiliary constraints both $I_{encoder}(x;z) + D_{KL}(q(z)||p(z))$ based on Theorem 5.



(a) (b)

Figure 8: (a) $\tilde{I}_{encoder}(x;z)$ of different VAE models & (b) Number of influential generators of different VAE models

5.3.4 $\tilde{I}_{encoder}(x;z_h)$ effectively determines the "used" factors and information conservation hold when the disentanglement achieves in some extent.

As shown in Fig. (9) and (10), the indicator $\tilde{I}_{encoder}(x;z_h)$ determines the "used" factors. According to those figures, VAEs automatically suppress the auxiliary factors and learn the factor dimension as it still capable to have good reconstruct abilities and its $D_{KL}(q_\phi(z)||p_\theta(z))$ closer to zero under the suitable noise assumption. This phenomenon has already suggested by the information conservation theorem 1.

[12] It can be seen that in most cases the β-VAE with learnable Gaussian noise assumption is superior in learning a relatively better hypothesis with higher EVLB to that with pre-specified σ^2 ones. This can be easily interpreted by the fact that each dataset has its own level of noises, and noise modeling regime in VAE model tends to help the model better fit such noise and naturally conduct a better reconstruction result.

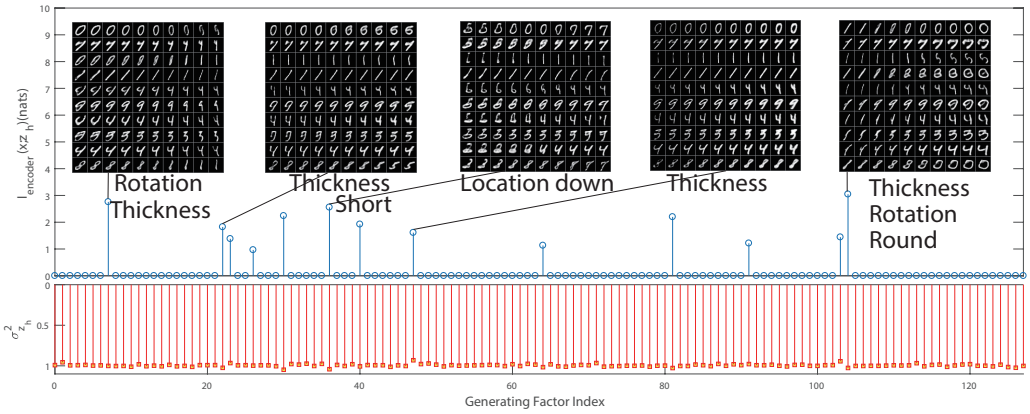


Figure 9: Noise learning β-VAE ($\beta = 10$, equivalent $\sigma^2 = 0.112$):$\tilde{I}_{encoder}(x; z_h)$, $\sigma^2_{z_h}$ and qualitatively influential factor traversals. **The top pulse subgraph**: $\tilde{I}_{encoder}(x; z_h)$ of each factor. **The bottom reverse pulse subgraph**: the estimated variance $\sigma^2_{z_h}$ of each factor. **The montages:** influential factor traversals. In all figures of factor traversal each montage corresponds to the traversal of a single factor while keeping others fixed to their inferred. Each row corresponds to a different seed image used to infer latent factor value in the VAE-based models. The traversal is over [-3, 3]. Note that all the factors with $\tilde{I}_{encoder}(x; z_h)$ not close to zero (> 0.1) can be visually told the existence of their generating effect. We select those factor traversals with visually most interpretable/comprehensive effects to present and the whole influential factor traversals are listed in appendix 17. The mutual information of "used" factor learnt by noise learning β-VAE can be found relatively balanced. The phenomenon of the multiple semantic change induced by the same learnt factor and the encoding of same semantic among different learnt factor tallies with factor equivalence class theorem 2.

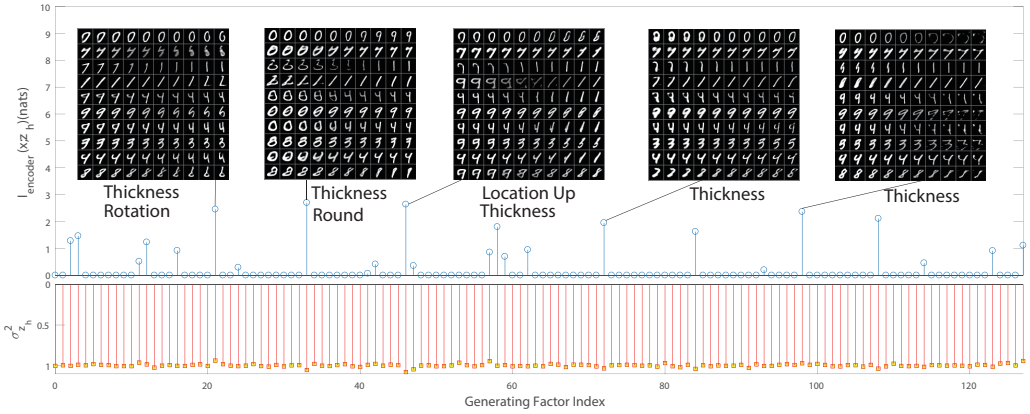


Figure 10: Noise specified (β-)VAE with equivalent $\sigma^2 = 0.11$: $\tilde{I}_{encoder}(x; z_h)$, $\sigma^2_{z_h}$ and qualitatively influential factor traversals. The mutual information of "used" factor learnt by noise specified β-VAE can be found more diverse than that in figure 9.

5.4 Experiments on Extended Yale Face Database B

In this series, we compare the β-VAEs with noise Gaussian, MoG, and network parameterized Gaussian noise components.The Table 1 quantitatively compare the EVLB performance indicator of different methods.

5.4.1 MoG is more dedicated.

β-VAE with MoG noise modeling ($\beta = 1$) learns an evidently better distribution hypothesis compared with the Gaussian one. It's comprehensive that MoG-VAE learns two different noise level

Table 1: Yale Face Database B Model Comparison [with 128 Latents (H=128)]

Noise	σ_k^2	π_k^2	EVLB
G	0.00040	1	76892
Network	-	-	90506
Network(MoG-2)	[0.00011 0.0053]	[0.728 0.272]	72778
Network(G)	0.0016	1	57195
MoG-2	[0.0001 0.0029]	[0.783 0.217]	81200

The result can be influenced by the specification of the initialization of the σ_k^2 and π_k. Variance was clipped to 0.0001 to guarantee the stable optimization. The results of Network(MoG-2) and Network(G) are derived using the fixed $G(z)$ but changing the noise hypothesis and recalculating the noise parameters.

component according to Fig. (2). One of them could be interpreted as the intrinsic physical Gaussian noise and the other might be the part which is hard to be reconstructed. However, if the noise is assumed to be one Gaussian, then it can hardly decompose such an elaborate description of noise configurations, as shown in Fig. (2).

5.4.2 Network Parameterized Gaussian inclines to overfit.

For the parameterized VAE, although the network parameterized noise VAE achieves the highest distribution hypothesis, however the qualitatively reconstructions of the network are not as good as its EVLB. The model generates more blurred reconstructions, which can be observed on two typical faces shown in Fig. (5).

We further plug the generator $G(z)$ into MoG-2 noise hypothesis, according to the Table 1, its EVLB decreases significantly. Besides, the network parameterized noise VAE suffers severely from the numerical instability such that we could not finish a complete training(2002 epoch) due to the its objective collapsed illustrated in Section 4.3. All of the aforemention evidences suggest that the model learns a relatively dedicate hypothesis for the noise rather than the deterministic part(oracle signal).

5.5 Experiments on CelebA

The reflection of the generating equivalence properties 3.3, that is, single factor could result in multiple semantic concepts change and same semantic concept could be encoded in different factors, are again well demonstrated by Fig. (11) and Table 2.

