Discovering Influential Factors in Variational Autoencoders

Shiqi Liu[a,1], Jingxin Liu[b,1], Qian Zhao[a], Xiangyong Cao[a], Huibin Li[a], Deyu Meng[a], Hongying Meng[b], Sheng Liu[c]

[a]*Xi'an Jiaotong University, Xi'an, Shaan'xi Province, P. R. China*
[b]*Brunel University London, London, United Kingdom*
[c]*State University of New York at Buffalo, NY 14214, United States*

Abstract

In the field of machine learning, it is still a critical issue to identify and supervise the learned representation without manually intervening or intuition assistance to extract useful knowledge or serve for the downstream tasks. 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 of discovering the influential factors. We find the VAE objective inclines to induce mutual information sparsity in factor dimension over the data intrinsic dimension and therefore result 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 downstream classification task. To make such indicator applicable, we design an algorithm for 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.

Keywords: Variational Autoencoder, Mutual Information, Generative Model

[1] Shiqi Liu and Jingxin Liu made equal contributions to this work.

1. 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 ([1],[2])) has been widely used in primary data analysis and its variant has been applied in face identification ([3]) and classical linear independent representation (Independent Component Analysis ([4],[5]) have been used in blind source separation ([6]) and EEG signal processing ([7]). Nonlinear dimension reduction (e.g. Autoencoder [8],[9],[10]) begins to further learn abstract representation([11],[12]) and has been used in semantic hashing ([13] and many other tasks([14],[15],[16]). Recently, a new technique, called variational autoencoder ([17, 18]) 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 ([19],[20],[21])[2] and interpretable time series variants ([23],[24]) and this method can serve for generating signals with abundant diversities in a "factor-controllable" way ([25],[26],[27]). The related techniques enable the knowledge transferring through shared factors among different tasks ([28]).

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 [3] ([8],[29]), 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

[2] According to [22], although the method proves effective at making the disentangled factors not correlated, the learned disentangled factors are still correlated with each other.

[3] Montage (D) in Fig.(1) is a typical traversal of the unused factor.

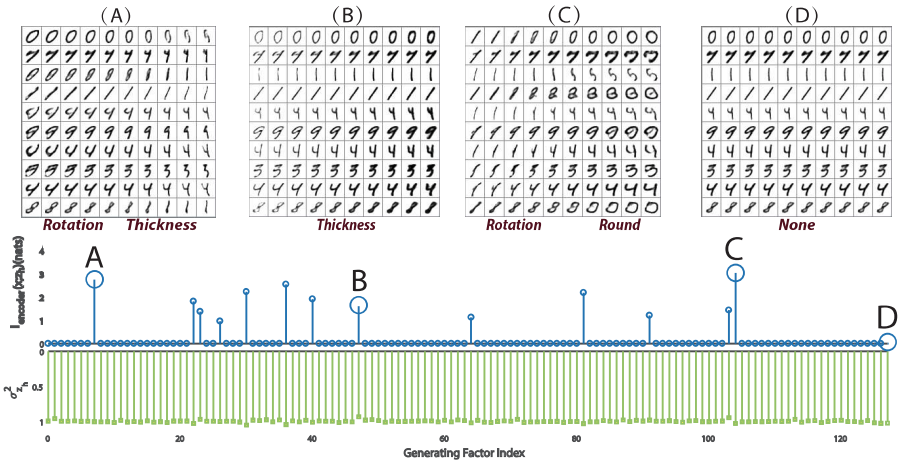


Figure 1: **Estimated $I(\mathbf{X}; Z_{\mathbf{enc}h})$ determines the influential factors;** $I(\mathbf{X}; Z_{\mathbf{enc}h})$, $\sigma^2_{z_h}$ and qualitatively influential factor traversals of $\beta(= 10)$-VAE on MNIST. The top pulse subgraph: $I(\mathbf{X}; Z_{\mathbf{enc}h})$ of each factor. The bottom reverse pulse subgraph: the estimated variance $\sigma^2_{z_h}$ 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.(A.7) in Appendix Appendix A. The montages 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.

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 ([30]), CelebA ([31]) and DEAP ([32]). Especially, some discovered

factors by the proposed indicator 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.

Throughout the paper, we denote a random variable in upper case, e.g., Z; a random vector in bold upper case, e.g., $\mathbf{Z}$, whose h^{th} component is denoted as Z_h; a general vector as bold lower case, e.g., $\mathbf{z}$. More notations appearing in the following contents are listed in Table 1 for easy reference.

2. VAE model

VAE ([17], [18]) is a scalable unsupervised representation learning model ([33]): VAE assumes that input $\mathbf{X}$ is generated by several independent Gaussian random variables $\mathbf{Z}$, that is $p_{\mathbf{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, I_H)$. 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 $p_{\mathbf{dec}}(\mathbf{x}|\mathbf{z})$ and the inference/encoding process $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), diag(\sigma_1(\mathbf{x}), \cdots, \sigma_H(\mathbf{x})))$ is treated as the approximate posterior distribution. Note that it yields that $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) = q_{\mathbf{enc}}(z_1|\mathbf{x}) \cdots q_{\mathbf{enc}}(z_H|\mathbf{x})$. We assume both of them are parameterized by the neural network with parameter $\mathbf{enc}$ and $\mathbf{dec}$.

