Fisher Information-based Metrics for Representation Learning

Do Nguyen Dang Thi[†], Le Quoc Anh[†], Tran Trong Duy[†‡], Le Vu Ha[†], Nguyen Linh Trung[†,*]
[†] VNU University of Engineering and Technology, Hanoi, Vietnam
[‡] CentraleSupélec, Université Paris-Saclay, CNRS, L2S, Gif-sur-Yvette, France

***Abstract*—Generative models have recently made remarkable progress in enhancing the disentanglement and interpretability of their latent representations. While mutual information-based (MI-based) metrics have been widely used to evaluate disentanglement, they have several limitations, such as high computational complexity, inability to deal with nonlinear relations, and sensitivity to noise. To overcome these shortcomings, we propose a novel approach that leverages Fisher Information (FI) to evaluate two critical aspects of disentanglement: explicitness and modularity. The proposed metrics have a clear intuition based on the information inequality from estimation theory. It is demonstrated via numerical experiments that the FI-based metrics provide more stable and interpretable evaluations under noise and nonlinearity, with lower computational overhead than traditional MI-based metrics.**

I. Introduction

In recent years, generative models have become a crucial tool in machine learning, with the ability to learn and present probability distributions of data, thereby generating new data similar to the original [1], [2]. A key component of these models is the latent representation, which is designed to be compact and structural. A major challenge lies in making these representations interpretable, meaning that each dimension in the latent space corresponds to a specific data attribute [3], [4]. To achieve this, the latent representation must not only capture essential features of the data but also be disentangled, i.e., different dimensions in the latent space should be independent and unaffected by one another [3]. For example, in a facial-image dataset, each latent dimension should map to a concrete attribute, such as eye color, hairstyle, or expression, instead of encoding a mixed blend of unclear characteristics [4]. If a model can learn this, one can easily adjust a single dimension to change the desired attribute without affecting others, thereby providing greater flexibility in generating new samples [5].

To address this challenge, several deep learning models have been proposed to increase the disentanglement of the latent representation, making its dimensions more independent and interpretable [6]. Notably, β-VAE [7], an extension of the Variational AutoEncoder (VAE), encourages the independence of latent dimensions by adjusting a hyperparameter β in the loss function that controls the trade-off between encoding capacity and disentanglement. InfoGAN [8] extends the Generative Adversarial Network (GAN) by adding a classifier to maximize MI between the latent dimensions and the data factors. InfoGAN is capable of learning disentangled representations without complex architectures. However, these methods only work for relatively simple imagery data, such as faces or single 3D objects [6].

The evaluation of the interpretability of latent representations is crucial, and many evaluation metrics have been proposed to measure properties in the disentanglement such as explicitness, modularity, compactness, and the information content of the latent space [6], [9]. These metrics can be classified into three categories: information-based, predictor-based, and intervention-based [9]. Information-based metrics estimate the MI between the factors and the codes to compute a disentanglement score. Predictor-based metrics are measured by using regressors or classifiers to predict factors from codes. While intervention-based metrics evaluate disentanglement by comparing codes within subsets of data with fixed factors.

Information-based metrics are particularly emphasized in this work due to their principled foundation in information theory and their widespread adoption in the literature. Furthermore, information-based metrics offer several advantages, such as requiring fewer hyperparameters and making no assumptions about the relationships between the factors and the encoding [10]. One of the most commonly used metrics is MIG (Mutual Information Gap), which utilizes MI to measure the degree of disentanglement between the dimensions in the latent space, ensuring that each dimension corresponds to a specific data attribute without mixing with other attributes [11]. Additionally, DCI (Disentanglement, Completeness, and Informativeness) is a more comprehensive set of metrics, evaluating not only disentanglement but also the ability to fully represent the important attributes of the data in the latent space [12].

However, these metrics have considerable limitations. First, none of them specifically measure the explicitness of the representation. Second, current metrics often calculate modularity based on the difference between the two largest MI values between factors and codes. Here, factors refer to the underlying attributes or properties of the data (e.g., digit identity, thickness, rotation angle), and codes are the dimensions of the learned latent representation. Additionally, most current implementations use maximum likelihood-based MI estimation methods, which typically partition the factor and code spaces into small bins, calculating empirical probabilities for each bin and then applying discrete MI formulas. Consequently, these metrics are inherently sensitive to bin granularity. Moreover, many metrics require repeated MI estimations between factor-

*This research was supported by project QG.25.08 at Vietnam National University, Hanoi. Correspondence: Nguyen Linh Trung (linhtrung@vnu.edu.vn).