Table 2: Variants and factors correspondence on CelebA (MoG-2 β(=40)-VAE)

Variant	glass	height	blue to yellow	black to white	half bright half gloomy
Factor	73	37,45	13, 96,40,45,118	7	110

Variant	face(big to small)	lighting	face lighting	skin color(white to yellow)
Factor	63,77,82	73,90	120,110	102,96,28,63,82

Variant	head direction	neck length	hair color	gender	mouth open to close
Factor	26,31	102	120	28	8

6 Perspective

We further try to give some discussions which should be beneficial to our future works on this work.

Firstly, from the perspective of noise modeling:

- The physical noise in different practical scenarios, such as medical image processing/generating, can be taken into consideration while implementing the VAEs.

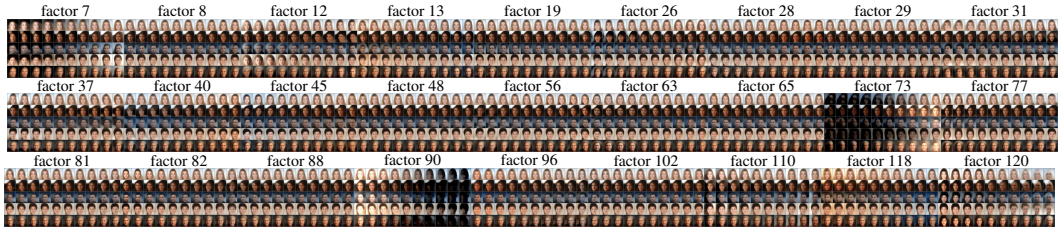


Figure 11: CelebA: Generating Factors Traversal of MoG-2 β(=40)-VAE.

- The noise modeling for other generative model and deep model is also an interesting direction.

From the perspective of representation learning:

- It is interesting that the topology properties of oracle signal are used to obtain the proof for the information conservation theorem. Other situations including that data owns several connected components can be further considered and would uncover the efficient coding properties of discrete factors.
- When $q_\phi(z|x)$ is far from deterministic, the discussion would be crucial for many general purposes induced by word disentanglement. Those study may further extend to the case that data containing different dimension manifolds and to the core VAEs' pattern separation/completion/generalization capabilities.

REFERENCES

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Dmitriy Aronov, Rhino Nevers, and David W Tank. Mapping of a non-spatial dimension by the hippocampal/entorhinal circuit. *Nature*, 543(7647):719, 2017.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *CoRR*, abs/1606.03657, 2016. URL `http://arxiv.org/abs/1606.03657`.

Yang Chen, Xiangyong Cao, Qian Zhao, Deyu Meng, and Zongben Xu. Denoising hyperspectral image with non-iid noise structure. *arXiv preprint arXiv:1702.00098*, 2017.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Carl Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Otto Fabius and Joost R van Amersfoort. Variational recurrent auto-encoders. *arXiv preprint arXiv:1412.6581*, 2014.

A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 23(6):643–660, 2001.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Irina Higgins, Arka Pal, Andrei A Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. *arXiv preprint arXiv:1707.08475*, 2017a.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. Scan: Learning abstract hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*, 2017b.

Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, Yoshua Bengio, et al. Denoising criterion for variational auto-encoding framework. In *AAAI*, pp. 2059–2065, 2017.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Diederik P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Advances in Neural Information Processing Systems*, pp. 4743–4751, 2016.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, pp. 2539–2547, 2015.

Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, pp. 1–101, 2016.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300*, 2015.

Yann Lcun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

Chongxuan Li, Jun Zhu, and Bo Zhang. Learning to generate with memory. In *International Conference on Machine Learning*, pp. 1177–1186, 2016.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Romain Lopez, Jeffrey Regier, Michael Cole, Michael Jordan, and Nir Yosef. A deep generative model for single-cell rna sequencing with application to detecting differentially expressed genes. *arXiv preprint arXiv:1710.05086*, 2017.

Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *arXiv preprint arXiv:1705.08821*, 2017.

Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian J. Goodfellow. Adversarial autoencoders. *CoRR*, abs/1511.05644, 2015. URL `http://arxiv.org/abs/1511.05644`.

Michael F Mathieu, Junbo Jake Zhao, Junbo Zhao, Aditya Ramesh, Pablo Sprechmann, and Yann LeCun. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, pp. 5040–5048, 2016.

Vladimir Maz'Ya and Gunther Schmidt. On approximate approximations using gaussian kernels. *Ima Journal of Numerical Analysis*, 16(1):13–29, 1996.

Deyu Meng and Fernando De La Torre. Robust matrix factorization with unknown noise. In *IEEE International Conference on Computer Vision*, pp. 1337–1344, 2014.

Tobias Plotz and Stefan Roth. Benchmarking denoising algorithms with real photographs. *CoRR*, abs/1707.01313, 2017. URL `http://arxiv.org/abs/1707.01313`.

Danilo Rezende, Ivo Danihelka, Karol Gregor, Daan Wierstra, et al. One-shot generalization in deep generative models. In *International Conference on Machine Learning*, pp. 1521–1529, 2016.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pp. 2234–2242, 2016.

Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pp. 835–851. Springer, 2016.

Ziyu Wang, Josh Merel, Scott Reed, Greg Wayne, Nando de Freitas, and Nicolas Heess. Robust imitation of diverse behaviors. *arXiv preprint arXiv:1707.02747*, 2017.

Wei Wei, Lixuan Yi, Qi Xie, Qian Zhao, Deyu Meng, and Zongben Xu. Should we encode rain streaks in video as deterministic or stochastic? In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Qi Xie, Dong Zeng, Qian Zhao, Deyu Meng, Zongben Xu, Zhengrong Liang, and Jianhua Ma. Robust low-dose ct sinogram preprocessing via exploiting noise-generating mechanism. *IEEE Transactions on Medical Imaging*, 36(12):2487–2498, 2017.

Hongwei Yong, Deyu Meng, Wangmeng Zuo, and Lei Zhang. Robust online matrix factorization for dynamic background subtraction. *arXiv preprint arXiv:1705.10000*, 2017.

Qian Zhao, Deyu Meng, Zongben Xu, Wangmeng Zuo, and Lei Zhang. Robust principal component analysis with complex noise. In *International Conference on Machine Learning*, pp. 55–63, 2014.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Learning hierarchical features from generative models. *arXiv preprint arXiv:1702.08396*, 2017.

7 Related Work

VAE was proposed by Kingma & Welling (2013) and Rezende et al. (2014) to implement the efficient learning and inference in directed probabilistic models regarding continuous latent variables with intractable posterior distributions and in scalable datasets. They introduced a network inference/recoginition model to represent the approximate posterior distribution and utilized reparameterization trick for stochastic joint optimization of a variational lower bound containing the parameters of both the generative/decoder and inference/recoginition/encoder models. While they also designed a network for parameterizing the noise for Gaussian MLP decoder, the capability of noise rectifying based on data has not been specifically emphasized, just similar to most of the latter VAE applications, which easily specify a fixed Gaussian noise (with preset variance parameter) but more or less underestimate the role of noise learning.

After being raised, many VAE variations have been proposed to boost VAE's capabilities in generation quality and/or disentanglement of the learned representation. In these methods, multiple efforts

were made by improving the generative and inference network structures. Typical works along this line include the convolution/de-convolution structure raised by Kulkarni et al. (2015) and ladder structure raised by Zhao et al. (2017)). Some other works advanced the mechanism under the VAE generation/inference processes. Typical works include the iterative attention generation/inference mechanism raised by Gregor et al. (2015), normalizing flow proposed by Rezende & Mohamed (2015) that enhanced the expressive ability of the approximate posterior and its variants (Kingma et al. (2016)).