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

Notation	Explanation
$\mathbf{X}$	random variables representing the data
$\mathbf{Y}$	random variables representing the data label
$\mathbf{Z}$	random variables with $p_{\mathbf{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}\|\mathbf{0}, I_H)$
$\mathbf{Z_{enc}}$	random variables with $q_{\mathbf{enc}}(\mathbf{z}) = \int q_{\mathbf{enc}}(\mathbf{z}\|\mathbf{x})p_{data}(\mathbf{x})d\mathbf{x}$
$Z_{\mathbf{enc}h}$	a random variable with index h of $\mathbf{Z_{enc}}$
$\mathbf{Y_{pre}}$	random variables with $p_{\mathbf{pre}}(\mathbf{y}) = \int p_{\mathbf{pre}}(\mathbf{y}\|\mathbf{z})q_{\mathbf{enc}}(\mathbf{z}\|\mathbf{x})p_{data}(\mathbf{x})d\mathbf{x}$
$\hat{\mathbf{X}}(\mathbf{Z_{enc}})$	a function named $\hat{\mathbf{X}}$ of $\mathbf{Z_{enc}}$, the estimator of $\mathbf{X}$
$H(\mathbf{X})$	$\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})} -\log p(\mathbf{x})$
$I(\mathbf{X};\mathbf{Z})$	$\mathop{\mathbb{E}}_{\mathbf{x}\sim p(\mathbf{x})} D_{KL}(p(\mathbf{z}\|\mathbf{x})\|\|p(\mathbf{z}))$
$\mathbf{X}_{rec}$	$\mathbf{dec}_\mu(\mathbf{Z_{enc}})$ where $\mathbf{dec}_\mu$ is such that $p_{\mathbf{dec}}(\mathbf{x}\|\mathbf{z}) = \mathcal{N}(\mathbf{x}\|\mathbf{dec}_\mu(\mathbf{z}), \mathbf{dec}_\sigma)$
$\mathbf{Z}_{major}$	a major set of $\mathbf{Z_{enc}}$ where $\mathbf{Z_{enc}} = [\mathbf{Z}_{major}, \mathbf{Z}_{minor}]$
$\mathbf{Z}_{minor}$	a minor set of $\mathbf{Z_{enc}}$ where $\mathbf{Z_{enc}} = [\mathbf{Z}_{major}, \mathbf{Z}_{minor}]$
$\mathbf{X}_{recc}$	$\mathbf{dec}_\mu(\mathbf{Z}_{major}, \mathbf{0})$
$\hat{\mathbf{Y}}$	a random variable, the estimator of $\mathbf{Y}$

Table 1: Notation Table

2.1. Generation

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

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

with the equality holds iff

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

In order to limit the information channel capacity ([33]), β-VAE introduces $\beta > 1$

to the second term of the objective,

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

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

2.2. Classification

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

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

Let $\mathbf{Y}_{\mathbf{pre}}$ denote the random variables with $p_{\mathbf{pre}}(\mathbf{y}) = \int p_{\mathbf{pre}}(\mathbf{y}|\mathbf{z}) q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) p_{data}(\mathbf{x}) d\mathbf{x}$.

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 ([8]), 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.

3. 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.

3.1. Ignored Factor Analysis

3.1.1. 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 there are two sets of H and P ($H \neq P$), $\mathbf{Z} = (Z_1, \cdots, Z_H)$ and $\mathbf{Y} = (Y_1, \cdots, Y_P)$, respectively, independent unit Gaussian random variables, 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*

$$\mathbf{Z} = g(\mathbf{Y}) \quad and \quad \mathbf{Y} = f(\mathbf{Z}).$$

Proof. Proof by Contradiction. Suppose those two function exist, and we will show that they will be inverse mapping of each other and f is a homeomorphism mapping of $\mathbb{R}^H$ and $\mathbb{R}^P$. Here, f is said to be a homeomorphism mapping if it satisfies the following three conditions:

- f is a bijection,
- f is continuous,
- the inverse function f^{-1} is continuous.

Since $\mathbb{R}^H$ and $\mathbb{R}^P$ have different topology structures ($P \neq H$), the homeomorphism mapping will not exist.

$$\mathbf{Z} = g(\mathbf{Y}) = g(f(\mathbf{Z})) \; \forall \mathbf{Z} \in \mathbb{R}^H \Rightarrow g \circ f = I_H$$

$$\mathbf{Y} = f(\mathbf{Z}) = f(g(\mathbf{Y})) \; \forall \mathbf{Y} \in \mathbb{R}^P \Rightarrow f \circ g = I_P$$

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

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.

3.1.2. 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 $\mathbf{z_{enc}}$ and $\mathbf{x}$, $I(\mathbf{X}; Z_{\mathbf{enc}h})$ can be taken as a rational indicator([34]) . That is,

$$I(\mathbf{X}; Z_{\mathbf{enc}h}) = \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(z_h|\mathbf{x})||q_{\mathbf{enc}}(z_h)). \tag{5}$$

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

3.1.3. 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* $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) \ll p_{\mathbf{dec}}(\mathbf{z})$[4]*, for any* $\mathbf{x}$*,* $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) = q_{\mathbf{enc}}(z_1|\mathbf{x}) \cdots q_{\mathbf{enc}}(z_H|\mathbf{x})$ *and* $p_{\mathbf{dec}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{0}, I_H)$ *then it yields the following decomposition:*

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

[4]That is the support of $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x})$ is contained in support of $p_{\mathbf{dec}}(\mathbf{z})$.

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

- *Further decomposition of an entity in the L_1 norm expression*[5]*:*

$$
\begin{aligned}
& \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(z_h|\mathbf{x})||p_{\mathbf{dec}}(z_h)) \\
= & I(\mathbf{X}; Z_{\mathbf{enc}h}) + D_{KL}(q_{\mathbf{enc}}(z_h)||p_{\mathbf{dec}}(z_h)). \qquad (7)
\end{aligned}
$$

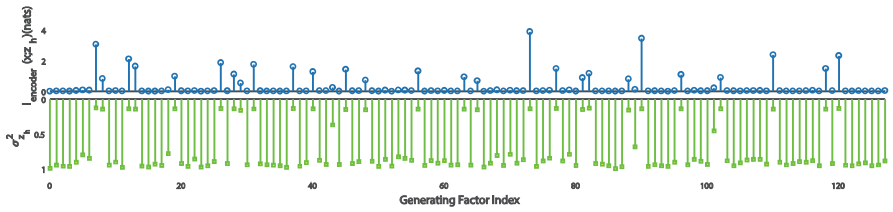
Proof. The L_1 norm expression is obvious. We prove the further decomposition of an entity in the L_1 norm expression:

$$
\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(z_h|\mathbf{x})||p_{\mathbf{dec}}(z_h)) &= \int q_{\mathbf{enc}}(z_h|\mathbf{x}) p_{data}(\mathbf{x}) \frac{q_{\mathbf{enc}}(z_h|\mathbf{x}) p_{data}(\mathbf{x})}{p_{\mathbf{dec}}(z_h) p_{data}(\mathbf{x})} d\mathbf{x} \\
&= \int q_{\mathbf{enc}}(z_h|\mathbf{x}) p_{data}(\mathbf{x}) \frac{q_{\mathbf{dec}}(z_h|\mathbf{x}) p_{data}(\mathbf{x})}{q_{\mathbf{enc}}(z_h) p_{data}(\mathbf{x})} \frac{q_{\mathbf{enc}}(z_h)}{p_{\mathbf{dec}}(z_h)} d\mathbf{x} \\
&= I(\mathbf{X}; Z_{\mathbf{enc}h}) + D_{KL}(q_{\mathbf{enc}}(z_h)||p_{\mathbf{dec}}(z_h)). \qquad (8)
\end{aligned}
$$