979-8-3315-7206-8/25/$31.00

code pairs, becoming problematic when dealing with high-dimensional latent spaces. This prohibits the application of more precise but computationally intensive MI estimation methods, such as MINE [13].

The relationship between FI and MI is complex [14], [15], with FI providing both lower and upper bounds for MI. In deep learning, the Fisher Information Matrix (FIM) has been utilized for optimization and spectral analysis [16]–[18]. However, in statistics and signal processing, FI is best known for its direct connection to the Information Inequality or the Cramér-Rao Bound (CRB), which is a well-known reference metric.

We propose the hypothesis that FIM, as a measure of how sensitive the model output is to changes in latent variables, can be used as a reliable tool to measure two important aspects of disentanglement: explicitness and modularity. Our proposed metrics – NEXF and MODF – retain the advantages of information-based metrics while addressing their limitations. We conduct several experiments demonstrating that the proposed metrics can better reflect the explicitness and modularity in noisy and nonlinear circumstances. Moreover, their computational complexities are significantly reduced, as compared to other information-based metrics, thanks to the condensed constructions of the proposed metrics.

The remainder of this paper is organized as follows. Section II provides theoretical background on information-theoretic concepts used to evaluate disentanglement, including MI and FI. Section III presents our proposed FI-based metrics for evaluating explicitness and modularity. Section IV describes the experimental setup and compares our metrics against existing disentanglement measures under various noise levels and nonlinearity conditions. Section V concludes the paper and outlines directions for future work.

II. Background

A. Mutual Information-based metrics

MI is a fundamental concept in information theory that quantifies the amount of information one random variable contains about another. Formally, the MI between two random variables X and Z is defined as

$$I(X,Z) = \int\int p(x,z)\log\left(\frac{p(x,z)}{p(x)p(z)}\right)dxdz. \tag{1}$$

A notable MI-based metric is the Mutual Information Gap (MIG), which measures the difference between the highest and second-highest MI values for each ground-truth factor. It is formally defined as

$$\mathrm{MIG}_i = \frac{I(v_i,z_\star) - I(v_i,z_o)}{H(v_i)}. \tag{2}$$

where $I(v_i,z_\star)$ is the MI between the ground-truth factor v_i and the latent variable $z_\star$ with the highest MI, $I(v_i,z_o)$ is the second-highest, and $H(v_i)$ denotes the entropy of the factor v_i. The final MIG score is computed by averaging over all generative factors. A higher MIG indicates that each factor is primarily captured by a distinct latent dimension, implying a more disentangled representation.

Other metrics in this category include Modularity and DCI, which also leverage MI or statistical independence to assess the alignment between latent dimensions and the underlying generative factors.

B. Fisher Information and the Cramér-Rao Bound

While MI captures statistical dependency between variables, FI quantifies the amount of information an observed random variable carries about an unknown parameter. Given a parametric distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{x}$ is a random variable and $\boldsymbol{\theta} \in \mathbb{R}^d$ is a vector of unknown parameters where d represents the dimension of the parameter $\boldsymbol{\theta}$. The FIM is defined as:

$$\boldsymbol{F}(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{x}|\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}\log p(\boldsymbol{x}|\boldsymbol{\theta})^\top\right]. \tag{3}$$

The FIM is always symmetric and positive definite. It provides insights into how sensitive the likelihood function, is to changes in model parameters, and therefore how much information the data carries about those parameters. FI is closely related to the CRB [19], which gives a lower bound on the variance of any unbiased estimator $\hat{\boldsymbol{\theta}}$. The CRB is defined as the inverse of the FI:

$$\mathrm{var}(\hat{\boldsymbol{\theta}}) \succeq CRB(\boldsymbol{\theta}) = \boldsymbol{F}^{-1}(\boldsymbol{\theta}), \tag{4}$$

where $\mathrm{var}(\hat{\boldsymbol{\theta}})$ denotes the covariance matrix of the estimator $\hat{\boldsymbol{\theta}}$, and $\succeq$ denotes the generalized inequality.

This result implies that higher FI leads to a lower value of bounds on variances of estimators, i.e., it enables more precise parameter estimation.