Despite the improvement to the VAE itself, some other efforts were made by the ensemble between GAN with VAE. E.g., Larsen et al. (2015) unified GAN and VAE to obtain a better reconstruction and a high-level abstracts visual features embedding. Mathieu et al. (2016) also unified GAN and VAE but put emphasis on disentangling factors of variation. GANs without auxiliary design would learn the data distribution disregarding its noise level though suffer from unstable training and mode collapsing (Salimans et al. (2016)) while VAEs would assume a decomposition of the noise and oracle clean datapoint regarding the noise data with an auxiliary prior on the distribution regarding the factors. If the wrong specified noise in VAE is combined into GAN which inclines to conduce unexpected images with corruptions, it's reasonable that the ensemble model may achieve a better hypothesis of true data distribution.

Besides, many efforts were made by regularization on the factor distribution or factor generating effect. E.g., Makhzani et al. (2015) introduced an adversarial loss into the latent space of the autoencoder which in idealistic case could learn any kind factor/lantent distribution including those contributing to the disentangled factors/representation. InfoGAN, raised by Chen et al. (2016), introduced the infomax principle to GAN by adding an auxiliary mutual information regularization which enabled the inference of GANs and led to a better disentangled representation as well.

Recently, there is a new VAE variation is proposed by Higgins et al. (2016) who introduced the β-VAE framework which enhanced the constraints regarding the KL-divergence of the posterior and prior distribution of VAE and showd a novel disentanglement performance. This method has obtained a better performance as compared with conventional VAE methods, especially on its flexible tuning a compromising a parameter beta between the KL-divergence term and the likelihood term (the variational lower bound). We thus select this VAE model to embed our noise modeling regime to improve its capability of adapting real noise in data.

In the other perspective of the decomposition of the noise and oracle clean data, deep denoising models, including denoising autoencoder (Goodfellow et al. (2016)), stacked denoising autoencoders (Vincent et al. (2010)), denosing variational encoder (Im et al. (2017)) etc., have also shown similar ability to the proposed VAE model with noise modeling. However, those models are trained by using paired/supervised data (both corrupted and clean data) while our model just learn in a purely unsupervised manner without supervision regarding the noise/or oracle data. Just similar to traditional VAE, such implementation paradigm is more similar to human learning process, and can be better generalized into real applications lacking supervision knowledge.

8 Appendix

8.1 Gaussian Noise Deduction

The VAE objective (i.e., the variational lower bound) can be treated as function of the θ and ϕ where noise variance parameters σ^2 are contained in θ,

$$\mathcal{L}(\theta, \phi, x^m) = \mathop{\mathbb{E}}_{z \sim q_\phi(z|x^m)} \log p_\theta(x^m|z) - D_{KL}(q_\phi(z|x^m)||p_\theta(z)). \tag{15}$$

Here the SGVB estimator in (Kingma & Welling (2013)), $\tilde{\mathcal{L}}^B(\theta, \phi, x^m) = [\frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x^m|z^l)] - D_{KL}(q_\phi(z|x^m)||p_\theta(z))$ is used. Note that the noise variance σ^2 is also taken as an optimization variable in the model, making the model capable of better adapting noise variation of data in practical cases in a totally automatic way, instead of a manually set manner.

Given multiple data points from a dataset X, we can construct an estimator of the mean marginal likelihood lower bound of the full dataset, based on minibatches

$$\tilde{\mathcal{L}}^M(\theta, \phi, X^M) = \frac{1}{M} \sum_{m=1}^{M} \tilde{\mathcal{L}}^B(\theta, \phi, x^m), \tag{16}$$

where the minibatch $X^M = \{x^m\}_{m=1}^M$ is a randomly drawn sample set of M datapoints from the full dataset X. Such a lower bound also constitutes an important indicator for model evidence in latter experiment. We call it the empirical variational lower bound (EVLB) in the following.

Note that $\tilde{\mathcal{L}}^B(\theta, \phi, x^m) \simeq \mathcal{L}(\theta, \phi, x^m)$ and we can deduce that

$$\tilde{\mathcal{L}}^M(\theta, \phi, X^M) \simeq \mathop{\mathbb{E}}_{x \sim p_{data}(x)} \mathcal{L}(\theta, \phi, x)$$

$$\leq \mathop{\mathbb{E}}_{x \sim p_{data}(x)} \log p_\theta(x) \leq \mathop{\mathbb{E}}_{x \sim p_{data}(x)} \log p_{data}(x). \tag{17}$$

The last inequality holds due to $D_{KL}(p_{data}(x)||p_\theta(x)) \geq 0$.

The alternative optimization strategy can be readily utilized to design the algorithm for solving the model by iteratively updating the noise parameter and the network ones. During the optimization process, the objective can be monotonically increasing, and thus the algorithm can be guaranteed to be convergent.

The algorithm is summarized as follows:

Optimization for parameters except σ^2:

Optimization for σ^2: $\sigma^2 = \frac{\sum_{l=1}^{L} \sum_{m=1}^{M} \|x^m - G(z^{m,l})\|_2^2}{dML}$. (Close form solution in regard to $\tilde{\mathcal{L}}^M(\theta, \phi, X^M)$.)

Direct gradient method to the transformed variable $log_simga = \log \sigma \in \mathbb{R}$ can be implemented to lift the lower bound $\tilde{\mathcal{L}}^M$ as a result to increase the likelihood as well.

8.2 MoG Noise Deduction

The noise ε in Eq. (11) in real situation might be more complex than a simple Gaussian, like that existed in real photographs (Plotz & Roth (2017)). We thus try to further ameliorate the noise setting as a mixture of Gaussian(MoG) noise. Such noise modeling strategy has been widely verified to be effective in applications, like matrix factorization (Meng & Torre (2014)) and robust principal component analysis (Zhao et al. (2014)). That is, we assume that

$$\varepsilon \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(0, \sigma_k^2). \tag{18}$$

Let $c_d \in \{0, 1\}^K$ be the latent indicator random one-hot variable, $\sum_{k=1}^{K} c_{dk} = 1$, for the MoG-noise component of pixel indexed by d. Let $\Pi = [\pi_1, \cdots, \pi_K]$ and $\Sigma = [\sigma_1^2, \cdots, \sigma_K^2]$ be the ratio and variance of each component, respectively. Let Π, Σ be contained in θ. The conditional joint distribution turns to be

$$p_\theta(c_d, x_d|z) = \prod_{k=1}^{K} \pi_k^{c_{dk}} \mathcal{N}(x_d|G(z)_d, \sigma_k)^{c_{dk}}. \tag{19}$$

The posterior distribution $q_\phi(z, c|x)$ can be factorized as $q_\phi(z|x)q(c|x, z)$, where $q_\phi(z|x)$ will be direct learnt and the alternative of $q(c|x, z)$, $q(c|x, e)$ will be set to the last step $p_\theta(c|x, e)$ in regard to EM procedure. The lower bound of $\log p_\theta(x)$ is then reformulated as follows:

$$\mathcal{L}(q_\phi(z, c|x)) = \mathop{\mathbb{E}}_{z \sim q_\phi(\tilde{z}|x)} \mathop{\mathbb{E}}_{c \sim q(\tilde{c}|x, \tilde{z}=z)} \log p_\theta(x, c|z) + H(q(c|x, z)) - D_{KL}(q_\phi(z|x)||p_\theta(z)). \tag{20}$$

Similar to the Gaussian case, the reparamerization trick is implemented,

$$\mathcal{L}(q(c|x,e),\phi,\theta,x^m)$$
$$= \mathop{\mathbb{E}}_{e\sim\mathcal{N}(0,1)} \mathop{\mathbb{E}}_{c\sim q(\tilde{c}|x,e)} \log p_\theta(x,c|\tilde{z}) + \mathcal{H}(q(c|x,\tilde{z})) - D_{KL}(q_\phi(z|x)||p_\theta(z)), \tag{21}$$

where $\tilde{z} = En(x) + {\Sigma_{z|x}}^{1/2}(x)e$.