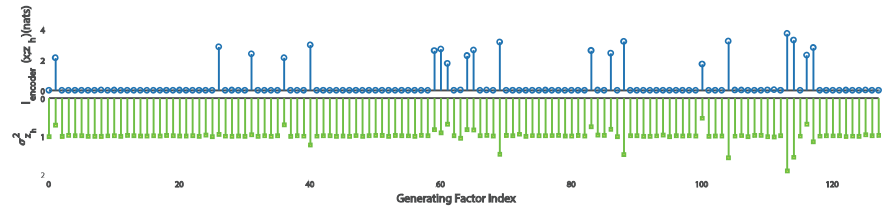
□

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 $I(\mathbf{X}; Z_{\mathbf{enc}h})$ and $D_{KL}(q_{\mathbf{enc}}(z_h)||p_{\mathbf{dec}}(z_h))$ together in h, clipping down the non-intrinsic factor dimension to some extent. The sparsity of Expectation $\mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(z_h|\mathbf{x})||p_{\mathbf{dec}}(z_h))$ actually leads to sparsity of both its summarization terms $I(\mathbf{X}; Z_{\mathbf{enc}h})$ and $D_{KL}(q_{\mathbf{enc}}(z_h)||p_{\mathbf{dec}}(z_h))$ 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

[5] It is similar to the result in [35].



(A) $I(\mathbf{X}; Z_{\mathbf{enc}_h})$, $\sigma^2_{z_h}$ plot of $\beta(=40)$-VAE on CelebA.



(B) $I(\mathbf{X}; Z_{\mathbf{enc}_h})$, $\sigma^2_{z_h}$ plot of $\beta(=6)$-VAE on DEAP.

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

conduct sparsity of mutual information $I(\mathbf{X}; Z_{\mathbf{enc}_h})$, 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 dimension and the factor ignored phenomenon occurs. On the one hand, with increase in the KL divergence regularization, even when the number of latent factors is 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 $Z_1, \cdots, Z_H$ be independent unit*

Gaussian distribution, and $Z_1, Z_2, \cdots, Z_H$ be conditional independent given $\mathbf{X}$. Then

$$\begin{aligned} I(\mathbf{X}; Z_1, \cdots, Z_H) &= \sum_{h=1}^{H} I(\mathbf{X}; Z_h) \\ &= \|(I(\mathbf{X}; Z_1), I(\mathbf{X}; Z_2), \cdots, I(\mathbf{X}; Z_H))\|_1. \end{aligned} \tag{9}$$

Proof.

$$\begin{aligned} I(\mathbf{X}; Z_1, \cdots, Z_H) &= \int p(z_1, \cdots, z_H, \mathbf{x}) \log \frac{p(\mathbf{x}, z_1, \cdots, z_H)}{p(z_1, \cdots, z_H)p(\mathbf{x})} dz_1 \cdots dz_H dx \\ &= \int p(\mathbf{x}, z_1, \cdots, z_H) \log \frac{\prod_{h=1}^{H} p(z_h|\mathbf{x})}{\prod_{h=1}^{H} p(z_h)} dz_1 \cdots dz_H d\mathbf{x} \\ &= \sum_{h=1}^{H} \int p(\mathbf{x}, z_h) \log \frac{p(z_h|\mathbf{x})}{p(z_h)} dz_h d\mathbf{x} = \sum_{h=1}^{H} I(\mathbf{X}; Z_h). \end{aligned}$$

□

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

3.2. Reconstruction and Classification Theoretical Supports

According to [36], 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 $\hat{\mathbf{X}}(\mathbf{Z_{enc}})$ be an estimation of $\mathbf{X}$, and give side information $\mathbf{Z_{enc}}$[6], and then it holds that*

$$\mathbb{E}(\mathbf{X} - \hat{\mathbf{X}}(\mathbf{Z_{enc}}))^2 \geq \frac{1}{2\pi e} e^{2(H(\mathbf{X}) - I(\mathbf{X}; \mathbf{Z_{enc}}))}. \tag{10}$$

Therefore, if we set $p_{\mathbf{dec}}(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\mathbf{dec}_\mu(\mathbf{z}), \mathbf{dec}_\sigma)$, then $\mathbf{X}_{rec} = \mathbf{dec}_\mu(\mathbf{Z_{enc}})$ has $\frac{1}{2\pi e} e^{2(H(\mathbf{X}) - I(\mathbf{X}; \mathbf{Z_{enc}}))}$ as the lower bound for recovering. Let $\mathbf{Z_{enc}} = [\mathbf{Z}_{major}, \mathbf{Z}_{minor}]$.

[6] Notice that $\mathbf{Z_{enc}}$ are random variables with $q_{\mathbf{enc}}(\mathbf{z}) = \int q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) p_{data}(\mathbf{x}) d\mathbf{x}$. $\hat{\mathbf{X}}(\mathbf{Z_{enc}})$ is a function named $\hat{\mathbf{X}}$ of $\mathbf{Z_{enc}}$.

Let us only use a major set of factors $\mathbf{Z}_{major}$, that is to construct a new estimator $\mathbf{X}_{recc} = \mathbf{dec}_{\mu}(\mathbf{Z}_{major}, \mathbf{0})$ with setting $Z_{minor} = \mathbf{0}$. With the assumption that $q_{\mathbf{enc}}(\mathbf{z})$ can factorize, it yields the separation of mutual information $I(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}}) = I(\mathbf{X}; \mathbf{Z}_{minor}) + I(\mathbf{X}; \mathbf{Z}_{major})$. It yields the following bound,

$$\begin{aligned} & \mathbb{E}(\mathbf{X} - \mathbf{X}_{recc})^2 \\ \geq & \frac{1}{2\pi e} e^{2(H(\mathbf{X}) - I(\mathbf{X}; \mathbf{Z}_{major}))} \\ \geq & \frac{1}{2\pi e} e^{2(H(\mathbf{X}) - I(\mathbf{X}; \mathbf{Z}_{major}))} e^{-2I(\mathbf{X}; \mathbf{Z}_{minor})}. \end{aligned} \tag{11}$$

