

# Discovering Influential Factors in Variational Autoencoder

Shiqi Liu<sup>1</sup>, Jingxin Liu<sup>2</sup>, Qian Zhao<sup>1</sup>, Xiangyong Cao<sup>1</sup>, Huibin Li<sup>1</sup>, Hongying Meng<sup>2</sup>, Sheng Liu<sup>3</sup>, Deyu Meng<sup>1</sup>

<sup>1</sup>Xi'an Jiaotong University, Xi'an, Shaan'xi Province, 710049, P. R. China

<sup>2</sup> Brunel University London, London, United Kingdom

<sup>3</sup> State University of New York at Buffalo, NY 14214, United States

Shiqi Liu and Jingxin Liu made equal contributions to this work.

## Abstract

In the field of machine learning, it is still a critical issue to identify and supervise the learned representation without manually intervention or intuition assistance to extract useful knowledge or serve for the latter tasks in machine learning. In this work, we focus on supervising the influential factors extracted by the variational autoencoder(VAE). The VAE is proposed to learn independent low dimension representation while facing the problem that sometimes pre-set factors are ignored. We argue that the mutual information of the input and each learned factor of the representation plays a necessary indicator. We find the VAE objective inclines to induce mutual information sparsity in factor dimension over the data intrinsic dimension and results in some non-influential factors whose function on data reconstruction could be ignored. We show mutual information also influences the lower bound of VAE's reconstruction error and latter classification task. To make such indicator applicable, we design an algorithm on calculating the mutual information for VAE and prove its consistency. Experimental results on Mnist, CelebA and Deap datasets show that mutual information can help determine influential factors, of which some are interpretable and can be used to further generation and classification tasks, and help discover the variant that connects with emotion on Deap dataset.

## Introduction

Learning efficient low dimension representation of data is important in machine learning and related applications. Efficient and intrinsic low dimension representation is helpful to exploit the underlying knowledge of data and serves for latter tasks including generation, classification and association. Early linear dimension reduction (Principle Component Analysis) has been widely used in primary data analysis and its variant has been applied in face identification (Yang et al. (2004)) and classical linear independent representation (Independent Component Analysis (Hyvärinen, Karhunen, and Oja (2004)) have been used in blind source separation (Jung et al. (2000)) and EEG signal processing (Makeig et al. (1996)). Nonlinear dimension reduction (e.g. Autoencoder (Goodfellow, Bengio, and Courville (2016))) begins to further learn abstract representation and has been used in semantic hashing (Salakhutdinov and Hinton (2009) and many other tasks. Recently, a new technique, called variational autoencoder (Kingma and Welling (2013); Rezende, Mo-

hamed, and Wierstra (2014)) has attracted much attention of researchers, due to its capability in extracting nonlinear independent representation. The method can further model causal relationship, represent disentangled visual variants (Mathieu et al. (2016)) and interpretable time series variants (Hsu, Zhang, and Glass (2017)) and this method can serve for generating signals with abundant diversities in a "factor-controllable" way (Suzuki, Nakayama, and Matsuo (2016); Higgins et al. (2017b)). The related techniques enable the knowledge transferring through shared factors among different tasks (Higgins et al. (2017a)).

However, the usage of VAE on extracting factors are unclear and we lack efficient methodologies to quantify the influence of each learned factor on data representation. In application, sometimes some pre-set factors remain unused<sup>1</sup> (Goodfellow, Bengio, and Courville (2016)), and the relation between the learned factors and original data has to be discovered by manually intervention (visual or aural observation). This leads to the waste on extra factors and hinders the factor selection for the subsequent tasks such as generating meaningful image/audio. Besides, some classical influence determination methods including estimating the variance of each factor lose its utility on the VAE. Therefore, identifying and monitoring the influential factor of VAE becomes a critical issue along this line of research.

In order to efficiently determine and supervise the learned factors, this paper has made the following efforts.

- We first adopt mutual information as the quantitative indicator of assessing the influence of each factor on data representation in the VAE model. Besides, in order to analyze the rationality of this indicator, we theoretically prove that how mutual information influence the lower bound of VAE's reconstruction error and subsequent classification task.
- We propose an estimation algorithms to calculate the mutual information for all the factors of VAE, and then we prove its consistency.
- We substantiate the effectiveness of the proposed indicator by experiments on Mnist (Lcun et al. (1998)), CelebA (Liu et al. (2015)) and Deap (Koelstra et al. (2012)). Especially, some discovered factors by the proposed indicator

<sup>1</sup>Montage (D) in Fig.(1) is a typical traversal of the unused factor.

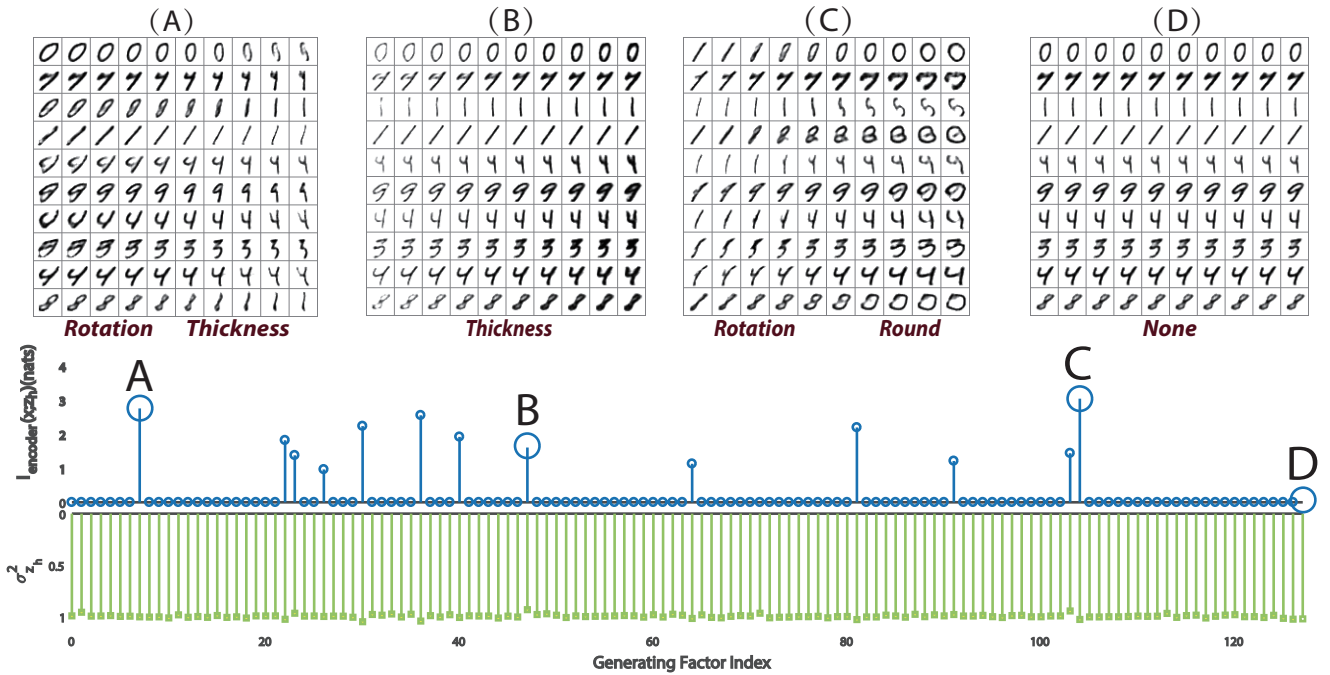


Figure 1: **Estimated  $I(\mathbf{x}; z_{\text{enc}_h})$  determines the influential factors;**  $I(\mathbf{x}; z_{\text{enc}_h})$ ,  $\sigma_{z_h}^2$  and qualitatively influential factor traversals of  $\beta(= 10)$ -VAE on Mnist. The top pulse subgraph:  $I(\mathbf{x}; z_{\text{enc}_h})$ ,  $\sigma_{z_h}^2$  and qualitatively influential factor traversals of  $\beta(= 10)$ -VAE on Mnist. The bottom reverse pulse subgraph: the estimated variance  $\sigma_{z_h}^2$  of each factor. The A,B,C montages: influential factor traversals corresponding to factor A,B,C noted in the pulse graph and the whole influential factor traversals are listed in Fig.(7) in Appendix . The montage D is the traversal of ignored factors with little estimated mutual information. According to the four montages, the variance can't determine the influential factors as mutual information indicator does.

are found meaningful and interpretable for data representation and other left ones are generally ignorable for the task. The capability of the selected factors on generalization and classification tasks are also verified.