By utilizing the SGVB estimator, we get,

$$\tilde{\mathcal{L}}^B(q(c|x,e),\phi,\theta,x^m) = [\frac{1}{L}\sum_{l=1}^{L} \mathop{\mathbb{E}}_{c\sim q(\tilde{c}|x^m,e^{(l)})} \log p_\theta(x^m,c|z^{m,l})$$
$$+\mathcal{H}(q(c|x^m,e^{(l)}))] - D_{KL}(q_\phi(z|x^m)||p_\theta(z)). \tag{22}$$

Given an input dataset X, we can then construct an estimator to the mean marginal likelihood lower bound of the full dataset, based on minibatches, as follows:

$$\tilde{\mathcal{L}}^M(q(c|x,e),\theta,\phi,X^M) = \frac{1}{M}\sum_{i=1}^{M} \tilde{\mathcal{L}}^B(q(c|x,e),\theta,\phi,x^m), \tag{23}$$

where $z^{m,l} = En(x^m) + {\Sigma_{z|x}}^{1/2}(x^m)e^{(l)}$ and the minibatch $X^M = \{x^m\}_{i=1}^M$is a randomly drawn sample of M datapoints from the full dataset X.

Then let

$$p^{old}_{\theta_{old}}(c_d,x_d|z_{old}) = \prod_{k=1}^{K} {\pi_k^{old}}^{c_{dk}} \mathcal{N}(x_d|G^{old}(z_{old})_d,\sigma_k^{old})^{c_{dk}}, \tag{24}$$

where $z_{old} = En^{old}(x) + {\Sigma^{old}_{z|x}}^{1/2}(x)e$, and we can get

$$p^{old}_{\theta_{old}}(c_d|x,z_{old}) = \frac{p^{old}_{\theta_{old}}(c_d,x_d|z_{old})}{\sum\limits_{c_d} p^{old}_{\theta_{old}}(c_d,x_d|z_{old})}. \tag{25}$$

The EM algorithm can be naturally employed to solve the model. The implementation steps are listed as follows:

Step 1. **Expectation Step**.

Set $q(c|x^m,e^{(l)}) = p^{old}_{\theta_{old}}(c|x^m, En^{old}(x^m) + \Sigma^{old}_{z|x}(x^m)e^{(l)})$ $i = 1,\cdots,m$, $l = 1,\cdots,L$.

Calculate the expectation of the latent variable c:

$$E(c_{dmlk}) = \gamma_{dmlk} = \frac{\pi_k \mathcal{N}(x_d^m|G(z^{m,l})_d,\sigma_k^2)}{\sum_{l=1}^{L}\sum_{m=1}^{M} \pi_k \mathcal{N}(x_d^m|G(z^{m,l})_d,\sigma_k^2)}, \tag{26}$$

where $z^{m,l} : z^{m,l}_{old} = En^{old}(x^m) + \Sigma^{old}_{z|x}(x^m)e^{(l)}$.

The Objective in Maximization Step is obtained as the following,

$$\tilde{\mathcal{L}}^M(q(c|x,e),\theta,\phi,x^m) = \frac{1}{M}\sum_{i=1}^{M} -D_{KL}(q_\phi(z|x^m)||p_\theta(z))$$
$$+\frac{1}{L}\sum_{l=1}^{L}\mathcal{H}(q^{old}(c|x^m,e^{(l)})) + \sum_{k=1}^{K}\sum_{d=1}^{D}\gamma_{dmlk}[\frac{(x_d^m - G(z^{m,l})_d)^2}{2\sigma_k^2} + \frac{1}{2}\log(2\pi)\sigma_k^2 + \ln\pi_k]. \tag{27}$$

Step 2. **Maximization Step:**

Fix: $q(c|x,e)$ determined in the Expectation Step.

$$\frac{1}{M}\sum_{i=1}^{M}-D_{KL}(q_\phi(z|x^m)||p_\theta(z))+\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{d=1}^{D}\gamma_{dmlk}[\frac{(x_d^m-G(z^{m,l})_d)^2}{2\sigma_k^2}+\frac{1}{2}\log(2\pi)\sigma_k^2+\ln\pi_k]. \tag{28}$$

Update $[\Pi,\Sigma]$ and $\{\theta,\phi\}/[\Pi,\Sigma]$ by alternative optimization strategy.

Update Π,Σ: note here $z^{m,l}$: $z_{old}^{m,l} = En^{old}(x^m)+\Sigma_{z|x}^{old}(x^m)e^{(l)}$, and we can easily get the closed-form updating formula for these parameters:

$$N_k=\sum_{d,m,l}\gamma_{dmlk} \qquad \pi_k=\frac{N_k}{\sum_{k=1}^{K}N_k} \qquad \sigma_k^2=\frac{1}{N_k}\sum_{d,m,l}\gamma_{dmlk}(x_d^m-G(z_{old}^{m,l})_d)^2. \tag{29}$$

Update $\{\theta,\phi\}/[\Pi,\Sigma]$: gradient methods with respect to $\{\theta,\phi\}/[\Pi,\Sigma]$. Note here $z^{m,l}=En(x^m)+\Sigma_{z|x}(x^m)e^{(l)}$.

The algorithm can then be summarized as follows:

1. Initialize the coefficient of $\{\theta,\phi\}/[\Pi,\Sigma]$ and the coefficient of noise ε: Π,Σ.
2. Sample e from $\mathcal{N}(0,I_H)$ to obtain $e_1,\cdots,e_M$ [One for each element sample in the mini batch in the next step (L here is set to 1)].
3. Sample a mini batch X^M from $p_{data}(x)$.
4. Implement EM algorithms as aforementioned (approximate inference for $q(c,z|x)$):
   Expectation: calculate γ_{dmk}.
   Maximization: update $[\Pi,\Sigma]$, Update $\{\theta,\phi\}/[\Pi,\Sigma]$ with gradient methods.
5. Goto 3: Until Trigger End-Criterion.

8.3 PROOF

Proof. For theorem 1. Proof by Contradiction. Suppose those two function exist, and we will show that they will be inverse mapping of each other and the homeomorphism mapping of $\mathbb{R}^H$ and $\mathbb{R}^P$. Since $\mathbb{R}^H$ and $\mathbb{R}^P$ have different topology structures ($P\neq H$), the homeomorphism mapping will not exist.

$$z=g(y)=g(f(z))\ \forall z\in\mathbb{R}^H\Rightarrow g\circ f=I_H$$

$$y=f(z)=f(g(x))\ \forall y\in\mathbb{R}^P\Rightarrow f\circ g=I_P$$

Since both f and g are continuous, there is a homeomorphism mapping between $\mathbb{R}^H$ and $\mathbb{R}^P$ and it leads to the contradiction. □

Proof. For theorem 2. We only need to test the mean and variance of y.

$$\mathbb{E}(y)=\mathbb{E}(Qz)=Q\,\mathbb{E}(z)=0$$

$$Cov(y,y)=QCov(z,z)Q^T=QIQ^T=I$$

Therefore, y is another set of H independent unit Gaussian random variables. Since $x=Q^Ty$, z and y can generate each other with an linear homeomorphism mapping. □

Proof. For theorem 4,

$$\mathcal{I}(z_1,\cdots,z_H;x)=\int p(z_1,\cdots,z_H,x)\log\frac{p(z_1,\cdots,z_H,x)}{p(z_1,\cdots,z_H)p(x)}dz_1\cdots dz_H dx$$

$$= \int p(z_1, \cdots, z_H, x) \log \frac{\prod_{h=1}^{H} p(z_h|x)}{\prod_{h=1}^{H} p(z_h)} dz_1 \cdots dz_H dx = \sum_{h=1}^{H} \int p(z_h, x) \log \frac{p(z_h|x)}{p(z_h)} dz_h dx$$

$$= \sum_{h=1}^{H} \mathcal{I}(z_h; x).$$

□

Proof. For theorem 5.

$$\mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||p_\theta(z)) = \int q_\phi(z|x) p_{data}(x) \frac{q_\phi(z|x) p_{data}(x)}{p_\theta(z) p_{data}(x)} dx$$

$$= \int q_\phi(z|x) p_{data}(x) \frac{q_\phi(z|x) p_{data}(x)}{q_\phi(z) p_{data}(x)} \frac{q_\phi(z)}{p_\theta(z)} dx = \mathcal{I}_{encoder}(x; z) + D_{KL}(q_\phi(z)||p_\theta(z)). \quad (30)$$

□

8.3.1 Auxiliary Explanations for Indicators

Corollary 1. *The terminology follows the aforemention definitions and if the involved KL-divergence and mutual information be well defined then*

$$\mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||q^*(z)) = \mathcal{I}_{encoder}(x; z) + D_{KL}(q_\phi(z)||q^*(z)). \quad (31)$$

The proof of corollary 1 is the same as that of theorem 5. This corollary suggests that the estimation in definition 2 provides another upper bound for the capacity of the encoder network. Empirically, this estimation is a much tighter estimation than using the estimation in definition 1.