The theorem implies that the mutual information carried by the selecting factors directly influences on the lower bound of the best reconstruction 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, $\mathbf{Y} \to \mathbf{X} \to \mathbf{Z}_{\mathbf{enc}} \to \mathbf{Y}_{pre}$, holds.[7] According to the Fano's inequality ([36]) 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 $\mathbf{Y} \to \mathbf{Z}_{\mathbf{enc}} \to \hat{\mathbf{Y}}$, with $P_e = Pr(\hat{\mathbf{Y}} \neq \mathbf{Y})$, we have*

$$H(P_e) + P_e log|\mathcal{Y}| \geq H(\mathbf{Y}) - I(\mathbf{Y}; \mathbf{Z}_{\mathbf{enc}}) \geq H(\mathbf{Y}) - I(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}}) \tag{12}$$

where $\mathcal{Y}$ is the alphabet of $\mathbf{Y}$. Since the number of class is no smaller than 2, it naturally holds that $log(|\mathcal{Y}|) > 0$. This inequality can then pbe weakened to

$$1 + P_e log|\mathcal{Y}| \geq H(\mathbf{Y}) - I(\mathbf{Y}; \mathbf{Z}_{\mathbf{enc}}) \geq H(\mathbf{Y}) - I(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}}). \tag{13}$$

or

$$P_e \geq \frac{H(\mathbf{Y}) - I(\mathbf{Y}; \mathbf{Z}_{\mathbf{enc}}) - 1}{log|\mathcal{Y}|} \geq \frac{H(\mathbf{Y}) - I(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}}) - 1}{log|\mathcal{Y}|}. \tag{14}$$

[7] This condition implies $\mathbf{Y} \to \mathbf{X} \to \mathbf{Z}_{\mathbf{enc}}$ which guarantees that $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}, \mathbf{y}) = q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x})$. It also implies $\mathbf{Y} \to \mathbf{Z}_{\mathbf{enc}} \to \mathbf{Y}_{\mathbf{pre}}$ which guarantees that $\mathbf{Y}_{\mathbf{pre}}$ can be then taken as a rational estimator of $\mathbf{Y}$ based on Theorem 5.

Note that according to information processing inequality $I(\mathbf{X}; Z_{\mathbf{enc}h}) \geq I(\mathbf{y}; z_{\mathbf{enc}h})$. $I(\mathbf{X}; Z_{\mathbf{enc}h}) = 0 \Rightarrow I(\mathbf{y}; z_{\mathbf{enc}h}) = 0$, and If $I(\mathbf{y}; z_{\mathbf{enc}h}) = 0$ the h^{th} factor will not influence the prediction. Let we regard $\mathbf{Y}_{pre}$ as $\hat{\mathbf{Y}}$. With the assumption that $q_{\mathbf{enc}}(\mathbf{z})$ can factorize, since $\mathbf{Y} \rightarrow \mathbf{Z}_{\mathbf{enc}} \rightarrow \mathbf{Y}_{pre}$, 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 $I(\mathbf{X}; Z_{\mathbf{enc}h})$ without significantly lifting the lower bound of the prediction error.

3.3. Algorithms to Quantitatively Calculate the Proposed indicators

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

We can then list the indicators to be estimated as:

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

$$I_{est}(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}})_M = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}^m)||q^*(\mathbf{z})). \tag{15}$$

This estimation uses M sample according to the empirical form of Corollary 1.

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

$$I_{est}(\mathbf{X}; Z_{\mathbf{enc}h})_M = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(z_{\mathbf{enc}h}|\mathbf{x}^m)||q^*(z_{\mathbf{enc}h})). \tag{16}$$

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

$$\begin{aligned} &\min_q \mathop{\mathbb{E}}_{\mathbf{x} \sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x})||q(\mathbf{z})) \\ &\Leftrightarrow \min_q D_{KL}(q_{\mathbf{enc}}(\mathbf{z})||q(\mathbf{z}))d\mathbf{z}, \end{aligned} \tag{17}$$

and then we can prove the following result:

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

$$\begin{aligned} & \mathop{\mathbb{E}}_{\mathbf{x}\sim p_{data}(\mathbf{x})} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x})||q^*(\mathbf{z})) \\ = \quad & I(\mathbf{X};\mathbf{Z}_{\mathbf{enc}}) + D_{KL}(q_{\mathbf{enc}}(\mathbf{z})||q^*(\mathbf{z})). \end{aligned} \tag{18}$$

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:

$$q^*(\mathbf{z}) = \arg\min_q \frac{1}{M}\sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}^m)||q(\mathbf{z})). \tag{19}$$

The above minimization problem can be solved with a closed-form solution as follows[8]:

$$\sigma_i^* = \frac{\sum_{m=1}^{M}\sigma_i(\mathbf{x}^m) + \mu_i^2(\mathbf{x}^m)}{M}.$$

The proof is as follows:

Notice that suppose we have two multivariate normal distributions, with means μ_0, μ_1 and with non-singular covariance matrices Σ_0, Σ_1 and the two distributions have the same dimension H, then it yields[37]

$$D_{KL}(\mathcal{N}_0||\mathcal{N}_1) = \frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\mu_1-\mu_0)^T\Sigma_1^{-1}(\mu_1-\mu_0) - H + \ln(\frac{\det\Sigma_1}{\det\Sigma_0})). \tag{20}$$

Note that we assume $q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, diag(\sigma_1, \cdots, \sigma_H))$ and

$q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), diag(\sigma_1(\mathbf{x}), \cdots, \sigma_H(\mathbf{x})))$. Thus, we have

$$\begin{aligned} & \frac{1}{M}\sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}^m)||q(\mathbf{z})) \\ = \quad & \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2}\sum_{i=1}^{H}\frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln\frac{\sigma_i}{\sigma_i(\mathbf{x}^m)} \\ = \quad & \frac{1}{2M}\sum_{i=1}^{H}\sum_{m=1}^{M}\frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln\frac{\sigma_i}{\sigma_i(\mathbf{x}^m)}. \end{aligned}$$

[8]The above minimization problem can also be solved by gradient descent.