This paper is organized as the following. We introduce the VAE model for generation and classification in Section 2. We argue the necessity of mutual information as a indicator in Section 3. Specifically, we introduce the mutual information of input data and factors, analyze the cause through the perspective of mutual information and data intrinsic dimension, discuss the relationship of mutual information and recover as well as the classification and propose the estimator and prove its consistency. We review the related work on supervising the factors of VAE in Section 4. The experiments are in Section 5.

## VAE model

VAE (Kingma and Welling (2013), Rezende, Mohamed, and Wierstra (2014)) is a scalable unsupervised representation learning model (Higgins et al. (2016)): VAE assumes that input  $\mathbf{x}$  is generated by several independent Gaussian random variables  $\mathbf{z}$ , that is  $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, I_H)$ . Since Gaussian distribution can be continuously and reversibly mapping to many other distributions, the theoretical analysis on it might be also instructive for other continuous-latent VAEs. The generating/decoding process is modeled as

$p_{\text{dec}}(\mathbf{x}|\mathbf{z})$  and the inference/encoding process  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_H(\mathbf{x})))$  is treated as the approximate posterior distribution. Note that it yields that  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = q_{\text{enc}}(z_1|\mathbf{x}) \cdots q_{\text{enc}}(z_H|\mathbf{x})$ . We assume both of them are parameterized by the neural network with parameter  $\text{enc}$  and  $\text{dec}$ .

**Factor:** Let  $\mathbf{z}_{\text{enc}}$  denote random variables determined by  $q_{\text{enc}}(\mathbf{z}) = \int q_{\text{enc}}(\mathbf{z}|\mathbf{x})p_{\text{data}}(\mathbf{x})d\mathbf{x}$  and a factor in the latter literature refers to a dimension of  $\mathbf{z}_{\text{enc}}$ .

## Generation

In VAE setting, the approximate inference method is applied to maximizing the variational lower bound of  $\log p_{\text{dec}}(\mathbf{x}) = \log \int p_{\text{dec}}(\mathbf{x}|\mathbf{z})p_{\text{dec}}(\mathbf{z})d\mathbf{z}$ ,

$$\begin{aligned} \mathcal{L}_{\text{rec}} &= \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{z}|\mathbf{x})} \log p_{\text{dec}}(\mathbf{z}|\mathbf{x}) - D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x})||p_{\text{dec}}(\mathbf{z})) \\ &\leq \log p_{\text{dec}}(\mathbf{x}), \end{aligned} \quad (1)$$

with the equality holds iff

$$D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x})||p_{\text{dec}}(\mathbf{z}|\mathbf{x})) = 0. \quad (2)$$

In order to limit the information channel capacity (Higgins et al. (2016)),  $\beta$ -VAE introduces  $\beta > 1$  to the second term of the objective,

$$\begin{aligned} \mathcal{L}_{\text{rec}-\beta} &= \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{z}|\mathbf{x})} \log p_{\text{dec}}(\mathbf{z}|\mathbf{x}) - \\ &\quad \beta D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x})||p_{\text{dec}}(\mathbf{z})) \\ &< \log p_{\text{dec}}(\mathbf{x}). \end{aligned} \quad (3)$$

After training the objective, by sampling from the  $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, I_H)$  or setting  $\mathbf{z}$  with purpose, the learned  $p_{\text{dec}}(\mathbf{x}|\mathbf{z})$  can generate new samples.

## Classification

The  $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$  can further support latter tasks such as classification. Let  $p_{\text{pre}}(\mathbf{y}|\mathbf{z})$  denote the predicting process, and the classification objective is the following,

$$\mathcal{L}_{\text{pre}} = \mathbb{E}_{\mathbf{z} \sim q_{\text{enc}}(\mathbf{z}|\mathbf{x})} \log p_{\text{pre}}(\mathbf{y}|\mathbf{z}). \quad (4)$$

In real implementation the above objectives should further take expectation on the data distribution. However, sometimes only part factors are manually found useful for the generation (Goodfellow, Bengio, and Courville (2016)), and the factor which is irrelevant to  $\mathbf{x}$  can not support classification either. Therefore, some approaches to automatically find the influential factor beneficial to the latter tasks are demanded.

## Mutual Information as A Necessary Indicator

By exploring why factors are ignored, we argue that mutual information is a necessary indicator to find the influential factor.

### Ignored Factor Analysis

**Low Intrinsic Dimension of Data** One aim of the VAE is to learn the data intrinsic factors but intrinsic dimension keeps the same under the continuous reversible mapping suggested by Theorem 1.

**Theorem 1** (Information Conservation). *Suppose that  $\mathbf{z} = (z_1, \dots, z_h)$  and  $\mathbf{y} = (y_1, \dots, y_P)$  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  $f: \mathbb{R}^H \rightarrow \mathbb{R}^P$  and  $g: \mathbb{R}^P \rightarrow \mathbb{R}^H$  such that*

$$\mathbf{z} = g(\mathbf{y}) \quad \text{and} \quad \mathbf{y} = f(\mathbf{z}).$$

Proof is listed in Appendix . Suppose the oracle data, denoted by random variable  $\mathbf{x}$ , is generated by  $\mathbf{y}$  (with  $P$  independent unit Gaussian random variables) with a homeomorphism mapping  $\mathbf{x} = \phi(\mathbf{y})$ . Factors  $\mathbf{z}$  (with  $H$  independent unit Gaussian random variables) generates the  $\mathbf{x}$  with a homeomorphism mapping  $\mathbf{x} = \psi(\mathbf{z})$ . It yields  $\mathbf{z} = \psi^{-1} \circ \phi(\mathbf{y})$  and  $\mathbf{y} = \phi^{-1} \circ \psi(\mathbf{z})$ . Then according to the information conservation theorem, it must hold that  $H = P$ .

For example, 10 Gaussian factors and 128 Gaussian factors can not generate each other. Analogically, if the data are generated by 10 intrinsic Gaussian factors, it intuitively would not be inferred to 128 Gaussian factors by VAE and some factors would be independent with data although we may pre-set in this way.

### Mutual Information Reflexes the Absolute Statistic Dependence

In order to quantify the dependence and estimate which factor influences the generating process or has no effect at all, the mutual information of  $\mathbf{z}_{\text{enc}}$  and  $\mathbf{x}$ ,

$I(\mathbf{x}; \mathbf{z}_{\text{enc}_h})$  can be taken as a rational indicator (Peng, Long, and Ding (2005)). That is,

$$I(\mathbf{x}; \mathbf{z}_{\text{enc}_h}) = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||q_{\text{enc}}(z_h)). \quad (5)$$

The mutual information can reflect the absolute statistic dependence:  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h}) = 0$  if and only if  $\mathbf{x}$  and  $\mathbf{z}_{\text{enc}_h}$  are independent. The larger  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h})$  is, the more information  $z_h$  conveys regarding  $\mathbf{x}$ , and the more influential factor it should be to represent the data.

**Sparsity in Mutual Information** Actually, mutual information is implicitly involved in the VAE objective. The following theorem further suggests the VAE objective induces the sparsity in mutual information. It then explains why factors are ignored from the perspective of mutual information.