Corollary 2. *The terminology follows the aforemention definitions and if the involved KL-divergence and mutual information be well defined then*

$$\mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||p_\theta(z)) - \mathop{\mathbb{E}}_{x \sim p_{data}(x)} D_{KL}(q_\phi(z|x)||q^*(z))$$

$$= D_{KL}(q_\phi(z)||p_\theta(z)) - D_{KL}(q_\phi(z)||q^*(z)) \leq D_{KL}(q_\phi(z)||p_\theta(z)). \quad (32)$$

The corollary is an direct result of theorem 5 and corollary 1. It suggests that the estimation in definition 4 is a lower bound for $D_{KL}(q_\phi(z)||p_\theta(z))$.

Definition 5 (Another Estimation for $D_{KL}(q_\phi(z)||p_\theta(z))$)**.**

$$\bar{D}_{KL}(q_\phi(z)||p_\theta(z)) = D_{KL}(q^*(z)||p_\theta(z)). \quad (33)$$

Empirically, $\bar{D}_{KL}(q_\phi(z)||p_\theta(z))$ and $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ shown the same estimation results on MNIST.

Definition 6 (Another estimation for $I_{encoder}(x; z)$)**.**

$$\bar{I}_{encoder}(x; z) = -D_{KL}(q^*(z)||p_\theta(z)) + \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z|x_m)||p_\theta(z)). \quad (34)$$

Definition 7 (Another estimation for $I_{encoder}(x; z_h)$ which quantifies the influence of each factor)**.**

$$\bar{I}_{encoder}(x; z_h) = -D_{KL}(q^*(z_h)||p_\theta(z_h)) + \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_\phi(z_h|x_m)||p_\theta(z_h)). \quad (35)$$

8.4 ANALYSIS ON THE "DISENTANGLEMENT" METRIC RAISED IN β-VAE (HIGGINS ET AL. (2016))

The terminology inherits those in the β-VAE paper. The main idea of that "disentanglement" metric is to create a statistic point z_{diff} relevant to the model for each simulated factor respectively and then to use a linear classifier to project the statistic point to the corresponding index of the simulated factor. If the statistic points induced by the model are easy to be separated then the model is thought to learn "disentangled" representation.

Here, we will argue that even for the idealistic VAE model that follows the disentanglement conditons 3.2.1 could still receive bad score under that performance metric in some situations.

Suppose that the true simulated factors v follows $\mathcal{N}(0, I_H)$. Then the learnt "used"[13] factor z can be in the equivalence class $[v]$ according to theorem 2. Concretely, there exists an orthogonal transformation Q such that $z = Qv$.

Suppose that the simulated factors with index y of v is fixed. Suppose $v^1_{y-fixed}$ and $v^2_{y-fixed}$ are two random variable representing the samples from the y-fixed v. Then the factors inferred by the idealistic VAE turns to be $z^1 = Qv^1_{y-fixed}$ and $z^2 = Qv^2_{y-fixed}$.

In order to calculate the statistic point $z_{diff}(y) = \mathbb{E}\,|z^1 - z^2|$, we first calculate the mean and variance of $(z^1 - z^2)$.

$$\mathbb{E}(z^1 - z^2) = Q\,\mathbb{E}(v^1_{y-fixed} - v^2_{y-fixed}) = 0 \tag{36}$$

$$\begin{aligned} Var(z^1 - z^2) &= Var(Q(v^1_{y-fixed} - v^2_{y-fixed})) = QCov(v^1_{y-fixed} - v^2_{y-fixed}, v^1_{y-fixed} - v^2_{y-fixed})Q^T \\ &= Qdiag(2, \cdots, 2, \underline{0}_y, 2, \cdots, 2)Q^T = 2I - Qdiag(0, \cdots, 0, \underline{2}_y, 0, \cdots, 0)Q^T = 2I - 2q_y q_y^T. \end{aligned} \tag{37}$$

Therefore, $z_{diff}(y)$ can be obtained through the diagonal value of $2I - 2q_y q_y^T$. That is,

$$z_{diff}(y) = \mathbb{E}(|z^1 - z^2|) = 2\sqrt{\frac{2}{\pi}}(\sqrt{(1 - q^2_{y1})}, \cdots, \sqrt{(1 - q^2_{yH})})^T. \tag{38}$$

From the above equation, the location of statistic point is unique determined by $(q^2_{y1}, \cdots, q^2_{yH})$. When $(q^2_{y1}, \cdots, q^2_{yH})$ is close to the vertex of the unit cubic for each y then all the statistic points turn to be easily separated.

However, from the perspective of the objective, all the orthogonal Qs are with the same potential to be learnt. It seems not to be with a small probability that statistic points of different indexes take similar location. For instance, if $H = 2$ and $Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$ then $z_{diff}(1) = z_{diff}(2)$ cannot be separated while the representation still follows the disentanglement conditions.

Among different trials, the Q might contribute to that "disentanglement" metric but also might not. That explains why that metric is unstable.

8.5 EXPERIMENT DETAILS

We set L to 1, and minibatch size M to be 100 in all practices. All the pixels value have been linear normalized in to [0,1].

8.5.1 CLARIFICATION ON THE CORRECTION ON RESULTS

We find in the last version that the code on calculating $q^*(z)$ is wrong. Concretely, the objective of the KL-divergence of two Gaussian is incorrect calculated. That influences the estimation of

[13] The auxiliary unused factor is innocuous for the subsequent analysis.

$\tilde{I}_{encoder}(x;z)$, $\tilde{I}_{encoder}(x;z_h)$, $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ and $\sigma^2_{z_H}$ but has little influence on the determination of the "used" factors. So we redo all the experiments on MNIST and delete the relevant results regarding those wrong-calculated indictors on CelebA and Extended Yale Face B.

The new experiment results solve many our past confusions due to the wrong experiment. They are

- Why $\sigma^2_{z_h}$ is strongly correlated with $\tilde{I}(x;z_h)$?
- Why $\sigma^2_{z_h}$ of the used factor is always relative small?
- Why can VAE still learn the "disentangled" representation when the learnt $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ is such big?

Now we know that they are actually not the cases. Something better is that we find our experiment results are more close to our theoretical analysis: when guaranteeing the reconstruction quality in a tolerance range, the smaller $\tilde{D}_{KL}(q_\phi(z)||p_\theta(z))$ the less "used" factors are learnt. That coincide with information conservation theorem: the independent unit Gaussian of the factors assumption facilitates the most efficient coding.

8.5.2 MNIST

We split randomly 7000 datapoints according to ratio $[0.6 : 0.2 : 0.2]$ into training set, validation set (no use), testing set. All the indicators and $q^*(z)$ are evaluated/calculated on 10000 datapoints belonging to the testing set. All the seed images used to infer latent code and to draw the traversal come from the testing set.