The optimization can be divided into H optimization sub-problems as the following,

$$\sigma_i^* = \arg\min_{\sigma_i} \sum_{m=1}^{M} \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)}, i = 1, \cdots, H. \tag{21}$$

$$\nabla_{\sigma_i}(\sum_{m=1}^{M} \frac{\sigma_i(\mathbf{x}^m)}{\sigma_i} + \frac{\mu_i(\mathbf{x}^m)^2}{\sigma_i} - 1 + \ln \frac{\sigma_i}{\sigma_i(\mathbf{x}^m)}) = \sum_{m=1}^{M} -\frac{\sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{\sigma_i^2} + \frac{1}{\sigma_i}. \tag{22}$$

Since

$$\sum_{m=1}^{M} -\frac{\sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{\sigma_i^{*2}} + \frac{1}{\sigma_i^*} = 0, \tag{23}$$

it yields

$$\sigma_i^* = \frac{\sum_{m=1}^{M} \sigma_i(\mathbf{x}^m) + \mu_i(\mathbf{x}^m)^2}{M}. \tag{24}$$

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 $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}(\mathbf{x}), diag(\sigma_1(\mathbf{x}), \cdots, \sigma_H(\mathbf{x})))$
2: **Obtain:** $q^*(\mathbf{z}) = \arg\min_q \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x}^m)||q(\mathbf{z}))$.
3: **for** $i = h$ **to** H **do**
4: $\quad \sigma_i^* = \frac{\sum_{m=1}^{M} \sigma_i(\mathbf{x}^m) + \mu_i^2(\mathbf{x}^m)}{M}$.
5: **end for**
6: **Calculate:** $I_{est}(\mathbf{X}; Z_{\mathbf{enc}h})_M = \frac{1}{M} \sum_{m=1}^{M} D_{KL}(q_{\mathbf{enc}}(z_{\mathbf{enc}h}|\mathbf{x}^m)||q^*(z_{\mathbf{enc}h}))$
7: **for** $i = h$ **to** H **do**
8: $\quad I_{est}(\mathbf{X}; Z_{\mathbf{enc}h})_M = \frac{1}{M} \sum_{m=1}^{M} \log \frac{\sigma_h^*}{\sigma_h(\mathbf{x}^m)} + \frac{\sigma_h^2(\mathbf{x}^m) + \mu_h^2(\mathbf{x}^m)}{2\sigma_h^{*2}}$.
9: **end for**
10: **Calculate:** $I_{est}(\mathbf{X}; Z_{\mathbf{enc}})_M$
11: $I_{est}(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}})_M = \sum_{h=1}^{H} I_{est}(\mathbf{X}; Z_{\mathbf{enc}h})_M$.
12: **Output:** $I_{est}(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}})_M$, $I_{est}(\mathbf{X}; Z_{\mathbf{enc}h})_M$, $q^*(\mathbf{z})$

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

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

$$|I_{est}(\mathbf{X};\mathbf{Z_{enc}})_M - I(\mathbf{X};\mathbf{Z_{enc}})| < \varepsilon. \tag{25}$$

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

$$|I_{est}(\mathbf{X};\mathbf{Z_{enc}})_M - I(\mathbf{X};\mathbf{Z_{enc}})| < \varepsilon. \tag{26}$$

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

$$|I_{est}(\mathbf{X};\mathbf{Z_{enc}})_M - \tilde{I}[q^*]| < \varepsilon/2, \tag{27}$$

$$\begin{aligned} & |I_{est}(\mathbf{X};\mathbf{Z_{enc}})_M - I(\mathbf{X};\mathbf{Z_{enc}})| \\ \leq\ & |I_{est}(\mathbf{X};\mathbf{Z_{enc}})_M - \tilde{I}[q^*]| + |\tilde{I}[q^*] - I(\mathbf{X};\mathbf{Z_{enc}})| \\ <\ & \frac{\varepsilon}{2} + |D_{KL}(q_{\mathbf{enc}}(\mathbf{z})||q^*(\mathbf{z}))| < \varepsilon. \end{aligned} \tag{28}$$

□

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 $q_{\mathbf{enc}}(\mathbf{z})$ and the number of the sample is bigger enough. Besides, the minimization of $D_{KL}(q_{\mathbf{enc}}(\mathbf{z})||q^*(\mathbf{z}))$ in theorem 6 inspires the derivation of $q^*(\mathbf{z})$.

4. 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 ([8], [33]) or aural ([23]) observation. However, it might be labor-intensive to select factors for latter tasks.

In ([38]), $q_{\mathbf{enc}}(\mathbf{z}|\mathbf{x})$ 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 ([26]) 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 indictor 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.

5. Experimental Results

5.1. Datasets

MNIST is a database of handwritten digits ([30]). 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 ([31]) 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 ([32]) and we use 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 A.

5.2. 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

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

of estimated mutual information can be found with little generation effects and factors with large values of mutual information 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 $I(\mathbf{X}; Z_{\mathbf{enc}h}) > 0.5$ of CelebA shown in the Fig.(A.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.

5.3. 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 2 shows the detailed total information and the reconstruction error corresponding to the different ratio of factor. The top 10% factors contain almost the whole

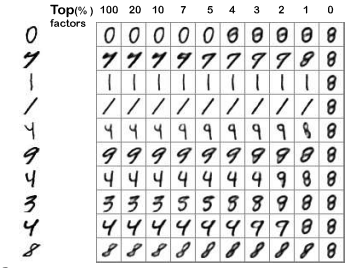


Figure 4: Generation plot with different ratio of factors.

Table 2: mutual information and reconstruction error plot

Top(%) factors in used	100	20	10	7	5	4	3	2	1	0
$I(\mathbf{X}; \mathbf{Z}_{\mathbf{enc}_{used}})$	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

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.

5.4. 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 3, 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).

Table 3: Mutual information and EEG-emotion classification with $\beta(=6)$-VAE

Top(%) factors in used	100	50	10	7	5	4	3	2	1	0
$I(\mathbf{X};\mathbf{Z}_{\mathbf{enc}_{used}})$	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

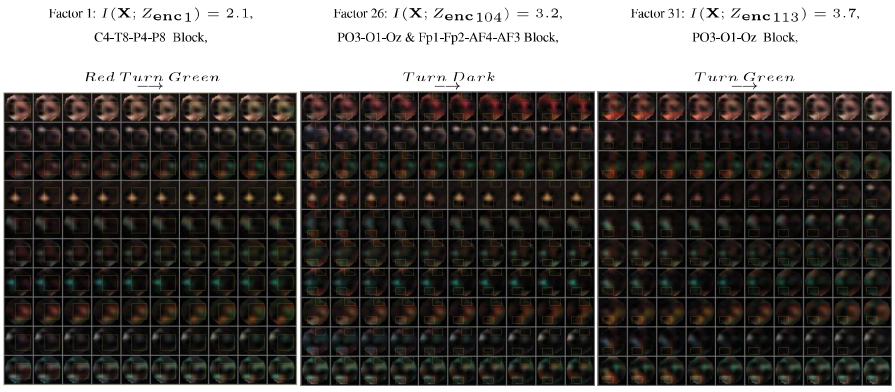


Figure 5: Emotion relevant factors discovery. We present 3 influential factors determined by estimated mutual information. The whole influential factor traversals are listed in Fig.(A.8) in appendix.