**Theorem 2** (Objective Decomposition). *If  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) \ll p_{\text{dec}}(\mathbf{z})^2$ , for any  $\mathbf{x}$ ,  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = q_{\text{enc}}(z_1|\mathbf{x}) \cdots q_{\text{enc}}(z_H|\mathbf{x})$  and  $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, I_H)$  then it yields the following decomposition:*

- $L_1$  norm expression of the KL-divergence term in VAE:

$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x})||p_{\text{dec}}(\mathbf{z})) \\ &= \sum_{h=1}^H \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) \\ &= \left\| \left( \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_1|\mathbf{x})||p_{\text{dec}}(z_1)), \right. \right. \\ & \quad \left. \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_2|\mathbf{x})||p_{\text{dec}}(z_2)), \dots, \right. \\ & \quad \left. \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) \right\|_1. \quad (6) \end{aligned}$$

- Further decomposition of an entity in the  $L_1$  norm expression:

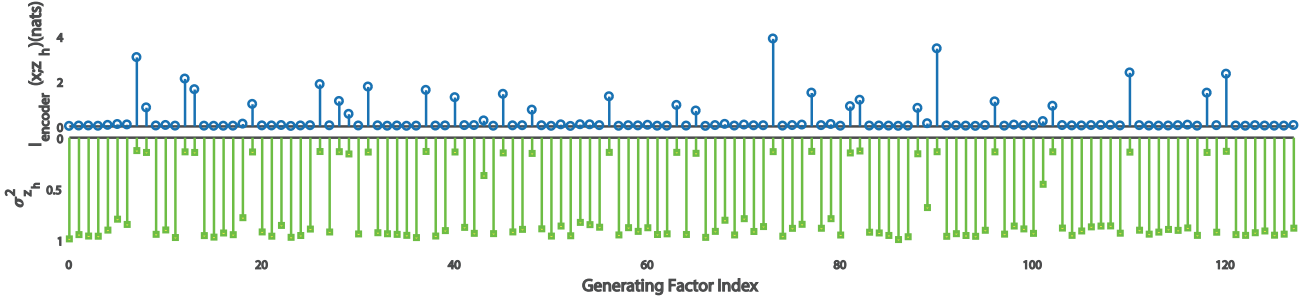
$$\begin{aligned} & \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) \\ &= I(\mathbf{x}; \mathbf{z}_{\text{enc}_h}) + D_{KL}(q_{\text{enc}}(z_h)||p_{\text{dec}}(z_h)). \quad (7) \end{aligned}$$

Proof is given in Appendix . The theorem demonstrates that the expectation of the second term in variation lower bound in Eq. (1) can be represented in the form of  $L_1$  norm which inclines to induce the sparsity of  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h})$  and  $D_{KL}(q_{\text{enc}}(z_h)||p_{\text{dec}}(z_h))$  together in  $h$ , clipping down the non-intrinsic factor dimension to some extent. The sparsity of Expectation  $\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h))$

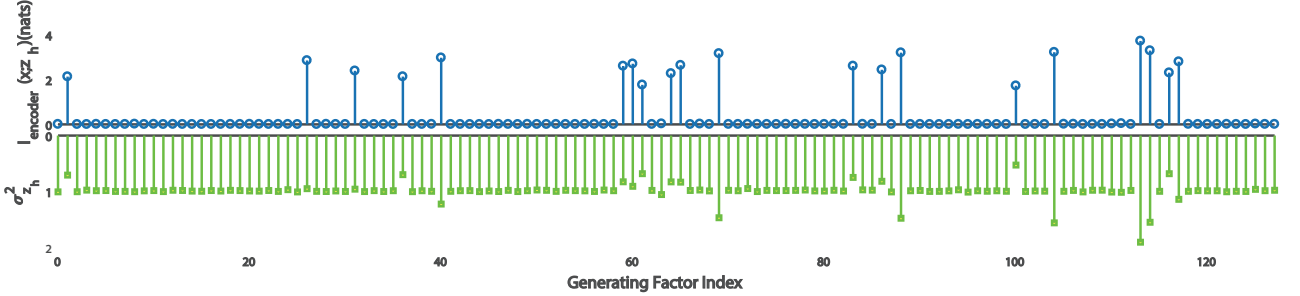
actually leads to sparsity of both its summarization terms  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h})$  and  $D_{KL}(q_{\text{enc}}(z_h)||p_{\text{dec}}(z_h))$  together in  $h$ , since both of them are non-negative. For any zero value summarization, both of its elements should also be zero. Thus this regularization term inclines to intrinsically conduct sparsity of mutual information  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h})$ , which have been comprehensively substantiated by all our experiments, as can be easily seen in Fig.(1) and Fig.(2).

Therefore VAE objective inclines to induce mutual information sparsity in factor dimension over the data intrinsic

<sup>2</sup>That is the support of  $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$  is contained in support of  $p_{\text{dec}}(\mathbf{z})$ .



(A)  $I(\mathbf{x}; z_{\text{enc}_h}), \sigma_{z_h}^2$  plot of  $\beta(=40)$ -VAE on CelebA.



(B)  $I(\mathbf{x}; z_{\text{enc}_h}), \sigma_{z_h}^2$  plot of  $\beta(=6)$ -VAE on Deap.

Figure 2: **Mutual information sparsity occurs on CelebA and Deap.**

dimension and the factor ignored phenomenon occurs. On the one hand, with this KL divergence regularization, even when the number of latent factors are set large, unlike auto-encoder, the over-fitting issue still tends not to occur. On the other hand, this helps us get influential factors to represent the variants of data, and facilitate an efficient generalization of data by varying these useful factors while neglecting others.

By the way, the following theorem suggests the condition that we can use  $I(\mathbf{x}; z_h)$  to estimate the whole mutual information.

**Theorem 3 (Mutual Information Separation).** *Let  $z_1, \dots, z_h$  be independent unit Gaussian distribution, and  $z_1, z_2, \dots, z_h$  be conditional independent given  $\mathbf{x}$ . Then*

$$\begin{aligned} I(\mathbf{x}; z_1, \dots, z_h) &= \sum_{h=1}^H I(\mathbf{x}; z_h) \\ &= \|(I(\mathbf{x}; z_1), I(\mathbf{x}; z_2), \dots, I(\mathbf{x}; z_h))\|_1 \end{aligned} \quad (8)$$

Proof is presented in Appendix . This theorem suggests that if the learnt  $q_{\text{enc}}(\mathbf{z})$  can factorize and the  $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$  can factorize, then we could use the sum of  $I(\mathbf{x}; z_{\text{enc}_h})$  to direct estimate the whole mutual information.

## Reconstruction and Classification Theoretical Supports

According to Cover and Thomas (2012), the mutual information can also provide a lower bound for the best mean recover error.

**Theorem 4.** *Suppose  $\mathbf{x}$  is with differential entropy  $H(\mathbf{x})$ , then let  $\hat{\mathbf{x}}(\mathbf{z}_{\text{enc}})$  be an estimation of  $\mathbf{x}$ , and give side information  $\mathbf{z}_{\text{enc}}$ , and then it holds that*

$$\mathbb{E}(\mathbf{x} - \hat{\mathbf{x}}(\mathbf{z}_{\text{enc}}))^2 \geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}))}. \quad (9)$$

Therefore, if we set  $p_{\text{dec}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\text{dec}(\mathbf{z}), \sigma)$ , then  $\mathbf{x}_{\text{recc}} = \text{dec}(\mathbf{z}_{\text{enc}})$  has  $\frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}))}$  as the lower bound for recover. Let us only use a major set of factors  $\mathbf{z}_{\text{major}}$ :  $\mathbf{x}_{\text{recc}} = \text{dec}(\mathbf{z}_{\text{major}}, \mathbf{0})$ . With the assumption that  $q_{\text{enc}}(\mathbf{z})$  can factorize, it yields the separation of mutual information  $I(\mathbf{x}; \mathbf{z}_{\text{enc}}) = I(\mathbf{x}; \mathbf{z}_{\text{minor}}) + I(\mathbf{x}; \mathbf{z}_{\text{major}})$ . It yields the following bound,

$$\begin{aligned} \mathbb{E}(\mathbf{x} - \mathbf{x}_{\text{recc}})^2 &\geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{major}}))} \\ &\geq \frac{1}{2\pi e} e^{2(H(\mathbf{x}) - I(\mathbf{x}; \mathbf{z}_{\text{major}}))} e^{-2I(\mathbf{x}; \mathbf{z}_{\text{minor}})}. \end{aligned} \quad (10)$$

The theorem implies that the mutual information carried by the selecting factors directly influences on the lower bound of the best recover and we may select some top influential factors carrying the most information to represent and generate the data with less reconstruction distortion.

We further provide some theoretical supports for the proposed mutual information as the factor indicator in classification.

Suppose that Markov chain condition,  $\mathbf{y} \rightarrow \mathbf{x} \rightarrow \mathbf{z}_{\text{enc}} \rightarrow \mathbf{y}_{\text{pre}}$ , holds, according to the Fano's inequality (Cover and

Thomas (2012)) and the information processing inequality the mutual information also correlates with the classification error.