In all figures of latent code traversal each block corresponds to the traversal of a single latent variable while keeping others fixed to either their inferred (β-VAE, VAE). Each row represents a different seed image used to infer the latent values in the VAE-based models. β-VAE and VAE traversal is over the $[-3, 3]$ range.

The assumed variance σ^2 of noise specified Gaussian of VAE models is enumerated from $[0.0005, 0.001 : 0.001 : 0.012, 0.02 : 0.01 : 0.11]$. The β setting for the noise learning β-VAE is enumerated from $[0.1, 0.5, 1, 2 : 2 : 18]$.

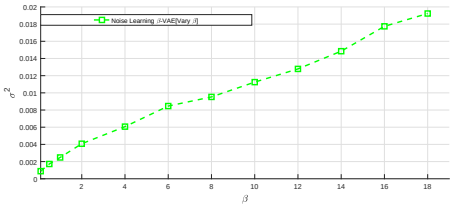


Figure 12: Learned σ^2 of different β setting

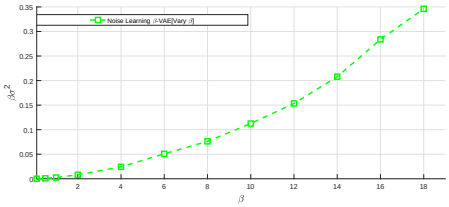


Figure 13: $\beta\sigma^2$ of different β setting

8.5.3 EXTENDED YALE FACE B

We split randomly 2424 datapoints according to ratio $[0.8 : 0.1 : 0.1]$ into training set, validation set (no use), testing set. The model is training on the training set. All the seed images used to infer latent code and to draw the traversal come from the 100 datapoints from the testing set.

In all figures of latent code traversal each block corresponds to the traversal of a single latent variable while keeping others fixed to either their inferred (β-VAE, VAE). Each row represents a different seed image used to infer the latent values in the VAE-based models.

β-VAE and VAE traversal is over the $[-3, 3]$ range.

The β setting for the noise learning β-VAE is enumerated from $[1, 40, 80, 120, 160]$.

8.5.4 CELEBA

We split randomly roughly 200000 datapoints according to ratio $[0.8 : 0.1 : 0.1]$ into training set, validation set (no use), testing set. The model is training on the training set. All the indicators and

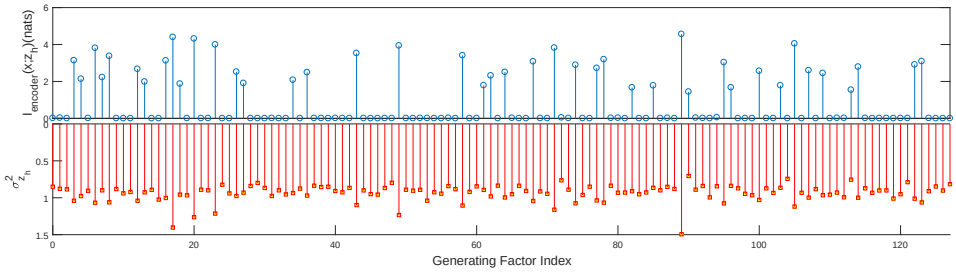


Figure 14: Noise learning β-VAE ($\beta = 1$, equivalent $\sigma^2 = 0.00248$): estimation of $I_{encoder}(x; z_h)$, σ_h^2.

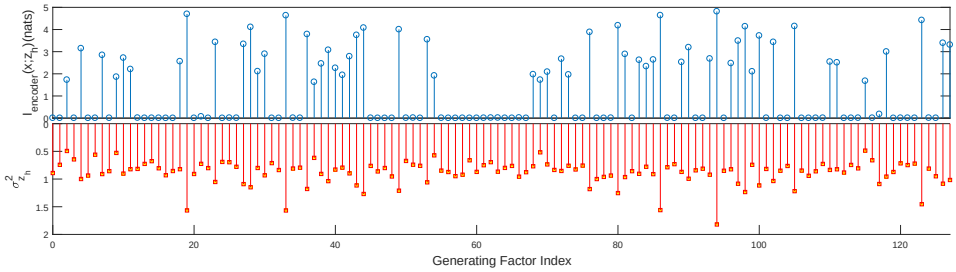


Figure 15: Noise learning β-VAE ($\beta = 0.5$, equivalent $\sigma^2 = 0.00086$): estimation of $I_{encoder}(x; z_h)$, σ_h^2.

$q^*(z)$ are evaluated/calculated on 10000 datapoints selected from testing set. All the seed images used to infer latent code and to draw the traversal come from the 100 datapoints from the testing set.

In all figures of latent code traversal each block corresponds to the traversal of a single latent variable while keeping others fixed to either their inferred (β-VAE, VAE). Each row represents a different seed image used to infer the latent values in the VAE-based models.

β-VAE and VAE traversal is over the $[-3, 3]$ range.

The β setting for the noise learning β-VAE is enumerated from $[1, 30, 40]$.

8.5.5 NETWORK STRUCTURE

Dataset	Optimiser	Architecture	
Mnist	Adam $1e-3$ Epoch 200	Input Encoder Latents Decoder	28x28x1 Conv 32x4x4,32x4x4 (stride 2). FC 256. ReLU activation. 128 FC 256. Linear. Deconv reverse of encoder. ReLU activation.
CelebA	Adam $1e-4$ Epoch 20	Input Encoder Latents Decoder	64x64x3 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (stride 2). FC 256. ReLU activation. 128/32 FC 256. Linear. Deconv reverse of encoder. ReLU activation.
Extended Yale Face B	Adam $1e-4$ Epoch 2002	Input Encoder Latents Decoder	192x168x1 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (stride 2). FC 256. ReLU activation. 128 FC 256. Linear. Deconv reverse of encoder. ReLU activation.
Extended Yale Face B (Network Parameterized Noise)	Adam $1e-4$ Epoch 1460	Input Encoder Latents Decoder	192x168x1 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (stride 2). FC 256. ReLU activation. 128 FC 256. Linear. Deconv reverse of encoder. ReLU activation.

8.6 AUXILIARY GENERATING PICTURE

Note that only the factors with $\tilde{I}_{encoder}(x; z_h) > 0.5$ are shown.