6. 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 mutual 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 automatically find vairiants and extract knowledge under the data and it can be applied to various variants of VAE including β-VAE ([33]), FactorVAE ([39]), β-TCVAE ([40]) and DIP-VAE ([41]). It 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.

7. 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

[1] C. M. Bishop, Pattern recognition and machine learning, springer, 2006.

[2] Q. Zhao, D. Meng, Z. Xu, Robust sparse principal component analysis, Science China Information Sciences 57 (9) (2014) 1–14.

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

[4] A. Hyvärinen, J. Karhunen, E. Oja, Independent component analysis, Vol. 46, John Wiley & Sons, 2004.

[5] A. Hyvärinen, P. Pajunen, Nonlinear independent component analysis: Existence and uniqueness results, Neural Networks 12 (3) (1999) 429–439.

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

[7] S. Makeig, A. J. Bell, T.-P. Jung, T. J. Sejnowski, Independent component analysis of electroencephalographic data, in: Advances in neural information processing systems, 1996, pp. 145–151.

[8] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

[9] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of Machine Learning Research 11 (Dec) (2010) 3371–3408.

[10] Y. J. Fan, Autoencoder node saliency: Selecting relevant latent representations, Pattern Recognition 88 (2019) 643–653.

[11] G. Bhatt, P. Jha, B. Raman, Representation learning using step-based deep multi-modal autoencoders, Pattern Recognition.

[12] Z. Zhang, D. Chen, Z. Wang, H. Li, L. Bai, E. R. Hancock, Depth-based subgraph convolutional auto-encoder for network representation learning, Pattern Recognition 90 (2019) 363–376.

[13] R. Salakhutdinov, G. Hinton, Semantic hashing, International Journal of Approximate Reasoning 50 (7) (2009) 969–978.

[14] K. G. Lore, A. Akintayo, S. Sarkar, Llnet: A deep autoencoder approach to natural low-light image enhancement, Pattern Recognition 61 (2017) 650–662.

[15] G. Liu, L. Li, L. Jiao, Y. Dong, X. Li, Stacked fisher autoencoder for sar change detection, Pattern Recognition 96 (2019) 106971.

[16] L. Hou, V. Nguyen, A. B. Kanevsky, D. Samaras, T. M. Kurc, T. Zhao, R. R. Gupta, Y. Gao, W. Chen, D. Foran, et al., Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images, Pattern recognition 86 (2019) 188–200.

[17] D. P. Kingma, M. Welling, Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114.

[18] D. J. Rezende, S. Mohamed, D. Wierstra, Stochastic backpropagation and approximate inference in deep generative models, arXiv preprint arXiv:1401.4082.

[19] M. F. Mathieu, J. J. Zhao, J. Zhao, A. Ramesh, P. Sprechmann, Y. LeCun, Disentangling factors of variation in deep representation using adversarial training, in: Advances in Neural Information Processing Systems, 2016, pp. 5040–5048.

[20] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, O. Winther, Autoencoding beyond pixels using a learned similarity metric, arXiv preprint arXiv:1512.09300.

[21] I. Higgins, L. Matthey, X. Glorot, A. Pal, B. Uria, C. Blundell, S. Mohamed, A. Lerchner, Early visual concept learning with unsupervised deep learning, arXiv preprint arXiv:1606.05579.

[22] F. Locatello, S. Bauer, M. Lucic, S. Gelly, B. Schölkopf, O. Bachem, Challenging common assumptions in the unsupervised learning of disentangled representations, arXiv preprint arXiv:1811.12359.

[23] W.-N. Hsu, Y. Zhang, J. Glass, Unsupervised learning of disentangled and interpretable representations from sequential data, in: Advances in neural information processing systems, 2017, pp. 1876–1887.

[24] E. L. Denton, et al., Unsupervised learning of disentangled representations from video, in: Advances in neural information processing systems, 2017, pp. 4414–4423.

[25] M. Suzuki, K. Nakayama, Y. Matsuo, Joint multimodal learning with deep generative models, arXiv preprint arXiv:1611.01891.

[26] I. Higgins, N. Sonnerat, L. Matthey, A. Pal, C. P. Burgess, M. Botvinick, D. Hassabis, A. Lerchner, Scan: Learning abstract hierarchical compositional visual concepts, arXiv preprint arXiv:1707.03389.

[27] Z. Hu, Z. Yang, X. Liang, R. Salakhutdinov, E. P. Xing, Toward controlled generation of text, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1587–1596.

[28] I. Higgins, A. Pal, A. A. Rusu, L. Matthey, C. P. Burgess, A. Pritzel, M. Botvinick, C. Blundell, A. Lerchner, Darla: Improving zero-shot transfer in reinforcement learning, arXiv preprint arXiv:1707.08475.

[29] A. van den Oord, O. Vinyals, et al., Neural discrete representation learning, in: Advances in Neural Information Processing Systems, 2017, pp. 6309–6318.

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

[31] Z. Liu, P. Luo, X. Wang, X. Tang, Deep learning face attributes in the wild, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 3730–3738.

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

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

[34] H. Peng, F. Long, C. Ding, Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy, IEEE Transactions on pattern analysis and machine intelligence 27 (8) (2005) 1226–1238.

[35] M. D. Hoffman, M. J. Johnson, Elbo surgery: yet another way to carve up the variational evidence lower bound, in: Workshop in Advances in Approximate Bayesian Inference, NIPS, 2016.

[36] T. M. Cover, J. A. Thomas, Elements of information theory, John Wiley & Sons, 2012.

[37] J. Duchi, Derivations for linear algebra and optimization, Berkeley, California 3.

[38] A. A. Alemi, I. Fischer, J. V. Dillon, K. Murphy, Deep variational information bottleneck, arXiv preprint arXiv:1612.00410.

[39] H. Kim, A. Mnih, Disentangling by factorising, in: International Conference on Machine Learning, 2018, pp. 2654–2663.

[40] T. Q. Chen, X. Li, R. B. Grosse, D. K. Duvenaud, Isolating sources of disentanglement in variational autoencoders, in: Advances in Neural Information Processing Systems, 2018, pp. 2610–2620.

[41] A. Kumar, P. Sattigeri, A. Balakrishnan, Variational inference of disentangled latent concepts from unlabeled observations, arXiv preprint arXiv:1711.00848.

[42] P. Bashivan, I. Rish, M. Yeasin, N. Codella, Learning representations from eeg with deep recurrent-convolutional neural networks, arXiv preprint arXiv:1511.06448.