**Theorem 5** (Fano's inequality). *For any estimation  $\hat{\mathbf{y}}$  such that  $\mathbf{y} \rightarrow \mathbf{z}_{\text{enc}} \rightarrow \mathbf{y}_{\text{pre}}$ , with  $P_e = \Pr(\hat{\mathbf{y}} \neq \mathbf{y})$ , we have*

$$H(P_e) + P_e \log |\mathcal{Y}| \geq H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) \geq H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}). \quad (11)$$

*This inequality can be weakened to*

$$1 + P_e \log |\mathcal{Y}| \geq H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) \geq H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}). \quad (12)$$

or

$$P_e \geq \frac{H(\mathbf{y}) - I(\mathbf{y}; \mathbf{z}_{\text{enc}}) - 1}{\log |\mathcal{Y}|} \geq \frac{H(\mathbf{y}) - I(\mathbf{x}; \mathbf{z}_{\text{enc}}) - 1}{\log |\mathcal{Y}|}. \quad (13)$$

Note that according to information processing inequality  $I(\mathbf{x}; z_{\text{enc}h}) \geq I(\mathbf{y}; z_{\text{enc}h})$ . If  $I(\mathbf{x}; z_{\text{enc}h}) = 0 \Rightarrow I(\mathbf{y}; z_{\text{enc}h}) = 0$ , the  $h$  factor will not influence the prediction. With the assumption that  $q_{\text{enc}}(\mathbf{z})$  can factorize, the theorem suggests the mutual information carried by the selecting factors directly influences the lower bound of the classification error and therefore we can remove minor factors according to the mutual information  $I(\mathbf{x}; z_{\text{enc}h})$  without significantly lifting the lower bound of the prediction error.

### Algorithms to Quantitatively Calculate the Proposed indicators

In order to calculate  $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$  in practice, we assume that  $q^*(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \text{diag}(\sigma_1^*, \dots, \sigma_H^*))$  is a factorized zero mean Gaussian estimation for  $q_{\text{enc}}(\mathbf{z})$ .

We can then list the indicators to be estimated as:

**Definition 1** (Estimation for  $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$ : the information conveyed by whole factors).

$$I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q^*(\mathbf{z})). \quad (14)$$

This estimation uses  $M$  sample according to the empirical form of Corollary 2.

**Definition 2** (Estimation for  $I(\mathbf{x}; z_{\text{enc}h})$ : the information conveyed by a factor).

$$I_{\text{est}}(\mathbf{x}; z_{\text{enc}h})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(z_{\text{enc}h}|\mathbf{x}^m) || q^*(z_{\text{enc}h})). \quad (15)$$

This indicator quantifies mutual information of a specific factor and input data.

Note that the above indicators need the value of  $q^*(\mathbf{z})$ , and thus we need to design algorithms to calculate this term. Based on Theorem 2 through the minimization equivalence, we know that

$$\begin{aligned} \min_q \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q(\mathbf{z})) \\ \Leftrightarrow \min_q \int D_{KL}(q_{\text{enc}}(\mathbf{z}) || q(\mathbf{z})) d\mathbf{z}, \end{aligned} \quad (16)$$

and then we can prove the following result:

**Corollary 1.** *if  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) \ll q^*(\mathbf{z})$  then*

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q^*(\mathbf{z})) \\ = I(\mathbf{x}; \mathbf{z}_{\text{enc}}) + D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z})). \end{aligned} \quad (17)$$

The proof of Corollary 1 is the same as that of Theorem 2. This corollary suggests that the estimation defined in Definition 1 provides another upper bound for the capacity of the encoder network. Empirically, this estimation is a much tighter estimation than the second term of the Objective (1).

$q^*(\mathbf{z})$  can then be obtained by solving the following optimization problem:

$$q^*(\mathbf{z}) = \arg \min_q \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q(\mathbf{z})). \quad (18)$$

The above procedure is summarized and presented in Algorithm 1 to calculate the proposed indicators.

---

### Algorithm 1 Mutual Information Estimation

---

- 1: **Input:** Sampled Data  $\{\mathbf{x}^m\}_{m=1}^M$ , Encoder Network  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), \text{diag}(\sigma_1(\mathbf{x}), \dots, \sigma_H(\mathbf{x})))$
  - 2: **Obtain:**  $q^*(\mathbf{z}) = \arg \min_q \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}^m) || q(\mathbf{z}))$ .
  - 3: **Calculate:**  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}h})_M = \frac{1}{M} \sum_{m=1}^M D_{KL}(q_{\text{enc}}(z_{\text{enc}h}|\mathbf{x}^m) || q^*(z_{\text{enc}h}))$
  - 4: **for**  $i = h$  **to**  $H$  **do**
  - 5:  $I_{\text{est}}(\mathbf{x}; z_{\text{enc}h})_M = \frac{1}{M} \sum_{m=1}^M \log \frac{\sigma_h^*}{\sigma_h^2(\mathbf{x}^m) + \mu_h^2(\mathbf{x}^m) / 2\sigma_h^{*2}}$
  - 6: **end for**
  - 7: **Calculate:**  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$
  - 8:  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M = \sum_{h=1}^H I_{\text{est}}(\mathbf{x}; z_{\text{enc}h})_M$ .
  - 9: **Output:**  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M, I_{\text{est}}(\mathbf{x}; z_{\text{enc}h})_M, q^*(\mathbf{z})$
- 

The following definition and theorem clarify the consistency of the estimation on mutual information.

**Definition 3** (Consistency). *The estimator  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$  is consistent to  $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$  if and only if:  $\forall \varepsilon > 0 \forall \delta > 0, \exists N, \text{ and } q^*(\mathbf{z}), \forall M > N, \text{ with probability greater than } 1 - \delta, \text{ we have}$*

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| < \varepsilon. \quad (19)$$

**Theorem 6.** *The estimator  $I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M$  is consistent to  $I(\mathbf{x}; \mathbf{z}_{\text{enc}})$ . That is, if the choice of  $q^*(\mathbf{z})$  satisfied the condition that  $D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z})) < \varepsilon/2$ , then  $\forall \delta > 0, \exists N, \forall M > N, \text{ with probability greater than } 1 - \delta, \text{ we have}$*

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| < \varepsilon. \quad (20)$$

*Proof.* Let  $\tilde{I}[q^*] = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x}) || q^*(\mathbf{z}))$ . According to the law of big number, we have  $\forall \delta > 0, \exists N, \forall M > N, \text{ with probability greater than } 1 - \delta, \text{ we have}$

$$|I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - \tilde{I}[q^*]| < \varepsilon/2, \quad (21)$$

$$\begin{aligned} & |I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| \\ & \leq |I_{\text{est}}(\mathbf{x}; \mathbf{z}_{\text{enc}})_M - \tilde{I}[q^*]| + |\tilde{I}[q^*] - I(\mathbf{x}; \mathbf{z}_{\text{enc}})| \\ & < \frac{\varepsilon}{2} + |D_{KL}(q_{\text{enc}}(\mathbf{z}) || q^*(\mathbf{z}))| < \varepsilon. \end{aligned} \quad (22)$$

□

This theorem suggests the estimation under a high probability could be arbitrary close to the real mutual information provided that the estimation  $q^*(\mathbf{z})$  is arbitrary close to the learned  $q_{\text{enc}}(\mathbf{z})$  and the number of the sample is bigger enough. Besides, the minimization of  $D_{KL}(q_{\text{enc}}(\mathbf{z})||q^*(\mathbf{z}))$  in theorem 6 inspires the derivation of  $q^*(\mathbf{z})$ .

## Related Work

There are not too many works on such indicator designing issue to discover influential factors in VAE. A general and easy approach for determine the VAE’s factor influence is through intuitive visual (Goodfellow, Bengio, and Courville (2016), Higgins et al. (2016)) or aural (Hsu, Zhang, and Glass (2017)) observation. However, it might be labor-intensive to select factors for latter tasks.

In (Alemi et al. (2016)),  $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$  are visualized by plotting the 95% confidence interval as an ellipse to supervise the behavior of network and it reflects the factor influence directly. However, it still needs human to interpret the plot.

In classical PCA, it’s common to select factor with high variance and (Higgins et al. (2017b)) suggests that the variance of factor may indicate the usage of the factors. However, the variance could not always represent the absolute statistical relationship between the factors and data, which can be easily observed by Fig.(1) and Fig.(2).

Our work emphasizes mutual information which conveys the absolute statistical relationship between the factors and the data and uses it as an indicator to find the influential factors, substantiated with the relationship of the total information of selected factors and the reconstruction and relationship of mutual information and classification. All our experiments substantiate that designed indicator can discover the influential factors significantly relevant for data representation.

