How Not to Measure Disentanglement

Anna Sepliarskaia [1] **Julia Kiseleva** [2] **Maarten de Rijke** [3]

Abstract

To evaluate disentangled representations several metrics have been proposed. However, theoretical guarantees for conventional metrics of disentanglement are missing. Moreover, conventional metrics do not have a consistent correlation with the outcomes of qualitative studies. In this paper we analyze metrics of disentanglement and their properties. We conclude that existing metrics of disentanglement were created to reflect different characteristics of disentanglement and do not satisfy two basic desirable properties: (1) assign a high score to representations that are disentangled according to the definition; and (2) assign a low score to representations that are entangled according to the definition.

1. Introduction

Algorithms for learning representations are crucial for a variety of machine learning tasks, including image classification (Vincent et al., 2008; Hinton & Salakhutdinov, 2006) and image generation (Goodfellow et al., 2014; Makhzani et al., 2015). One type of representation learning algorithm is designed to create a disentangled representation. While there is no standardized definition of a disentangled representation, the key intuition is that a disentangled representation should capture and separate the generative factors (Bengio et al., 2013; Higgins et al., 2018). In this paper, we assume that the *generative factors* of the dataset are interpretable factors that describe every sample from the dataset.

Consider, for example, a dataset containing rectangles of different shapes. The disentanglement of the representation depends on the chosen set of generative factors. One possible set of generative factors on this dataset are the length and width (see Fig. 1). In the disentangled latent representation with respect to the length and width we can choose two latent factors. One of these factors is an invertible function of the length of the rectangles. Another is an invertible function of the width of the rectangles.

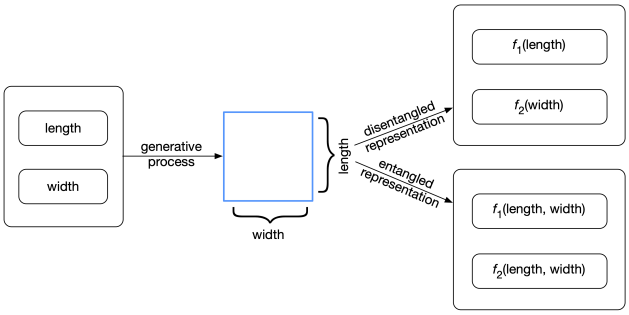


Figure 1. Example of a dataset containing rectangles

Learning a disentangled representation is an important step towards better representation learning because a disentangled representation contains information about elements in a dataset in an interpretable and compact structure (Bengio et al., 2013; Higgins et al., 2018). Therefore, the development of algorithms that learns disentangled representations has become an active area of research (Detlefsen & Hauberg, 2019; Dezfouli et al., 2019; Lorenz et al., 2019). The conventional way to measure the quality of these algorithms is to provide the results according to one of following metrics (Locatello et al., 2018): *BetaVAE* (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018), DCI (Eastwood & Williams, 2018), SAP score (Kumar et al., 2017), and MIG (Chen et al., 2018). However, it has been shown that the outcomes of these metrics are inconsistent with the outcomes of a qualitative study of the disentanglement of learned representations (Abdi et al., 2019); moreover, it is not clear which metric should be preferred.

In this paper, we theoretically analyze the conventional metrics. The outcome of our analysis is an understanding of the reasons why conventional metrics do not always correlate with each other: different metrics were designed to reflect different characteristics of disentanglement. As a consequence, a metric for evaluating an algorithm that learns disentangled representations should be determined by characteristic of disentanglement that the method is designed

[1]Vienna University of Technology, Vienna, Austria. [2]Microsoft Research AI, Seattle, WA, USA. [3]University of Amsterdam, Amsterdam, The Netherlands. Correspondence to: Anna Sepliarskaia <anna.sepliarskaia@tuwien.ac.at>, Julia Kiseleva <Julia.kiseleva@microsoft.com>, Maarten de Rijke <m.derijke@uva.nl>.

to reflect. Moreover, we also discover why outcomes of conventional metrics are inconsistent with the definition of disentanglement. We check whether the metrics satisfy two basic desirable properties: (1) assign a high score to representations that are disentangled according to the definition; and (2) assign a low score to representations that are entangled according to the definition. We show that the majority of the metrics do not satisfy these two conditions.