III. Main work

To address the limitations of existing information-based metrics, we propose two novel metrics that directly quantify the interpretability properties of latent representations using FI. These metrics are designed to be both theoretically grounded and computationally efficient, enabling reliable evaluation even in high-dimensional latent spaces. The first metric focuses on explicitness, measuring how well the latent codes describe the underlying generative factors. The second metric captures modularity, quantifying the degree of independence between factor representations.

A. Explicitness Metric

Explicitness is defined as the property where factors of interest must be fully described by the code space. Perfect explicitness occurs when a generalizable relationship exists between factors and codes. Typically, desirable relationships are linear or monotonic for continuous factors and partitioned for categorical factors [20].

Considering factors and codes vectors $\boldsymbol{v}, \boldsymbol{z} \in \mathbb{R}^d$, by the CRB theorem, one has

$$\mathbb{E}\left[(\hat{\boldsymbol{v}}-\boldsymbol{v})^\top(\hat{\boldsymbol{v}}-\boldsymbol{v})\right] \succeq \boldsymbol{F}^{-1}, \tag{5}$$

where $\hat{\boldsymbol{v}}$ is an arbitrary estimator of $\boldsymbol{v}$ based on $\boldsymbol{z}$, and we use $\boldsymbol{F} = \boldsymbol{F}(\boldsymbol{v})$ as an abuse of notation. For a specific factor v_i:

$$\mathbb{E}\left[(\hat{v}_i - v_i)^2\right] \succeq F_{ii}^{-1}. \tag{6}$$

Equation (6) directly follows from the diagonal entries of the matrix inequality in Equation (5), which corresponds to the scalar form of the CRB [19], [21]. The higher value of FI indicates a better estimation of factors from codes, which implies greater explicitness. We propose an explicitness metric called EXF, defined as the trace of the FIM:

$$\mathrm{EXF}(\boldsymbol{v}) = \mathrm{Tr}(\boldsymbol{F}) = \sum_i F_{ii}. \tag{7}$$

EXF values are positive and unbounded, with zero indicating that the codes do not bring information about the factors. Given variances of the factors, the normalized EXF is defined as:

$$\mathrm{NEXF}(\boldsymbol{v}) = 1 - \frac{1}{d}\sum_i \frac{\mathrm{var}(v_i)}{F_{ii}}, \tag{8}$$

ranging from zero to one, where one indicates perfect explicitness.

Compared to metrics like JEMMIG [22] and DCIMIG [23], EXF does not require pairwise (mutual) information estimation between factors and codes, significantly reducing complexity and enabling more accurate information estimation methods.

B. Modularity Metric

Modularity or independence between factors is another critical property. Good modularity implies each factor affects only a distinct code subspace and vice versa [20]. To evaluate modularity, we utilize the off-diagonal entries of the FIM, which reflect asymptotic covariance estimation errors between parameters in maximum likelihood estimation. These entries relate directly to partial correlations of estimation errors [24]. Intuitively, if the factors are independent, their estimation errors should not be correlated. We propose the modularity metric MODF as

$$\mathrm{MODF}(v_i) = 1 - \frac{1}{d-1}\sum_{j\neq i} \frac{|F_{ij}|}{\sqrt{F_{ii}F_{jj}}}, \tag{9}$$

$$\mathrm{MODF}(\boldsymbol{v}) = \frac{1}{d}\sum_i MODF(v_i). \tag{10}$$

Additionally, a normalized modularity metric SMODF based on maximum partial correlations is defined as

$$\mathrm{SMODF}(v_i) = 1 - \max_j \frac{|F_{ij}|}{\sqrt{F_{ii}F_{jj}}}, \tag{11}$$

$$\mathrm{SMODF}(\boldsymbol{v}) = \frac{1}{d}\sum_i SMODF(v_i). \tag{12}$$

MODF and SMODF indicate high modularity when their values are close to one. Unlike existing metrics like MIG-sup [25] and modularity score [26], which measure relative modularity through MI differences, MODF and SMODF directly measure absolute modularity, offering lower computational complexity.

IV. Experiment and Results

We conduct four experiments to evaluate the proposed metrics. Following the experimental design in [9], the first three experiments evaluate the proposed metrics under varying levels of explicitness, modularity, and nonlinearity.