## Experimental Results

### Datasets

MNIST is a database of handwritten digits (Lcun et al. (1998)). We estimate all mutual information of factors learned from it and then use different ratio of top influential factors for the latter generation task.

CelebA (Liu et al. (2015)) is a large-scale celebfaces attributes datasets and we only use its images to sustain influential factor discovery.

DEAP is a publicly famous multi-modalities emotion recognition dataset proposed by (Koelstra et al. (2012)) and we used the transformed the signal -to-video sequence for 4-class emotion prediction by using different ratio of top influential factors and for emotion relevant influential factor extraction.

More details are presented in Appendix .

### Influential Factor Discovery Tests

According to Fig.(1), the proposed mutual information estimator effectively determines the influential as well as the non-influential factors. The factors with small values of estimated mutual information can be found with little generation effects and factors with large values of mutual informa-

tion can be found with influential generation effects. Comparatively, it can be observed that the variance as used in classical methods can not significantly indicate the usage of factors.

In order to substantiate the validity of our mutual information estimator, we use it to automatically select influential factors with estimated  $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h}) > 0.5$  of CelebA shown in the Fig.(6) and many of them are possess the interpretable variants such as background color, smile and face angle etc. This verifies that mutual information is an effective indicator to automatically determine the influential factors in VAE setting.

### Generation Capability Test for Discovered Factors

Estimated mutual information can instruct the latter generation task with few but influential factors. We select the different ratios of the top influential factors according to the quantity of the mutual information to generate the later image. The factors are sorted according to the values of its mutual information indicators and the other non-influential factors estimated by the indicator are constantly set to zero in the generating process. According to Fig.(4), we can find that by on using 10% of the top influential factors discovered by the proposed algorithm, the VAE model can still generate images almost similar to the one reconstructed by using whole factors.

Table 1 shows the detailed total information and the reconstruction error corresponding to the different ratio of factor. The top 10% factors contain almost the whole information and therefore their reconstructions have the almost the same reconstruct error compared to using all the factors. As suggested by the information and reconstruction relationship, the less information is contained in the used factors, the higher minimum reconstruction loss bound is raised.

### Classification Capability Test by Discovered Factors

Estimated mutual information can instruct the latter classification task with few but influential factors. We select the different ratio of the top influential factors according to the quantity of the mutual information to predict emotions. The factors are sorted according to its mutual information and the estimated non-influential factors are constantly set to zero in the prediction procedure.

According to Table 2, by only using half of the factors, the model still possesses the similar prediction accuracy. Besides, the estimated mutual information on the other side also helps us to determine the several variants which are relevant with the emotion classification as shown in the following Fig.(5).

## Conclusion

This paper explains the necessity of using mutual information of the input data and each factor as the indicator to estimate the intrinsic influence of a factor to represent data in the VAE model. The mutual information reflects the absolute statistical dependence. The second term in VAE objective and excess pre-set factors inclines to induce the mu-

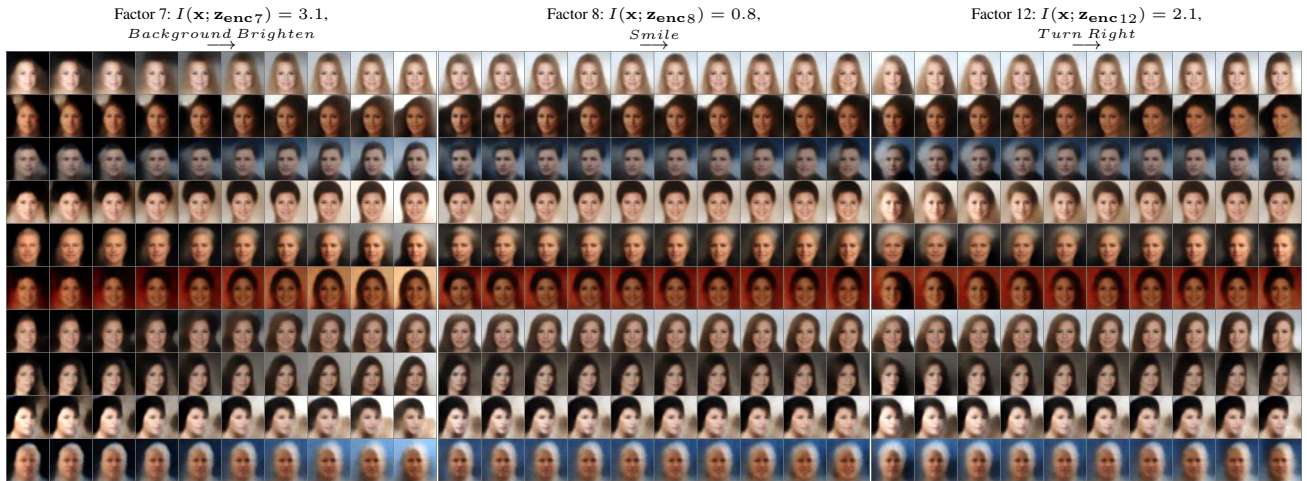


Figure 3: CelebA: Generating Factors Traversal of  $\beta(=40)$ -VAE. We present the first 3 influential factors determined by estimated mutual information. The whole influential factor traversals are listed in appendix 6.

Table 1: mutual information and reconstruction error plot

Top(%) factors in used	100	20	10	7	5	4	3	2	1	0
$I(\mathbf{x}; \mathbf{z}_{\text{enc}_{used}})$	24.3	24.3	24.3	19.6	16.5	14.7	10.6	8.4	5.8	0
mean square error	5.6	5.6	5.6	13.4	15.0	18.9	27.6	31.3	44.4	71.7

Table 2: Mutual information and EEG-emotion classification with  $\beta(=6)$ -VAE

Top(%) factors in used	100	50	10	7	5	4	3	2	1	0
$I(\mathbf{x}; \mathbf{z}_{\text{enc}_{used}})$	53.8	53.5	38.3	28.0	22.5	19.6	13.5	10.2	7.0	0
mean test accuracy	0.53	0.52	0.46	0.32	0.34	0.36	0.29	0.29	0.3	0.23

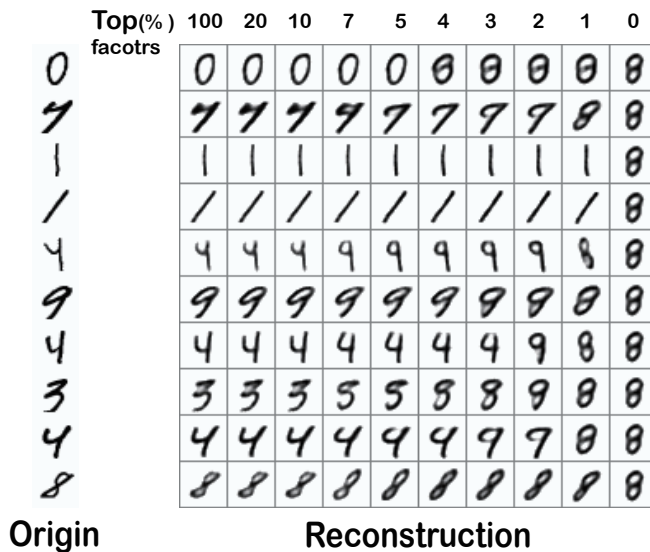


Figure 4: Generation plot with different ratio of factors.

tual information sparsity and helps achieve influential, as well as ignored, factors in VAE. We have also proved that the mutual information also involves in the lower bound of the mean square error of the reconstruction and of prediction error of the classification. We design a feasible algorithm to calculate the indicator for estimating the mutual information for all factors in VAE and proves its consistency. The experiments show that both the influential factors and non-influential factors can be automatically and effectively found. The interpretability of the discovered factors is substantiated intuitively, and the generalization and classification capability on these factors have also been verified. Specially, some variants relevant to classification are found. The experiments also inspire the idea that we can using a small amount of top influential factors for the latter data processing tasks including generation and classification by still keeping the performance of all factors, just similar to the dimensionality reduction capability as classical PCA, ICA and so on.

The VAE combined with mutual information indicator helps extract knowledge under the data and may be beneficial to extensive latter applications including blind source separation, interpretable feature learning, information bottleneck and data bias elimination. We will investigate these issues in our future research.