In summary, our key contribution in this paper is that we review existing metrics of disentanglement and discuss their fundamental properties.

2. Metrics of Disentanglement of Representations

The main purpose of this paper is to analyze conventional metrics of disentangled representations (the formal definitions of the metrics are given in the Appendix), which is done in this section. Though there is no universally accepted definition of disentanglement, most metrics are based on the definition proposed in (Bengio et al., 2013) and reflect characteristics of a disentangled representation in accordance with this definition. However, conventional metrics were designed to reflect **different** characteristics of disentangled representations: conventional metrics can be divided into two groups, depending on which characteristic they reflect. In this paper, we analyze whether conventional metrics satisfy the following fundamental properties:

Property 1. *A metric gives a high score to all representations that satisfy the characteristic that the metric reflects.*

Property 2. *A metric gives a low score for all representations that do not satisfy the characteristic that the metric reflects.*

2.1. BetaVAE, FactorVAE and DCI

In this subsection, we analyze metrics that reflect the following characteristic of disentangled representations.

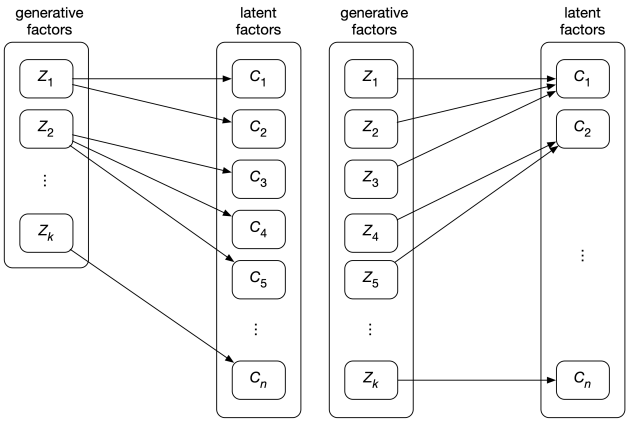
Characteristic 1. In a *disentangled representation* a change in one latent dimension corresponds to a change in one generative factor while being relatively invariant to changes in other generative factors (see Fig. 2a).

2.1.1. Analysis of whether metrics satisfy the property 1

Fact 1. BetaVAE *and* FactorVAE *do not satisfy Property 1.*

Proof. In a representation that satisfies Characteristic 1 there could be several generative factors that are not captured by any latent factors. In this case BetaVAE and FactorVAE cannot distinguish these generative factors. □

Fact 2. *DCI does not satisfy Property 1.*



(a) First characteristic of disentanglement. (b) Second characteristic of disentanglement.

Figure 2. Different characteristics of disentanglement.

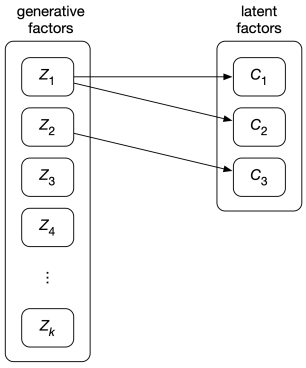


Figure 3. Example of the representation, satisfying Characteristic 1, but $BetaVAE = FactorVAE = \frac{3}{K}$.