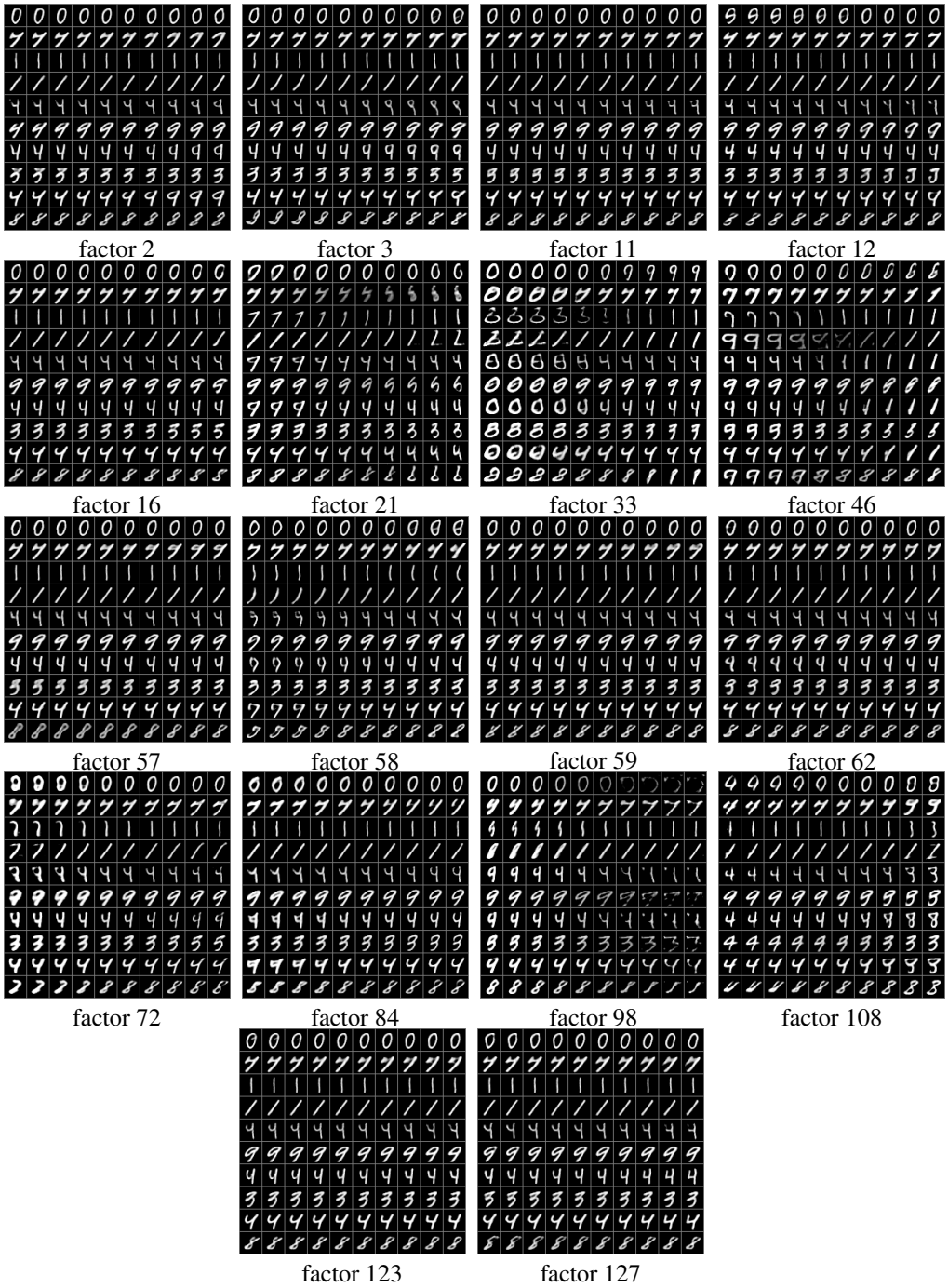


Figure 16: MNIST: Generating Factor Traversal of $\sigma^2 = 0.11$ Pre-specified VAE