### Acknowledgments

We would like to thank Zilu Ma and Tao Yu for discussing the information conservation theorems. We would like to thank Lingjiang Xie and Rui Qin for EEG data processing.

### References

Alemi, A. A.; Fischer, I.; Dillon, J. V.; and Murphy, K. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Bashivan, P.; Rish, I.; Yeasin, M.; and Codella, N. 2015. Learning representations from eeg with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*.

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

Goodfellow, I.; Bengio, Y.; and Courville, A. 2016. *Deep learning*. MIT press.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework.

Higgins, I.; Pal, A.; Rusu, A. A.; Matthey, L.; Burgess, C. P.; Pritzel, A.; Botvinick, M.; Blundell, C.; and Lerchner, A. 2017a. Darla: Improving zero-shot transfer in reinforcement learning. *arXiv preprint arXiv:1707.08475*.

Higgins, I.; Sonnerat, N.; Matthey, L.; Pal, A.; Burgess, C. P.; Botvinick, M.; Hassabis, D.; and Lerchner, A. 2017b. Scan: Learning abstract hierarchical compositional visual concepts. *arXiv preprint arXiv:1707.03389*.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, 1876–1887.

Hyvärinen, A.; Karhunen, J.; and Oja, E. 2004. *Independent component analysis*, volume 46. John Wiley & Sons.

Jung, T.-P.; Makeig, S.; Humphries, C.; Lee, T.-W.; Mckeown, M. J.; Iragui, V.; and Sejnowski, T. J. 2000. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37(2):163–178.

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

Koelstra, S.; Muhl, C.; Soleymani, M.; Lee, J.-S.; Yazdani, A.; Ebrahimi, T.; Pun, T.; Nijholt, A.; and Patras, I. 2012. Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3(1):18–31.

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

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, 3730–3738.

Makeig, S.; Bell, A. J.; Jung, T.-P.; and Sejnowski, T. J. 1996. Independent component analysis of electroencephalographic data. In *Advances in neural information processing systems*, 145–151.

Mathieu, M. F.; Zhao, J. J.; Zhao, J.; Ramesh, A.; Sprechmann, P.; and LeCun, Y. 2016. Disentangling factors of variation in deep representation using adversarial training. In *Advances in Neural Information Processing Systems*, 5040–5048.

Peng, H.; Long, F.; and Ding, C. 2005. Feature selection based on mutual information criteria of max-dependency,



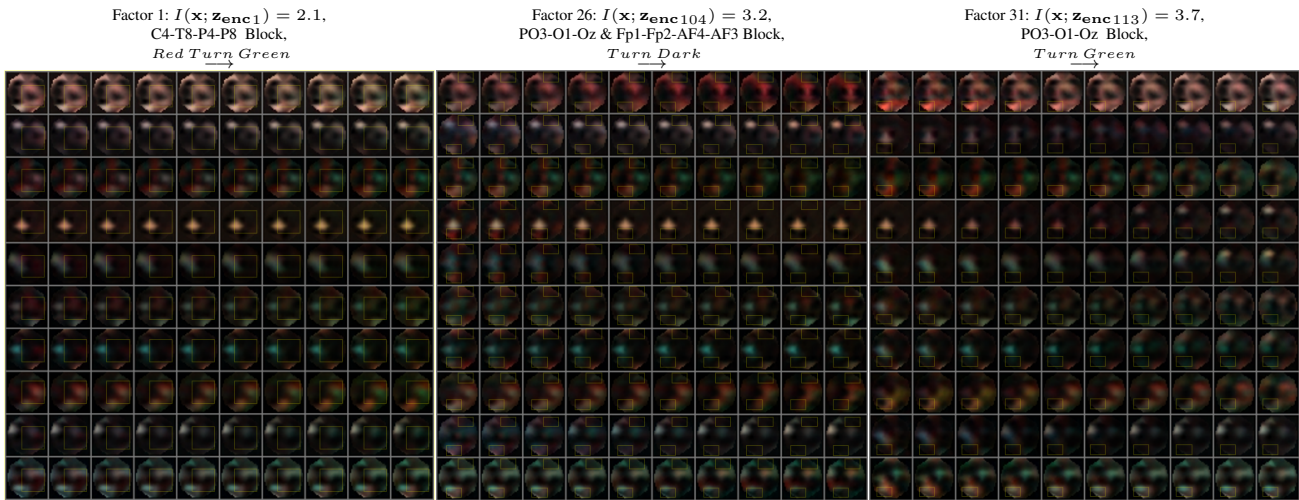


Figure 5: Emotion relevant factors discovery. We present 3 influential factors determined by estimated mutual information. The whole influential factor traversals are listed in Fig.(8) in appendix.

max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence* 27(8):1226–1238.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

Salakhutdinov, R., and Hinton, G. 2009. Semantic hashing. *International Journal of Approximate Reasoning* 50(7):969–978.

Suzuki, M.; Nakayama, K.; and Matsuo, Y. 2016. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*.

Yang, J.; Zhang, D.; Frangi, A. F.; and Yang, J.-y. 2004. Two-dimensional pca: a new approach to appearance-based face representation and recognition. *IEEE transactions on pattern analysis and machine intelligence* 26(1):131–137.

## Appendix A

**Theorem 1 (Information Conservation)** Suppose that  $\mathbf{z} = (z_1, \dots, z_h)$  and  $\mathbf{y} = (y_1, \dots, y_P)$  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  $f : \mathbb{R}^H \rightarrow \mathbb{R}^P$  and  $g : \mathbb{R}^P \rightarrow \mathbb{R}^H$  such that

$$\mathbf{z} = g(\mathbf{y}) \quad \text{and} \quad \mathbf{y} = f(\mathbf{z}).$$

*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.

$$\begin{aligned} \mathbf{z} &= g(\mathbf{y}) = g(f(\mathbf{z})) \quad \forall \mathbf{z} \in \mathbb{R}^H \Rightarrow g \circ f = I_H \\ \mathbf{y} &= f(\mathbf{z}) = f(g(\mathbf{y})) \quad \forall \mathbf{y} \in \mathbb{R}^P \Rightarrow f \circ g = I_P \end{aligned}$$

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.  $\square$

**Theorem 2 (Objective Decomposition)** If  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) \ll p_{\text{dec}}(\mathbf{z})$ , i.e., the support of  $q_{\text{enc}}(\mathbf{z}|\mathbf{x})$  is contained in support of  $p_{\text{dec}}(\mathbf{z})$ , for any  $\mathbf{x}$ ,  $q_{\text{enc}}(\mathbf{z}|\mathbf{x}) = q_{\text{enc}}(z_1|\mathbf{x}) \cdots q_{\text{enc}}(z_H|\mathbf{x})$  and  $p_{\text{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, I_H)$  then it yields the following decomposition:

- $L_1$  norm expression of the KL-divergence term in VAE:

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(\mathbf{z}|\mathbf{x})||p_{\text{dec}}(\mathbf{z})) &= \sum_{h=1}^H \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) = \\ &= \left( \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_1|\mathbf{x})||p_{\text{dec}}(z_1)), \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_2|\mathbf{x})||p_{\text{dec}}(z_2)), \dots, \right. \\ &\quad \left. \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) \right) \|_1. \end{aligned} \quad (23)$$

- Further decomposition of an entity in the  $L_1$  norm expression:

$$\mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) = I(\mathbf{x}; z_{\text{enc}h}) + D_{KL}(q_{\text{enc}}(z_h)||p_{\text{dec}}(z_h)). \quad (24)$$

*Proof.* The  $L_1$  norm expression is obvious. We prove the further decomposition of an entity in the  $L_1$  norm expression:

$$\begin{aligned} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} D_{KL}(q_{\text{enc}}(z_h|\mathbf{x})||p_{\text{dec}}(z_h)) &= \int q_{\text{enc}}(z_h|\mathbf{x}) p_{\text{data}}(\mathbf{x}) \frac{q_{\text{enc}}(z_h|\mathbf{x}) p_{\text{data}}(\mathbf{x})}{p_{\text{dec}}(z_h) p_{\text{data}}(\mathbf{x})} d\mathbf{x} \\ &= \int q_{\text{enc}}(z_h|\mathbf{x}) p_{\text{data}}(\mathbf{x}) \frac{q_{\text{dec}}(z_h|\mathbf{x}) p_{\text{data}}(\mathbf{x})}{q_{\text{enc}}(z_h) p_{\text{data}}(\mathbf{x})} \frac{q_{\text{enc}}(z_h)}{p_{\text{dec}}(z_h)} d\mathbf{x} = I(\mathbf{x}; z_{\text{enc}h}) + D_{KL}(q_{\text{enc}}(z_h)||p_{\text{dec}}(z_h)). \end{aligned} \quad (25)$$