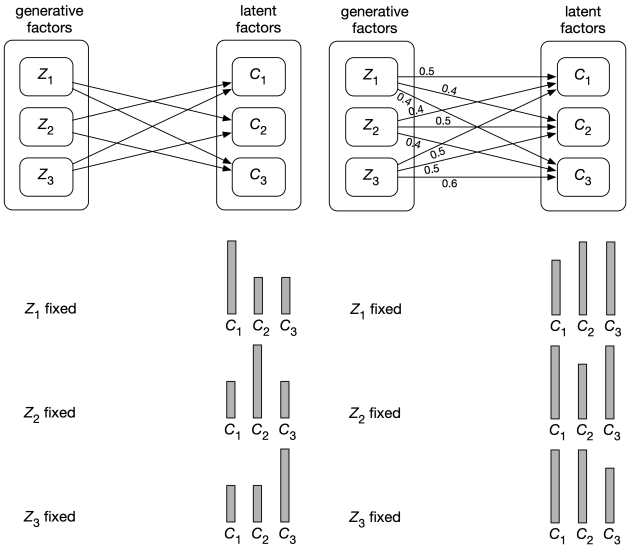
Proof. We argue that using entropy as a score of disentanglement of one latent variable is not correct. Indeed, a score of disentanglement of c_i should be high when c_i reflects one generative factor well, while it reflects other generative factors equally poorly. However, since the distribution may be close to uniform for these generative factors, the entropy is large. Let us provide an example that is built on this observation. Suppose there are 11 generative factors, and 11 is the dimension of the latent representation. Each latent factor c_i captures primarily a generative factor z_i:

$$I_{i,i} = 0.8,\ I_{i,k} = 0.02,\ k \neq i.$$

Then, the DCI score is 0.6, so the DCI assigns a small score to a representation that satisfies Characteristic 1. □

2.1.2. Analysis of whether metrics satisfy the property 2

Fact 3. BetaVAE *does not satisfy Property 2.*



(a) Entangled representation with BetaVAE equal to 1.

(b) Entangled representation with FactorVAE equal to 1.

Figure 4. Failures of BetaVAE and FactorVAE

Proof. As a proof, we give a counterexample (see Fig. 4a). Suppose there are 3 generative factors from a uniform distribution and the dimension of the latent representation is 3. Assume that the latent variables are equal to the generative factors with the following probabilities:

$$p_1 = (0.5, 0.5, 0),\ p_2 = (0, 0.5, 0.5),\ p_3 = (0.5, 0, 0.5).$$

We generate 10,000 training points with a batch size of 128. The accuracy of the linear classifier is equal to 0.9967 in this case, but the latent representation does not satisfy Characteristic 1. This shows that BetaVAE does not satisfy Property 2. □

Fact 4. FactorVAE *does not satisfy Property 2.*

Proof. Let us consider the following example (see Fig. 4b). Suppose there are 3 generative factors from a Gaussian distribution with $\mu = 0, \sigma = 1$, and each latent variable is a weighted sum of the generative factors:

$$c_1 = 0.5 \cdot z_1 + 0.4 \cdot z_2 + 0.5 \cdot z_3$$
$$c_2 = 0.4 \cdot z_1 + 0.5 \cdot z_2 + 0.5 \cdot z_3$$
$$c_3 = 0.4 \cdot z_1 + 0.4 \cdot z_2 + 0.6 \cdot z_3.$$

We generate 10,000 training points with a batch size of 128. The FactVAE disentanglement score is equal to 1 in this case, but the representation does not satisfy Characteristic 1. This shows that FactoVAE does not satisfy Property 2. □

Fact 5. *DCI does not satisfy Property 2.*

Proof. We give a counterexample, which is built on the fact that the weighted sum in Eq. 2 can be large if only one latent variable is disentangled, while the other latent variables do not capture any information about generative factors. Suppose there are 2 generative factors and the dimension of the latent representation is 2, and the matrix of informativeness is the following:

$$P_{0,0} = 1, P_{0,1} = 0, P_{1,1} = 0.09, P_{1,0} = 0.01.$$

In this case, the DCI score is 0.957. This counterexample shows that the DCI score can be close to 1 for the representation does not satisfy Characteristic 1. □

2.2. SAP and MIG metrics

In this subsection, we analyze metrics that reflect the following characteristic of disentangled representations.