In these experiments, the latent code $\boldsymbol{z} \in \mathbb{R}^8$ is constructed as a controlled function of a meaningful-factor vector $\boldsymbol{v} \in \mathbb{R}^8$ and an independent noise vector $\mathbf{n} \in \mathbb{R}^8$, with a scalar parameter $\alpha \in [0, 1]$ regulating the relative contributions of signal and noise. To estimate the FIM for each experiment, we employ FINE [27], [28], a neural FI estimator, in which the statistics network is a multilayer perceptron with a single hidden layer of sixteen ReLU units. Training is carried out using the Adam optimizer with a learning rate of 10^{-3}, a batch size of 100, and 2000 epochs per experiment.

The fourth experiment compares NEXF scores across different latent configurations in two generative models: GAN and InfoGAN.

In the remainder of this section, we conduct experiments to evaluate the effectiveness of the proposed metrics, comparing them to existing metrics for the same tasks.

A. Explicitness Evaluation Experiment

In this experiment, we examine how evaluation metrics change as representations transition from fully disentangled to completely random due to noise, meaning explicitness decreases from perfection. To simulate this transition, we define the relationship between meaningful factors and latent variables as

$$\boldsymbol{z} = f(\boldsymbol{v}) = (1-\alpha)\boldsymbol{v} + \alpha\mathbf{n}, \tag{13}$$

with $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$, $\boldsymbol{v} \sim \mathcal{N}(0, \sigma_v^2\mathbf{I})$ and α varied from 0.0 to 1.0 in steps of 0.2,

For the model defined in (13), the analytical FIM can be written as

$$\boldsymbol{F} = \frac{1}{\sigma_v^2} + \frac{1}{\sigma_n^2}\left(\frac{1-\alpha}{\alpha}\right)^2. \tag{14}$$

A detailed closed-form expression of this result is provided in Appendix .

As shown in Fig. 1, score of the proposed metrics according to the variation of the parameter α. When $\alpha = 0$, NEXF reaches its highest value of 1, indicating perfect explicitness when there is no noise in the representation. NEXF starts decreasing as α increases, and when $\alpha = 1$, NEXF is nearly 0, indicating the representation completely loses explicitness when the latent variable is represented only by noise.

The other MI-based metrics, such as JEMMIG and DCIMIG, which are used to measure explicitness, have also been illustrated when α increases from 0 to 1. When α increases, the score of both metrics decreases rapidly. It can be observed that both metrics are sensitive to noise and do not maintain good explicitness in high-noise conditions. Furthermore, when $\alpha = 1$, meaning that z is represented only by noise, JEMMIG still reaches an explicitness value of

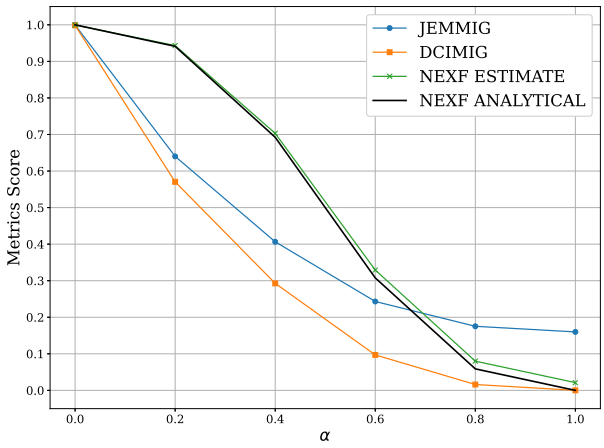

Fig. 1. Metric scores for perfectly disentangled representations under increasing noise influence

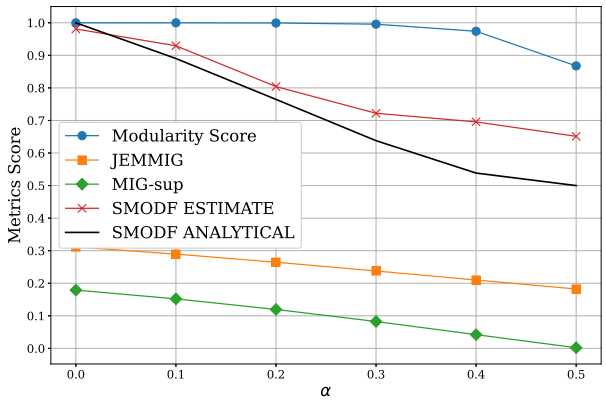

Fig. 2. Metric scores for perfectly disentangled representations under increasing coupling, reflecting modularity degradation.

approximately 0.2, although theoretically, there should be zero explicitness in the latent variable z.