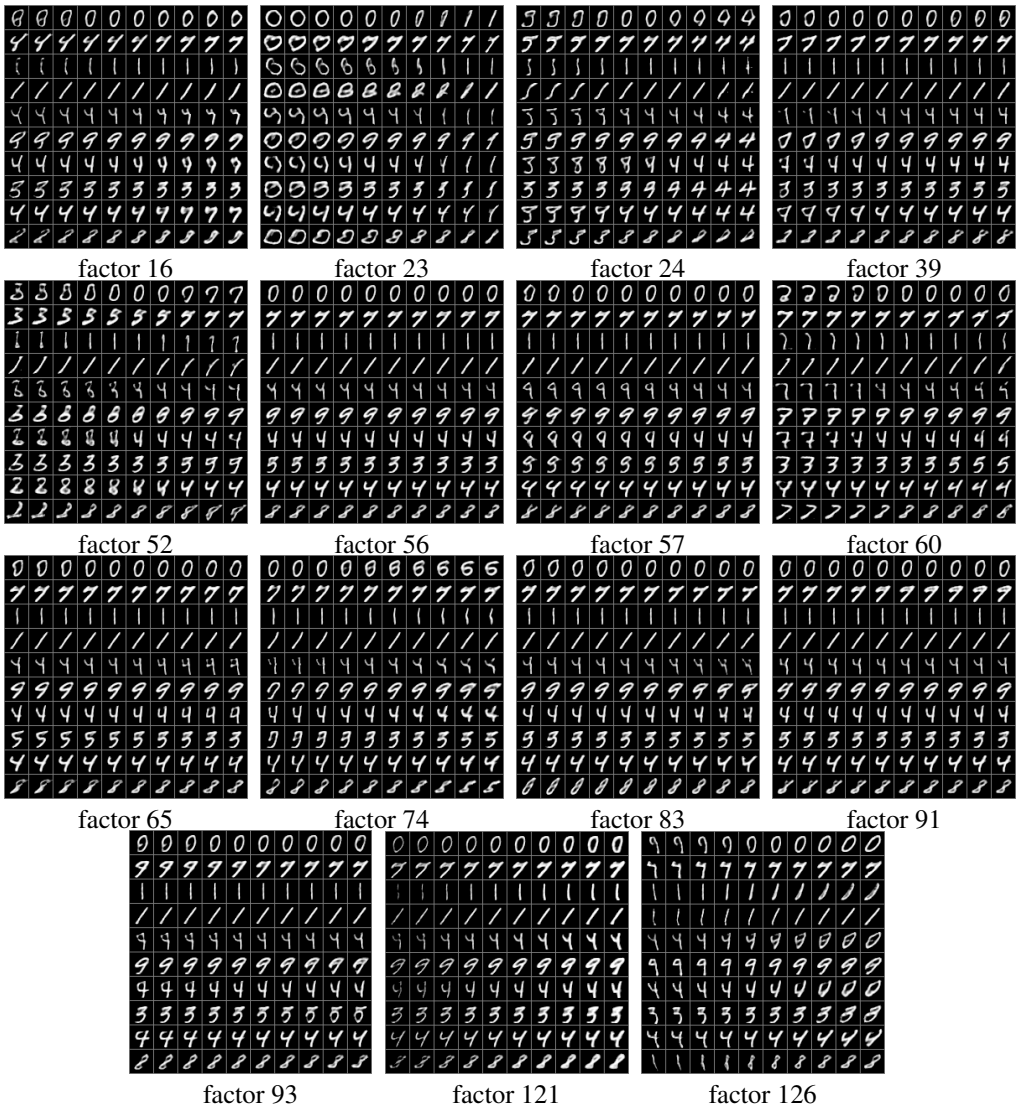


Figure 17: MNIST: Generating Factor Traversal of Noise Learning β(=10)-VAE

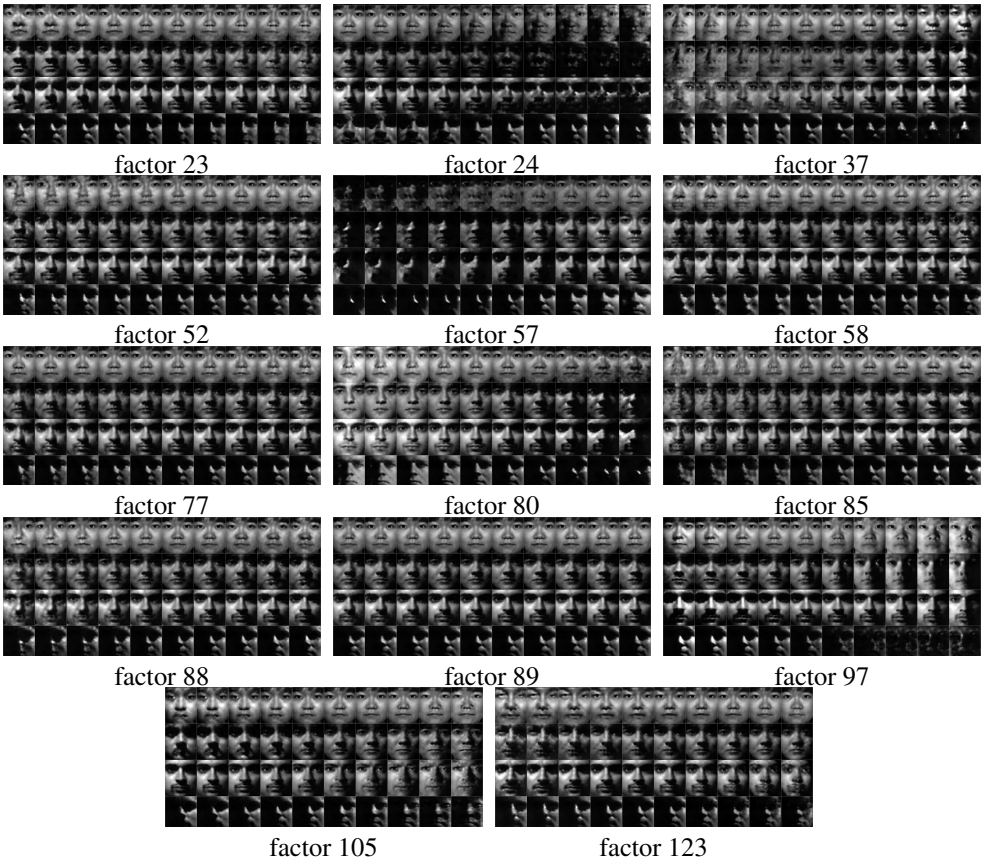


Figure 18: Extended Yale Face B: Generating Factor Traversal of Noise Learning $\beta(= 120)$-VAE

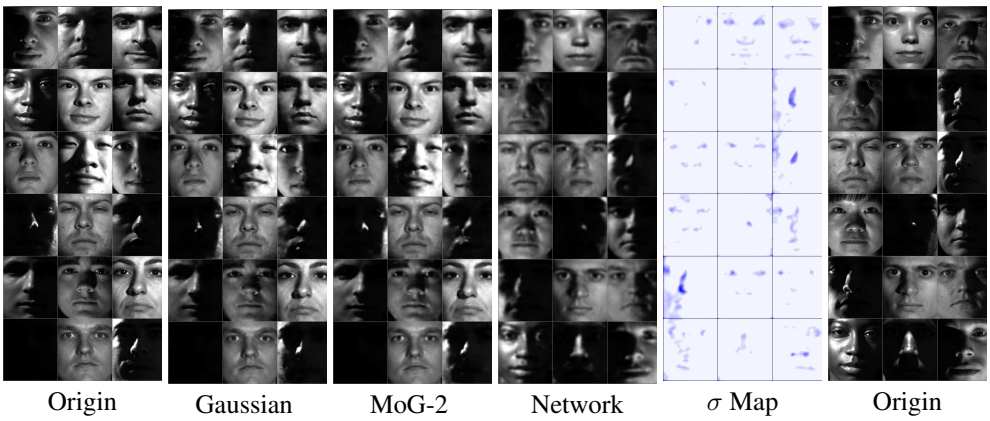


Figure 19: Reconstruction Visual Comparison 3 Column

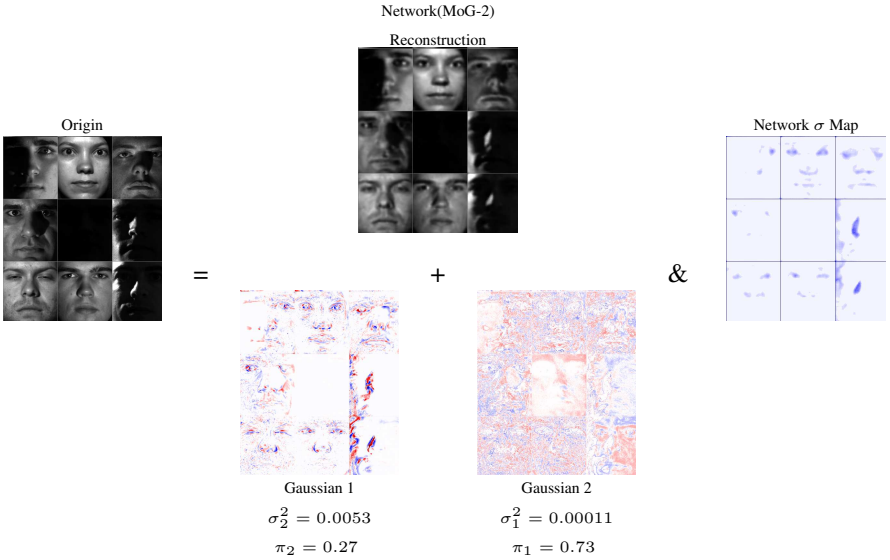


Figure 20: Network (MoG-2) Reconstruction and Residual Gaussian Components Membership

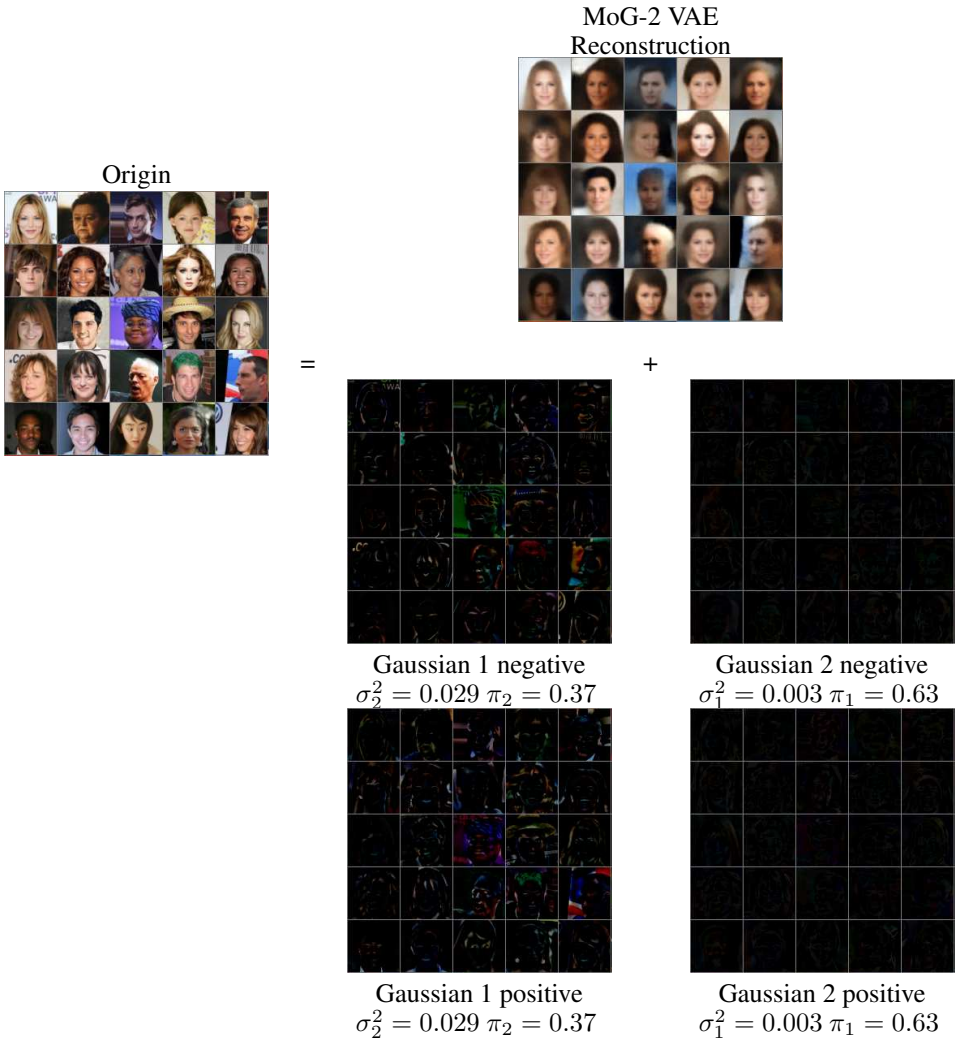


Figure 21: MoG-2 β(=40)-VAE reconstruction and residual Gaussian components membership