Characteristic 2. In a *disentangled representation* a change in a single generative factor leads to a change in a single factor in the learned representation (see Fig. 2b).[1]

2.2.1. Analysis of whether metrics satisfy property 1

Fact 6. *SAP does not satisfy Property 1.*

Proof. We claim that it is incorrect to use the R^2 score of linear regression as informativeness between latent variables and generative factors. Indeed, a linear regression cannot capture non-linear dependencies. Thus, informativeness, which is calculated using the R^2 score of a linear regression, may be low if each generative factor is a non-linear function of some latent variable. Let us give an example that is built on this observation. Suppose there are 2 generative factors from the uniform distribution $U([-1, 1])$ and the dimension of the latent representation is 2. Let us assume the latent variables are obtained from the generative factors according to the following equations:

$$c_1 = z_1^{15},\ c_2 = z_2^{15}.$$

For this representation, we generate 10,000 examples and obtain the SAP score equal to 0.32. It proves that SAP can assign a low score to a representation that satisfies Characteristic 2. □

Fact 7. *MIG satisfies Property 1.*

Proof. Indeed, in a disentangled representation each generative factor is primarily captured in only one latent dimension. This means that for each generative factor z_j, there is exactly one latent factor c_{i_j} for which z_j is a function of c_{i_j}:

[1] This property of representations is also called *completeness* (Eastwood & Williams, 2018).

$z_j \sim f(c_{i_j})$. Therefore,

$$I_{i_j,j} = H(z_j) - H(z_j|c_{i,j}) \sim H(z_j),$$

whereas for other latent variables $I_{k,j} = I(c_k, z_j) \sim 0$. Consequently, according to MIG, the score of disentanglement of each generation factor is close to 1:

$$\frac{I_{i_j,j} - \max_{k \neq i_j} I_{k,j}}{H(z_j)} \sim 1. \quad (1)$$

Therefore, the average of these scores is also close to 1. This shows that MIG always assigns a high score to a representation that satisfies Characteristic 2. □

2.2.2. Analysis of whether metrics satisfy property 2

Fact 8. *SAP does not satisfy Property 2.*

Proof. A high SAP score indicates that the majority of generative factors is captured linearly in only one latent dimension. However, the SAP metric does not penalize the existence of several latent factors that capture the same generative factor non-linearly. Let us consider the following example. Suppose there are 2 generative factors from the uniform distribution $U([-1, 1])$, and the dimension of the latent representation is 3. Let us assume that the latent factors are obtained from the generative factors according to the following equations:

$$c_1 = z_1,\ c_2 = z_1^{25} + z_2^{25},\ c_3 = z_2.$$

For this latent representation, a change in each generative factor leads to a change in several latent factors, but the SAP score is equal to 0.98. This shows that the SAP score can be close to 1 for a latent representation that does not satisfy Characteristic 1. □

Fact 9. *MIG satisfies Property 2.*

Proof. A high MIG score indicates that the majority of generative factors is captured in only one latent dimension. Consequently, a change in one of the generative factors entails a change primarily in only one latent dimension. □

A summary of the results of our analysis is given in Table 1.

2.3. Difference between Characteristics 1 and 2

The Characteristics 1 and 2 of a disentangled representation have important differences. Indeed, a representation in which several latent factors capture one common generative factor satisfies a Characteristic 1, but not a Characteristic 2. On the other hand, a representation in which a latent variable captures multiple generative factors while there are no other latent variables that capture these generative factors does not satisfy Characteristic 1, but satisfies Characteristic 2.

Table 1. Summary of facts about proposed metrics of disentangled representations.

Metric	Satisfies Property 1	Satisfies Property 2
BetaVAE	No	No
FactorVAE	No	No
DCI	No	No
SAP	No	No
MIG	Yes	Yes

Consider, for example, the following latent representation of dimension 4 of the dataset containing rectangles of different shapes shown in Fig. 1:

$$z_1 = x,\ z_2 = x^2,\ z_3 = y,\ z_4 = y^3,$$

where x is the length of a rectangle, while y is the width of a rectangle. It satisfies Characteristic 1, but not a Characteristic 2. Conversely, any one-dimensional latent representation of the same dataset would satisfy Characteristic 2, but not necessarily Characteristic 1.