In contrast, the proposed NEXF metric demonstrates a more stable ability to measure explicitness. The explicitness measured by NEXF decreases almost proportionally as noise amplitude increases. This indicates that NEXF is not only highly precise but also exhibits consistency across various noise conditions, unlike other metrics such as JEMMIG and DCIMIG.

B. Modularity Evaluation Experiment

We investigate the variation of evaluation metrics as the modularity of the representation decreases while maintaining the explicitness at its maximum. We define the relationship between meaningful factors and latent variables as

$$\boldsymbol{z} = f(\boldsymbol{v}) = \mathbf{R}\boldsymbol{v} + \mathbf{n}, \tag{15}$$

where: $\mathbf{R} = \begin{bmatrix} 1-\alpha & \alpha & 0 & \cdots & 0 \\ 0 & 1-\alpha & \alpha & \cdots & 0 \\ 0 & 0 & 1-\alpha & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha & 0 & 0 & \cdots & 1-\alpha \end{bmatrix}$,

$\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$, $\boldsymbol{v} \sim \mathcal{N}(0, \sigma_v^2\mathbf{I})$ and α are varied from 0.0 to 0.5 in steps of 0.1.

Thus, the FIM for latent variable $\boldsymbol{z}$ is given by:

$$\mathbf{F} = \frac{1}{\sigma_v^2}\mathbf{I} + \begin{cases} \frac{1}{\sigma_n^2}\left[(1-\alpha)^2 + \alpha^2\right] & \text{if } i = j, \\ \frac{1}{\sigma_n^2}(1-\alpha)\alpha & \text{if } j = i \pm 1, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Here, α is an adjustable parameter controlling interactions between entries in matrix $\mathbf{R}$. When $\alpha = 0$, $\mathbf{R}$ becomes an identity matrix, resulting in the highest modularity of the latent representation $\boldsymbol{z}$ with respect to $\boldsymbol{v}$. As α increases, $\mathbf{R}$ gradually shifts from an identity matrix toward a nearly symmetric matrix, where each representation dimension is associated with two factors, thus reducing modularity.

The analytical FIM results, computed according to our proposed metric definition and comparing to mutual-information-based metrics for varying α are shown in Fig. 2. At $\alpha = 0$, both the Modularity Score and SMODF attain 1, whereas the MI–based metrics start at lower baselines (JEMMIG $\approx$ 0.3, MIG-sup $\approx$ 0.2). As α increases to 0.3, the Modularity Score remains essentially unchanged ($\approx$1), failing to reflect the reduced independence between dimensions. By contrast, JEMMIG and MIG-sup decline from their initial, already low values, exhibiting limited dynamic range.

Our proposed SMODF metric decreases approximately linearly from 1 to 0.5 as α varies from 0 to 0.5, reflecting modularity changes. It is important to note that SMODF can reach zero only under perfect linear correlation among estimated errors; in our setup, balancing the FI at $\alpha = 0.5$ yields a minimal modularity of 0.5. These results demonstrate that SMODF provides both sensitivity and a full range in detecting modularity degradation.

C. Non-linear relations

We evaluate two proposed metrics with nonlinear relations between factor v and latent z. The relationship function between v and z is defined as

$$z = f(v) = 1000^{-\alpha} + 0.25\tan\left(\omega(v - 0.5)\right) + 0.5 + \mathbf{n}, \tag{17}$$

where

$$\omega = 2\arctan\left(\frac{1000^{\alpha} - 0.25}{2}\right), \tag{18}$$

with $\mathbf{n} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})$, $\boldsymbol{v} \sim \mathcal{U}(0, 0.1)^d$ and α varied from 0.0 to 1.0 in steps of 0.2.

Unlike the previous explicitness and modularity experiments, the nonlinear case cannot be computed in closed-form for the FIM. Therefore, only empirical estimation via FINE is used in this setting.