Appendix A. Experiment Details

Appendix A.1. 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 β-VAE, VAE). Each row is generated with a different seed image.

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

Appendix A.2. 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 β-VAE, VAE). Each row is generated with a different seed image.

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

Appendix A.3. DEAP

DEAP is a well-known public multi-modalities (e.g. EEG, video, etc.) dataset proposed by [32]. The EEG signals are recorded from 32 channels by 32 participants watching 40 videos for 63 seconds each. The EEG data was preprocessed which down-sampling into 128Hz and band range 4-45 Hz. By the same transformation idea from [42], 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 [42] 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. β(= 6)-VAE is trained on each frame and LSTM was used to combine all the frames together for each video.

The estimated mutual information is 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 β-VAE). Each row is generated with a different seed image.

Appendix A.4. Network Structure

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

Appendix A.5. Experiment Plot

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

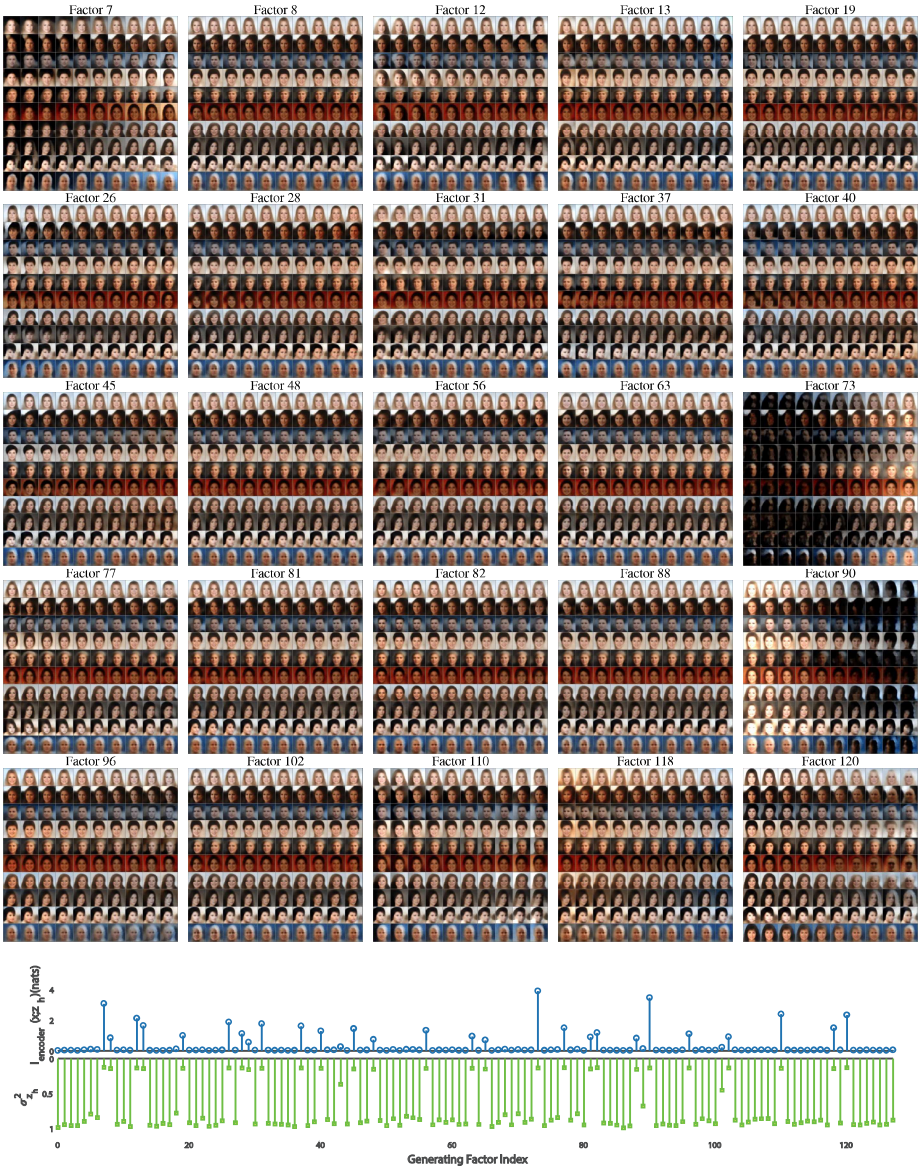


Figure A.6: **Mutual Information Sparsity in CelebA**: Generating Factors Traversal of β(=40)-VAE

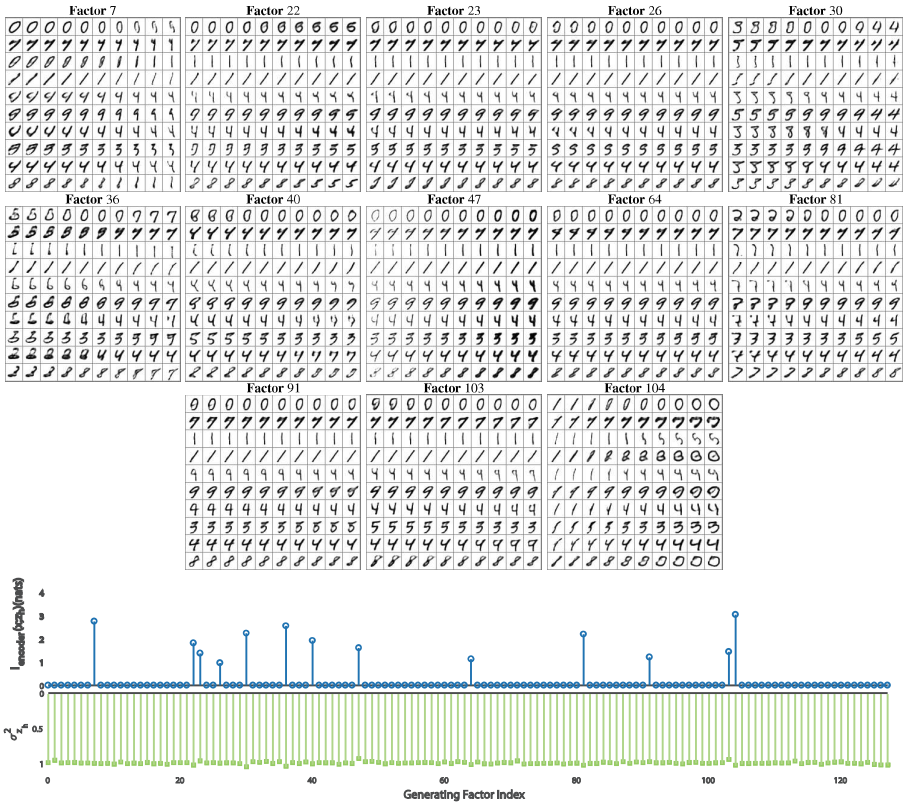


Figure A.7: **Mutual Information Sparsity in MNIST**: Generating Factor Traversal of β(=10)-VAE