3. Conclusion

In recent years, several models have been developed to obtain disentangled representations (Yu & Grauman, 2017; Hu et al., 2017; Denton et al., 2017; Kim & Mnih, 2018). Currently, there are five metrics that are commonly used to evaluate the models: *BetaVAE* (Higgins et al., 2017), FactorVAE (Kim & Mnih, 2018), DCI (Eastwood & Williams, 2018), SAP (Kumar et al., 2017) and MIG (Chen et al., 2018). Interestingly, all of these metrics are based upon the definition of disentangled representation proposed in (Bengio et al., 2013). However, three of the metrics were designed to reflect Characteristic 1 of disentangled representations, while two were designed to reflect Characteristic 2.

The primary goal of this paper has been to provide an analysis of the existing metrics of disentangled representations. We theoretically analyze how well the proposed metrics reflect the characteristics of disentangled representations that they are intended to reflect. In particular, we analyze each of the existing metrics of disentanglement by two properties: whether a metric is close to 1 when a representation satisfies the characteristic that the metric reflects and whether the metric is close to 0 when a representation does not satisfy the characteristic. Surprisingly, we found that most of the existing metrics do not satisfy these basic properties.

References

Abdi, A. H., Abolmaesumi, P., and Fels, S. A preliminary study of disentanglement with insights on the inadequacy of metrics. *arXiv preprint arXiv:1911.11791*, 2019.

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

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Denton, E. L. et al. Unsupervised learning of disentangled representations from video. In *Advances in Neural Information Processing Systems*, pp. 4414–4423, 2017.

Detlefsen, N. S. and Hauberg, S. Explicit disentanglement of appearance and perspective in generative models. *arXiv preprint arXiv:1906.11881*, 2019.

Dezfouli, A., Ashtiani, H., Ghattas, O., Nock, R., Dayan, P., and Ong, C. S. Disentangled behavioral representations. *bioRxiv*, pp. 658252, 2019.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hinton, G. E. and Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science*, 313 (5786):504–507, 2006.

Hu, Q., Szabó, A., Portenier, T., Zwicker, M., and Favaro, P. Disentangling factors of variation by mixing them. *arXiv preprint arXiv:1711.07410*, 2017.

Kim, H. and Mnih, A. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Lorenz, D., Bereska, L., Milbich, T., and Ommer, B. Unsupervised part-based disentangling of object shape and appearance. *arXiv preprint arXiv:1903.06946*, 2019.

Makhzani, A., Shlens, J., Jaitly, N., Goodfellow, I., and Frey, B. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103. ACM, 2008.

Yu, A. and Grauman, K. Semantic jitter: Dense supervision for visual comparisons via synthetic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5571–5580. IEEE, 2017.

Supplement to "How Not to Measure Disentanglement"

A. Metrics of Disentanglement of Representations

A.1. Definition of BetaVAE

The algorithm that calculates *BetaVAE* (Higgins et al., 2017) consists of the following steps:

1. Choose a generative factor z_r.
2. Generate a batch of pairs of vectors for which the value of z_r within the pair is equal, while other generative factors are chosen randomly:

$$(\mathbf{p}_1 = \langle z_{1,1}, \ldots, z_{1,K} \rangle, \mathbf{p}_2 = \langle z_{2,1}, \ldots, z_{2,K} \rangle), \\ z_{1,r} = z_{2,r}$$

3. Calculate the latent code of the generated pairs: $(\mathbf{c}_1 = f_e(g(\mathbf{p}_1)), \mathbf{c}_2 = f_e(g(\mathbf{p}_2)))$
4. Calculate the absolute value of the pairwise differences of these representations:

$$\mathbf{e} = \langle \left| c_{1,1} - c_{2,1} \right|, \ldots, \left| c_{1,N} - c_{2,N} \right| \rangle$$

5. The mean of these differences across the examples in the batch gives one training point for the linear regressor that predicts which generative factor was fixed.
6. BetaVAE is the accuracy of the linear regressor.