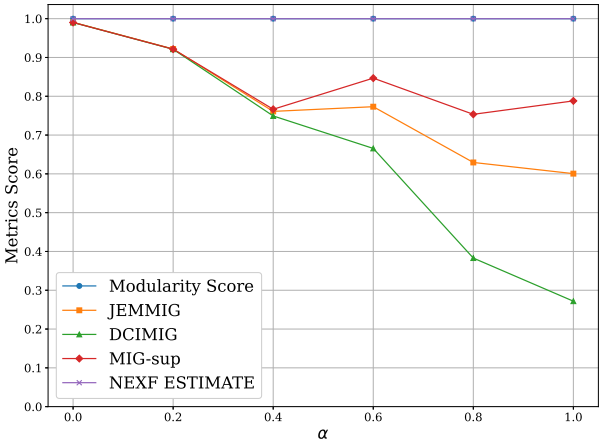

Fig. 3. Metric scores for perfectly disentangled representations within increasingly nonlinear correlation

Fig. 3 illustrates the comparison between various disentanglement metrics as the level of nonlinearity increases, controlled by the parameter α. As observed from the plot, when α increases—indicating that the relationship between factors and codes becomes more nonlinear—most existing metrics such as JEMMIG, DCIMIG, and MIG-sup exhibit a significant decline in their scores. In particular, DCIMIG shows the sharpest drop, decreasing from approximately 0.9 to below 0.3 when α reaches 1.0, suggesting that this metric fails to accurately reflect the modular structure under nonlinear conditions.

Similarly, JEMMIG and MIG-sup also experience a decline in score as α increases, although their degradation is less severe than that of DCIMIG. In contrast, NEXF estimates clearly exhibit their robustness by maintaining high scores, close to 1, even as α increases. This result highlights its ability to accurately assess disentanglement under complex and nonlinear conditions.

D. EXF Metric on GAN vs. InfoGAN

In this experiment, we use NEXF to evaluate the explicitness between the latent variable and factors in both GAN and InfoGAN models. We use the MNIST dataset as the target data distribution. This dataset contains grayscale images of handwritten digits (0-9), each with a resolution of 28x28. For InfoGAN, we evaluate NEXF in four cases: (i) With all latent variables, including 62 noise dimensions ($\mathbf{z}$), 10 discrete dimensions ($\mathbf{c}_{\text{dis}}$), and 2 continuous dimensions ($\mathbf{c}_{\text{cont}}$); (ii) with $\mathbf{c}_{\text{dis}}$ and $\mathbf{c}_{\text{cont}}$; (iii) with only $\mathbf{c}_{\text{cont}}$ and (iv) with $\mathbf{z}$ alone. For the baseline GAN, NEXF is computed over the entire 62-dimensional latent vector $\mathbf{z} \sim \mathcal{N}(0, I)$.

To evaluate NEXF for each case, we define factors as the class labels of the MNIST dataset predicted by a classifier. Specifically, the latent variables are passed through the generator to generate images, which are then input into the classifier to obtain the predicted class labels. A small Gaussian noise is added to the class labels to ensure the regularity condition of the FI. Importantly, this small amount of noise does not impact the accuracy of the information estimated by FINE.

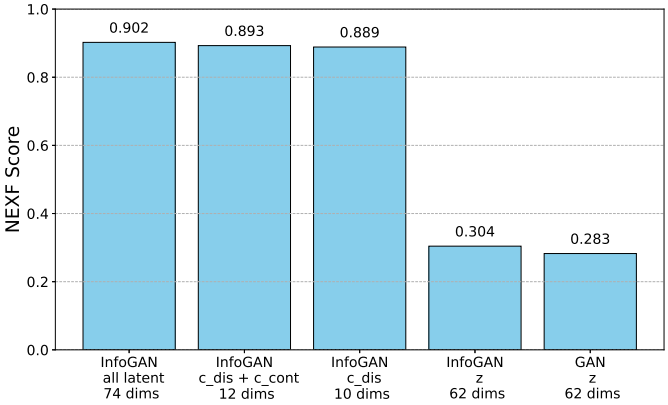

Fig. 4. NEXF scores of InfoGAN and GAN under different latent configurations.

The estimated results for the five cases are shown in Fig. 4. The NEXF score for all latent dimensions (i) in the InfoGAN model is 0.902, which is much higher than the score 0.283 in the GAN model. This shows that the latent variables in the InfoGAN model are more explicit and more interpretable compared to those in the GAN model, thanks to the training procedure that maximizes mutual information (MI).

We further analyze NEXF in different latent subspaces for InfoGAN. The values of NEXF drop slightly to 0.893 and 0.889 in case (ii) and case (iii), respectively. This result shows that discrete latent variables ($\mathbf{c}_{\text{dis}}$) are the main source of explicitness. In contrast, the NEXF in case (iv) drops significantly to 0.304, close to that of GAN. Hence, $\mathbf{z}$ alone could not explain the factors well; however, it implies that $\mathbf{z}$ and the latent variables $\mathbf{c}$ contain some overlapping information regarding the factors.