$\square$

**Theorem 3 (Mutual Information Separation)** Let  $z_1, \dots, z_h$  be independent unit Gaussian distribution, and  $z_1, z_2, \dots, z_h$  be conditional independent given  $\mathbf{x}$ . Then

$$I(\mathbf{x}; z_1, \dots, z_h) = \sum_{h=1}^H I(\mathbf{x}; z_h) = \|(I(\mathbf{x}; z_1), I(\mathbf{x}; z_2), \dots, I(\mathbf{x}; z_h))\|_1. \quad (26)$$

*Proof.*

$$\begin{aligned} I(\mathbf{x}; z_1, \dots, z_H) &= \int p(z_1, \dots, z_H, \mathbf{x}) \log \frac{p(\mathbf{x}, z_1, \dots, z_H)}{p(z_1, \dots, z_H) p(\mathbf{x})} dz_1 \cdots dz_H d\mathbf{x} \\ &= \int p(\mathbf{x}, z_1, \dots, z_H) \log \frac{\prod_{h=1}^H p(z_h|\mathbf{x})}{\prod_{h=1}^H p(z_h)} dz_1 \cdots dz_H d\mathbf{x} = \sum_{h=1}^H \int p(\mathbf{x}, z_h) \log \frac{p(z_h|\mathbf{x})}{p(z_h)} dz_h d\mathbf{x} \\ &= \sum_{h=1}^H I(\mathbf{x}; z_h). \end{aligned}$$

$\square$

## Appendix B

### Mnist

We split 7000 data points by ratio [0.6 : 0.2 : 0.2] into training, validation, testing set. The estimated mutual information and  $q^*(\mathbf{z})$  are calculated on 10000 data points in the testing set. Seed images from the testing set are used to infer factor value and draw the traversal.

In traversal figures, each block corresponds to the traversal of a single factor over the  $[-3, 3]$  range while keeping others fixed to their inferred (by  $\beta$ -VAE, VAE). Each row is generated with a different seed image.

The  $\beta$  setting for  $\beta$ -VAE is enumerated from [0.1, 0.5, 1, 2 : 2 : 18].

### CelebA

We split randomly roughly 200000 data points by ratio [0.8 : 0.1 : 0.1] into training, validation (no use), testing set.

The estimated mutual information and  $q^*(\mathbf{z})$  are calculated on 10000 data points in the testing set. Seed images from the testing set are used to infer factor value and draw the traversal.

In traversal figures, each block corresponds to the traversal of a single factor over the  $[-3, 3]$  range while keeping others fixed to their inferred (by  $\beta$ -VAE, VAE). Each row is generated with a different seed image.

The  $\beta$  setting for  $\beta$ -VAE is enumerated from [1, 30, 40].

### Deap

DEAP is a well-known public multi-modalities (e.g. EEG, video, etc.) dataset proposed by Koelstra et al. (2012). The EEG signals are recorded from 32 channels by 32 participants watching 40 videos for 63 seconds each. The EEG data was pre-processed which down-sampling into 128Hz and band range 4-45 Hz. By the same transformation idea from Bashivan et al. (2015), we applied fast Fourier transform (FFT) on 1-second EEG signal and convert it to an image. In this experiment, alpha (8-13Hz), beta (13-30Hz) and gamma (30-45Hz) are extracted as the frequency band which represented the activities related to brain emotion emerging. The next step is similar as Bashivan et al. (2015) work which mentioned in section II [PLEASE CHECK IT IN THE PAPER], by Azimuthal Equidistant Projection (AEP) and Clough-Tocher scheme resulting in three 32x32 size topographical activity maps corresponding to each frequency bands shown as RGB plot. The transformation work conduct the total of 1280 EEG videos where each has 63 frames. The two emotional dimensions are arousal and valence, which were labeled from the scale 1-9. For each of them, we applied 5 as the boundary for separating high and low level to generate 4 classes (e.g. high-arousal (HA), high-valence (HV), low-arousal (LA) and low-valence(LV)). In this paper we perform this 4-class classification task as same as the one in [baseline paper].

We split randomly roughly 1280 samples by ratio [0.8: 0.1: 0.1] into training, validation, testing set.  $\beta(= 6)$ -VAE is trained on each frame and LSTM was used to combine all the frames together for each video.

The estimated mutual information are calculated on 100\*63 imagewise(100 videos) data points in the testing set. Seed images from the testing set are used to infer factor value and draw the traversal.

In traversal figures, each block corresponds to the traversal of a single factor over the  $[-3, 3]$  range while keeping others fixed to their inferred (by  $\beta$ -VAE). Each row is generated with a different seed image.

### Network Structure

Dataset	Optimiser	Architecture	
Mnist	Adam $1e-3$	Input Encoder	28x28x1 Conv 32x4x4,32x4x4 (stride 2). FC 256. ReLU activation.
	Epoch 200	Latents Decoder	128 FC 256. Linear. Deconv reverse of encoder. ReLU activation. Gaussian.
CelebA	Adam $1e-4$	Input Encoder	64x64x3 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (stride 2). FC 256. ReLU activation.
	Epoch 20	Latents Decoder	128/32 FC 256. Linear. Deconv reverse of encoder. ReLU activation. Mixture of 2-Gaussian.
Deap	Adam $1e-4$	Input Encoder	32x32x3 Conv 32x4x4,32x4x4,64x4x4,64x4x4 (stride 2). FC 256. ReLU activation.
	Epoch 300	Latents Decoder	128/32 FC 256. Linear. Deconv reverse of encoder. ReLU activation. Gaussian.
		Input Recurrent Predictor	63x128 LSTM dim128. Time-Step 63. FC 4. ReLU activation.

## Experiment Plot

In the following subsection, we present the influential factor ( $I(\mathbf{x}; \mathbf{z}_{\text{enc}_h}) > 0.7$ ) traversals, mutual information and variance plot of different data sets.