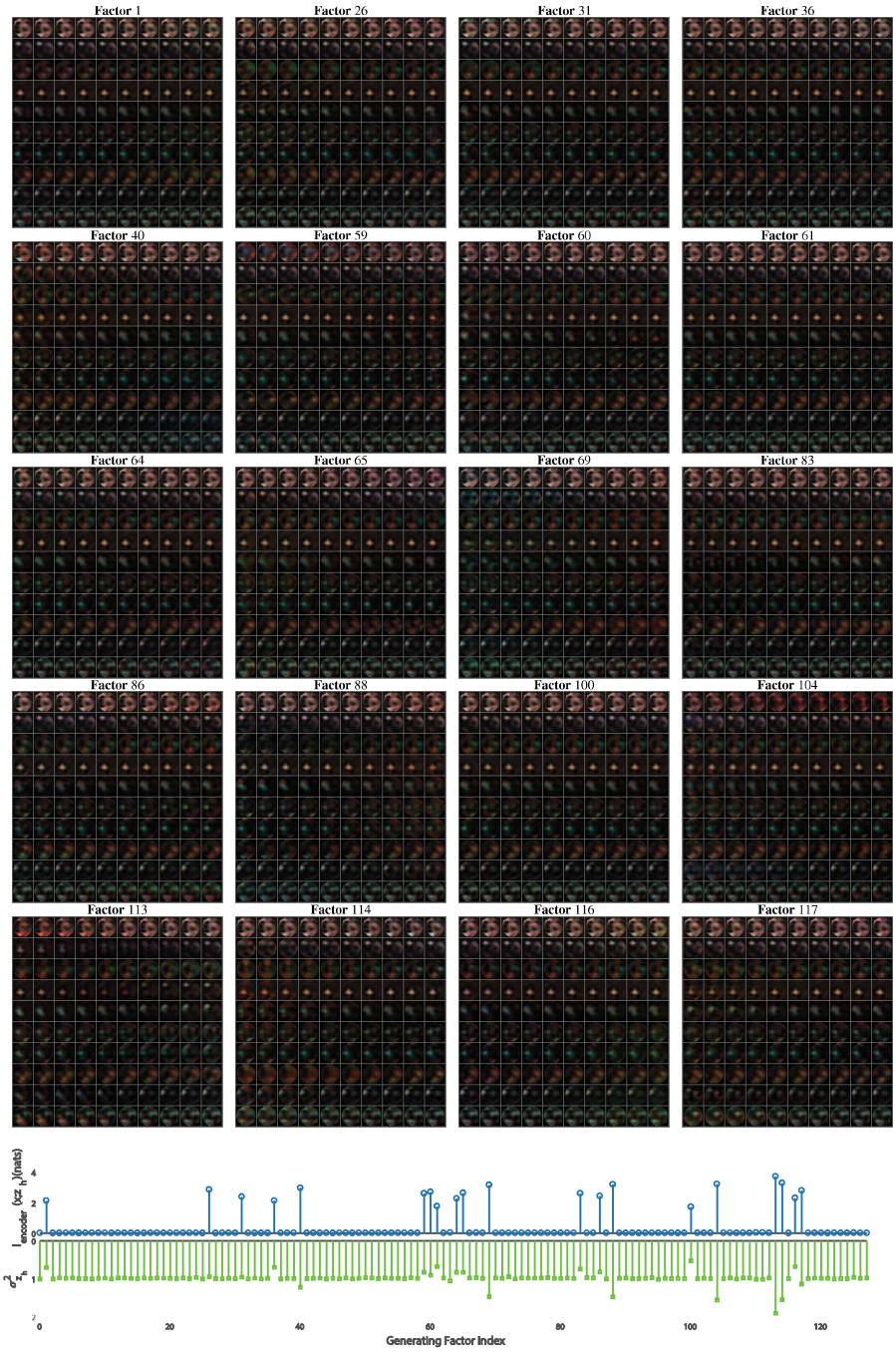


Figure A.8: **Mutual Information Sparsity in DEAP**: Generating Factor Traversal of β(=6)-VAE

Biography

Shiqi Liu is currently pursuing master degree in Applied Mathematics from Xi'an Jiaotong University. His research interests include self-paced learning, generative model, brain signal processing.

Jingxin Liu received Doctor of Engineering from Brunel University London. His research interests include brain signal processing, machine learning.

Qian Zhao received Ph.D. degree in Applied Mathematics from Xi'an Jiaotong University. He is currently a lecturer of Xi'an Jiaotong University. His research interests include low-rank matrix/tensor analysis, bayesian modeling, self-paced learning.

Xiangyong Cao is currently pursuing Ph.D. degree in Applied Mathematics from Xi'an Jiaotong University. His research interests include Bayesian method, Social network.

Huibin Li received master degree in Applied Mathematics from Xi'an Jiaotong University in 2009. He received Ph.D. in Department of Mathematics and Informatics from Ecole Centrale de Lyon. He is currently an assistant professor in School of Mathematics and Stasitics, Department of Informatics from Xi'an Jiaotong University. His research interests include 3D face recognition and 3D facial expression recognition; 3D non-rigid surface deformation modeling, registration and maching; Geometry based 3D Shape Analysis and Processing.

Deyu Meng received master degree in Applied Mathematics from Xi'an Jiaotong University in 2004. He received Ph.D. degree in Computer Science from Xi'an Jiaotong University in 2008. He is currently a professor of Xi'an Jiaotong University. His research interests include self-paced learning, noise modelling.

Hongying Meng He received his Ph.D. degree in Communication and Electronic Systems from Xian Jiaotong University and was a lecturer in Electronic Engineering Department of Tsinghua University in China. Dr Hongying Meng is currently a Senior Lecturer at the Department of Electronic and Computer Engineering, College of Engineering, Design and Physical Sciences, Brunel University London where he joined in as a lecturer in 2011. His research interests include digital signal processing, machine learning, human computer interaction, computer vision, embedded systems and communications.

Sheng Liu is currently pursuing Ph.D. degree in Computer Science from State University of New York at Buffalo. His research interests include video description, video event detection and video processing.