V. Conclusion

In this work, we have introduced two FI–based metrics, EXF/NEXF and MODF/SMODF, to quantify two key aspects of disentanglement, explicitness and modularity, respectively. By comparing our proposed metrics against MI–based metrics (MIG, DCI, DCIMIG, MIG-sup), we have demonstrated that the proposed metrics have lower computational complexity. Moreover, the metrics can capture the properties perfectly in the case of a nonlinear relationship between factors and codes. Additionally, NEXF corroborates the findings of InfoGAN by assigning much higher scores to InfoGAN compared to GAN, highlighting the effectiveness of maximizing MI in learning more explicit and interpretable representations. In future work, we will focus on further evaluations of the proposed metrics on real-world datasets and integrating them into the training objectives of generative models to jointly optimize generative quality and interpretability.

References

[1] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1.

[2] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.
[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013. DOI: 10.1109/TPAMI.2013.50.
[4] P. Soulos and L. Isik, "Disentangled deep generative models reveal coding principles of the human face processing network," *PLOS Computational Biology*, vol. 20, no. 2, e1011887, 2024.
[5] A. Pandey, M. Fanuel, J. Schreurs, and J. A. Suykens, "Disentangled representation learning and generation with manifold optimization," *Neural Computation*, vol. 34, no. 10, pp. 2009–2036, 2022.
[6] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.
[7] I. Higgins, L. Matthey, A. Pal, *et al.*, "Beta-vae: Learning basic visual concepts with a constrained variational framework.," *ICLR (Poster)*, vol. 3, 2017.
[8] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," *Advances in neural information processing systems*, vol. 29, 2016.
[9] M.-A. Carbonneau, J. Zaidi, J. Boilard, and G. Gagnon, "Measuring disentanglement: A review of metrics," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 7, pp. 8747–8761, 2022.
[10] L. Paninski, "Estimation of entropy and mutual information," *Neural computation*, vol. 15, no. 6, pp. 1191–1253, 2003.
[11] R. T. Q. Chen, X. Li, R. Grosse, and D. Duvenaud, *Isolating sources of disentanglement in variational autoencoders*, 2019. arXiv: 1802.04942 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1802.04942.
[12] C. Eastwood and C. K. I. Williams, "A framework for the quantitative evaluation of disentangled representations," in *International Conference on Learning Representations (ICLR)*, 2018.
[13] M. I. Belghazi, A. Baratin, S. Rajeshwar, *et al.*, "Mutual information neural estimation," in *International conference on machine learning*, PMLR, 2018, pp. 531–540.
[14] X.-X. Wei and A. A. Stocker, "Mutual information, fisher information, and efficient coding," *Neural Computation*, vol. 28, pp. 305–326, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:6154804.
[15] W. Górecki, X. Lu, C. Macchiavello, and L. Maccone, "Mutual Information Bounded by Fisher Information," Mar. 2024. arXiv: 2403.10248 [quant-ph].
[16] R. Pascanu, "Revisiting natural gradient for deep networks," *arXiv preprint arXiv:1301.3584*, 2013.
[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.
[18] R. Karakida, S. Akaho, and S.-i. Amari, "Universal statistics of fisher information in deep neural networks: Mean field approach," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2020, 2018.
[19] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.
[20] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," *Advances in neural information processing systems*, vol. 31, 2018.
[21] T. M. Cover, *Elements of information theory*. John Wiley & Sons, 1999.
[22] K. Do and T. Tran, "Theory and evaluation metrics for learning disentangled representations," *arXiv preprint arXiv:1908.09961*, 2019.
[23] A. Sepliarskaia, J. Kiseleva, M. de Rijke, *et al.*, "Evaluating disentangled representations," *arXiv preprint arXiv:1910.05587*, 2019.
[24] K. Baba, R. Shibata, and M. Sibuya, "Partial correlation and conditional correlation as measures of conditional independence," *Australian & New Zealand Journal of Statistics*, vol. 46, no. 4, pp. 657–664, 2004. DOI: https://doi.org/10.1111/j.1467-842X.2004.00360.x.
[25] Z. Li, J. V. Murkute, P. K. Gyawali, and L. Wang, "Progressive learning and disentanglement of hierarchical representations," *arXiv preprint arXiv:2002.10549*, 2020.
[26] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Jun. 2004, ISSN: 1550-2376. DOI: 10.1103/physreve.69.066133. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.69.066133.
[27] T. T. Duy, L. V. Nguyen, V.-D. Nguyen, N. L. Trung, and K. Abed-Meraim, "Fisher information neural estimation," in *2022 30th European Signal Processing Conference (EUSIPCO)*, IEEE, 2022, pp. 2111–2115.
[28] T. T. Duy, N. Van Ly, N. L. Trung, and K. Abed-Meraim, "Fisher information estimation using neural networks," *REV Journal on Electronics and Communications*, vol. 13, no. 1-2, 2023.