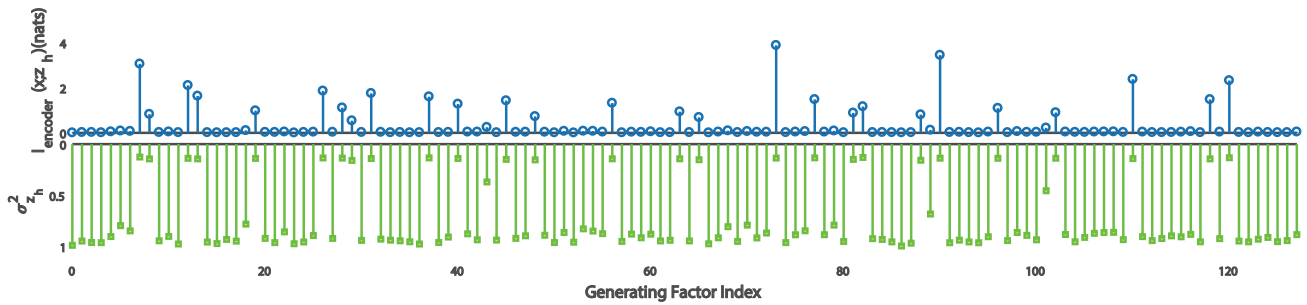
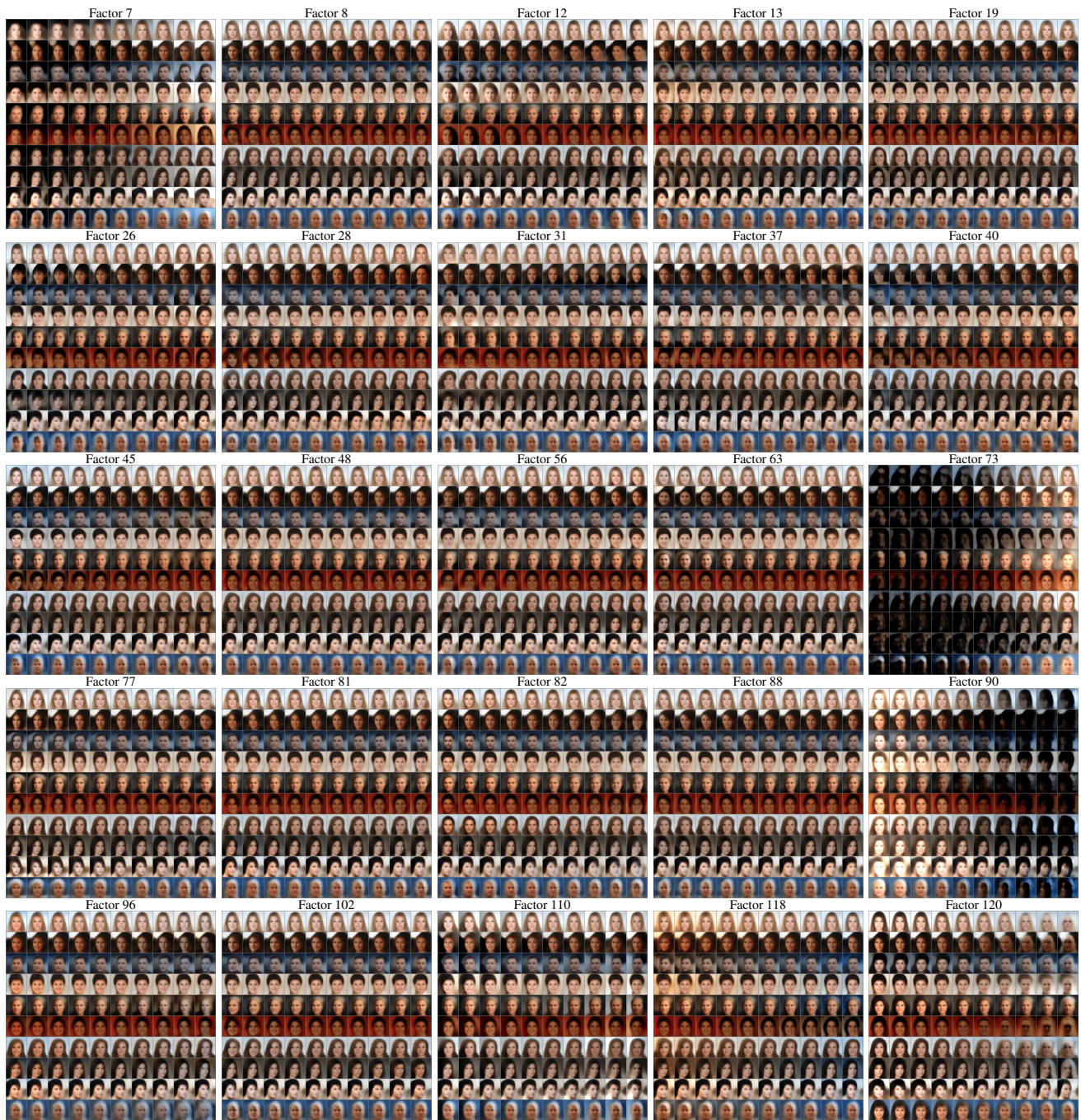


Figure 6: Mutual Information Sparsity in CelebA: Generating Factors Traversal of  $\beta(=40)$ -VAE

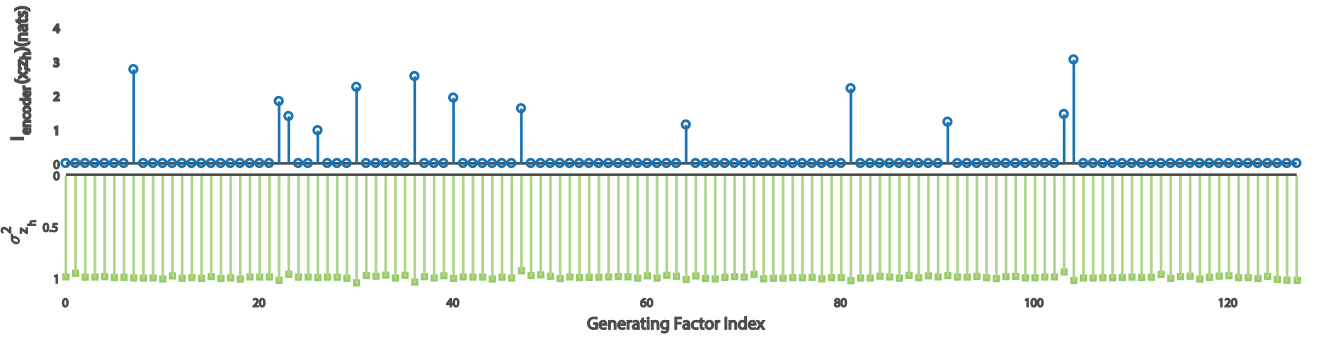
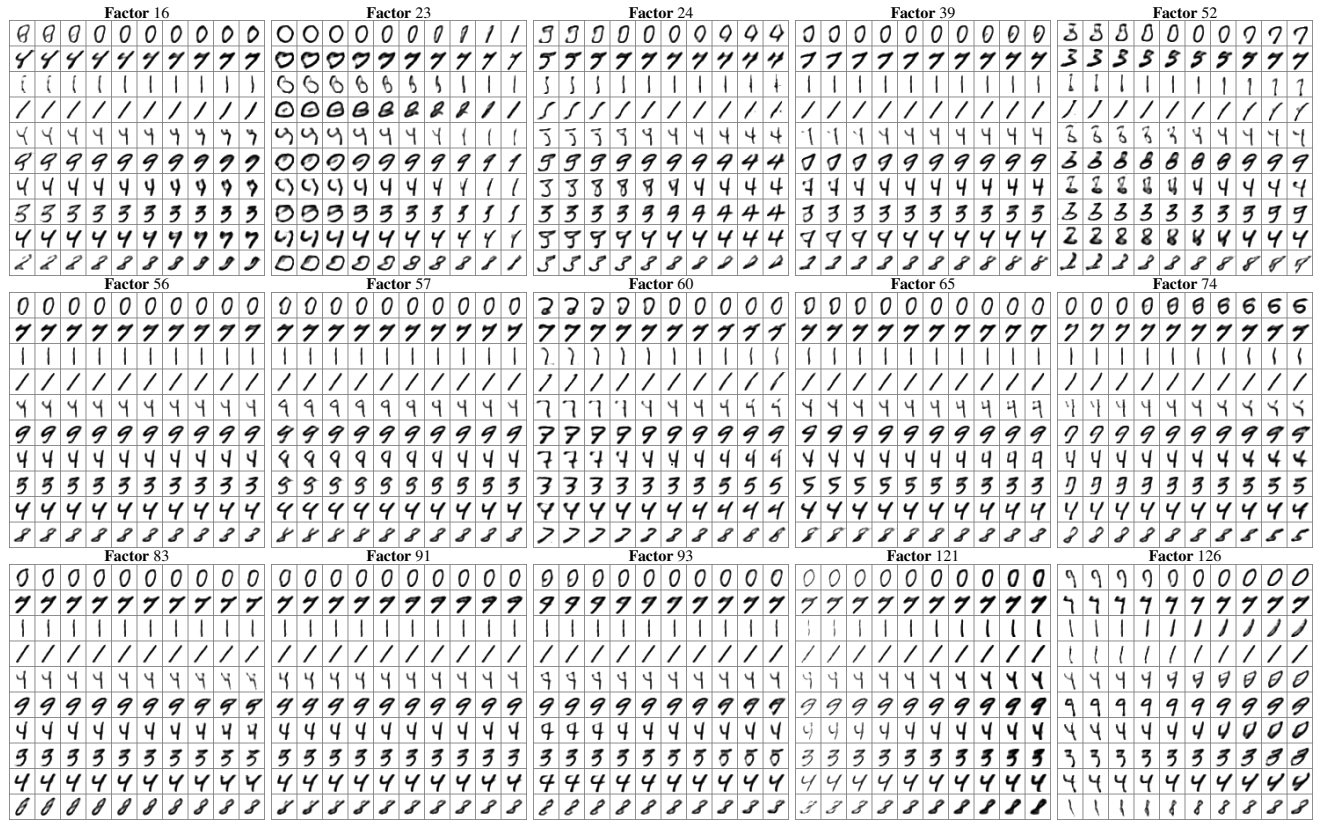


Figure 7: Mutual Information Sparsity in MNIST: Generating Factor Traversal of  $\beta(=10)$ -VAE