A.2. Definition of FactorVAE

The idea behind FactorVAE (Kim & Mnih, 2018) is very similar to BetaVAE. The main difference between them concerns how a batch of examples is generated to obtain a variation of latent variables when one generative factor is fixed. Another difference is the classifier that predicts which generative factor was fixed using the variation of latent variables. *FactorVAE* can be calculated by performing the following steps:

1. Choose a generative factor z_r.
2. Generate a batch of vectors for which the value of z_r within the batch is fixed, while other generative factors are chosen randomly.
3. Calculate latent codes of vectors from one batch.
4. Normalize each dimension in the latent representation by its empirical standard deviation over the full data.
5. Take the empirical variance in each dimension of these normalized representations.
6. The index of the dimension with the lowest variance and the target index r provides one training point for the classifier.
7. FactorVAE is the accuracy of the classifier.

A.3. DCI: Disentanglement, Completeness and Informativeness

Eastwood & Williams (2018) propose to use a metric of disentangled representations, which we call DCI, that is calculated as follows:

1. First, the *informativeness* between c_i and z_j is calculated. To determine the informativeness between c_i and z_j, Eastwood & Williams (2018) suggest training K regressors. Each regressor f_j predicts z_j given $\mathbf{c}$ ($\hat{z}_j = f_j(\mathbf{c})$) and can provide an importance score $P_{i,j}$ for each c_i. The normalized importance score obtained by regressor f_j for variable c_i is used as the informativeness between c_i and z_j:

$$I_{i,j} = \frac{P_{i,j}}{\sum_{k=0}^{k=K} P_{i,k}}.$$

2. For each latent variable its score of disentanglement is calculated as follows:

$$H_K(I_i) = 1 + \sum_{k=1}^{K} I_{i,k} \log_K I_{i,k}.$$

3. The weighted sum of the obtained scores of disentanglement for the latent variables is DCI:

$$\text{DCI}(\mathbf{c}, \mathbf{z}) = \sum_i \left(\rho_i \cdot H_K(I_i) \right), \qquad (2)$$

where $\rho_i = \sum_j P_{i,j} / \sum_{ij} P_{i,j}$.

A.4. SAP score: Separated Attribute Predictability

Kumar et al. (2017) provide a metric of disentanglement that is calculated as follows:

1. Compute a *matrix of informativeness* $I_{i,j}$, in which the ij-th entry is the linear regression or classification score of predicting the j-th generative factor using only the i-th variable in the latent representation.
2. For each column in the matrix of informativeness $I_{i,j}$, which corresponds to a generative factor, calculate the difference between the top two entries (corresponding to the top two most predictive latent factors). The average of these differences is the final score, which is called the SAP:

$$\text{SAP}(\mathbf{c}, \mathbf{z}) = \frac{1}{K} \sum_k \left(I_{i_k,k} - \max_{l \neq i_k} I_{l,k} \right),$$

where $i_k = \arg\max_i I_{i,k}$.

A.5. MIG: Mutual Information Gap

Chen et al. (2018) propose a disentanglement metric, Mutual Information Gap (MIG), that uses mutual information

between the j-th generative factor and the i-th latent variable as a notion of informativeness between them. The *mutual information* between two variables c and z is defined as

$$I(c; z) = H(z) - H(z|c),$$

where $H(z)$ is the entropy of the variable z. Mutual information measures how much knowing one variable reduces uncertainty about the other. A useful property of mutual information is that it is always non-negative $I(c; z) > 0$. Moreover, $I(c; z)$ is equal to 0 if and only if c and z are independent. Also, mutual information achieves its maximum if there exists an invertible relationship between c and z. The following algorithm calculates the MIG score:

1. Compute a *matrix of informativeness* $I_{i,j}$, in which the ij-th entry is the mutual information between the j-th generative factor and the i-th latent variable.
2. For each column of the score matrix $I_{i,j}$, which corresponds to a generative factor, calculate the difference between the top two entries, and normalize it by dividing by the entropy of the corresponding generative factor. The average of these normalized differences is the MIG score:

$$\text{MIG}(\mathbf{c}, \mathbf{z}) = \frac{1}{K} \sum_{k} \frac{I_{i_k,k} - \max_{l \neq i_k} I_{l,k}}{H(z_k)},$$

   where $i_k = \arg\max_i I_{i,k}$.