APPENDIX

This appendix presents a detailed derivation of the closed-form expression for the FIM corresponding to the model defined in Equation (14).

From (13) we have

$$\boldsymbol{z}' = \frac{\boldsymbol{z}}{\alpha} = \frac{(1-\alpha)}{\alpha}\boldsymbol{v} + \mathbf{n} = \boldsymbol{\theta} + \mathbf{n}$$

The conditional probability density function $f(\boldsymbol{z}'|\boldsymbol{v})$ is written as:

$$f(\boldsymbol{z}'|\boldsymbol{v}) = \frac{1}{\sqrt{(2\pi)^d|\Sigma_{\mathbf{n}}|}}\cdot \exp\left(-\frac{1}{2}(\mathbf{z}' - \tfrac{1-\alpha}{\alpha}\boldsymbol{v})^\top \Sigma_{\mathbf{n}}^{-1}(\mathbf{z}' - \tfrac{1-\alpha}{\alpha}\boldsymbol{v})\right)$$

The log of $f(\boldsymbol{z}'|\boldsymbol{v})$:

$$\log f(\boldsymbol{z}'|\boldsymbol{v}) = \log\left(\frac{1}{(2\pi)^{d/2}}\sigma_n^2\right) - \frac{1}{2\sigma_n^2}(\mathbf{z}' - \frac{1-\alpha}{\alpha}\boldsymbol{v})^\top(\mathbf{z}' - \frac{1-\alpha}{\alpha}\boldsymbol{v})$$

Taking the derivative with respect to $\boldsymbol{v}$:

$$\begin{aligned}\frac{\partial}{\partial v_i}\log f(\boldsymbol{z}' \mid \boldsymbol{v}) &= -\frac{1}{2\sigma_n^2}\frac{\partial}{\partial v_i}\sum_{i=1}^{d}(z_i' - \frac{1-\alpha}{\alpha}v_i)^2 \\ &= \frac{-1}{\sigma_n^2}\sum_{k=1}^{d}(z_i' - \bar{z_i'})\frac{\partial \bar{z}_i}{\partial v_i} \\ &= \frac{1}{\sigma_n^2}\cdot\frac{1-\alpha}{\alpha}\sum_{i=1}^{d}(z_i' - \bar{z}_i')\end{aligned}$$

Second derivative:

$$\begin{aligned}\frac{\partial^2}{\partial v_i^2}\log f(\boldsymbol{z}' \mid \boldsymbol{v}) &= \frac{\partial}{\partial v_i}\left(\frac{1}{\sigma_n^2}\cdot\frac{1-\alpha}{\alpha}\sum_{i=1}^{d}(z_i' - \bar{z'}_i)\right) \\ &= \frac{1}{\sigma_n^2}\cdot\frac{1-\alpha}{\alpha}\cdot\left(-\frac{1-\alpha}{\alpha}\right)\end{aligned}$$

Thus, the FIM of the latent variable $\boldsymbol{z}$ is computed as:

$$\begin{aligned}\mathbf{F} &= \mathbf{F}(\pi) + E_{\boldsymbol{v}}\mathbf{F}(\boldsymbol{v}) \\ &= \frac{1}{\sigma_v^2} + E_{\boldsymbol{v}}\left[-\frac{\partial^2}{\partial v_j \partial v_i}\log(f(\boldsymbol{z}'|\boldsymbol{v}))\right] \\ &= \frac{1}{\sigma_v^2} + \frac{1}{\sigma_n^2}\cdot\left(\frac{1-\alpha}{\alpha}\right)^2\end{aligned}$$

The derivation of (16) follows similar steps and is omitted for brevity.