Journal of Machine Learning Research 24 (2023) 1-30 Submitted 9/21; Revised 1/23; Published 12/23

Be More Active! Understanding the Differences Between Mean and Sampled Representations of Variational Autoencoders

Lisa Bonheme LB732@KENT.AC.UK
School of Computing
University of Kent
Canterbury, UK

Marek Grzes M.GRZES@KENT.AC.UK
School of Computing
University of Kent
Canterbury, UK

Editor: David Wipf

Abstract

The ability of Variational Autoencoders to learn disentangled representations has made them appealing for practical applications. However, their mean representations, which are generally used for downstream tasks, have recently been shown to be more correlated than their sampled counterpart, on which disentanglement is usually measured. In this paper, we refine this observation through the lens of selective posterior collapse, which states that only a subset of the learned representations, the active variables, is encoding useful information while the rest (the passive variables) is discarded. We first extend the existing definition to multiple data examples and show that active variables are equally disentangled in mean and sampled representations. Based on this extension and the pre-trained models from `disentanglement_lib`, we then isolate the passive variables and show that they are responsible for the discrepancies between mean and sampled representations. Specifically, passive variables exhibit high correlation scores with other variables in mean representations while being fully uncorrelated in sampled ones. We thus conclude that despite what their higher correlation might suggest, mean representations are still good candidates for downstream tasks applications. However, it may be beneficial to remove their passive variables, especially when used with models sensitive to correlated features.

Keywords: Representation learning, Disentangled representations, Deep generative models, Variational autoencoders, Posterior collapse

1 Introduction

Variational Autoencoders (VAEs) are considered state-of-the-art techniques to learn unsupervised disentangled representations, that is, representations encoding separately the different factors of variations (Bengio et al., 2013). Disentangled representations are very attractive in terms of interpretability and fairness (Locatello et al., 2019a), and can be beneficial for downstream tasks such as abstract reasoning (van Steenkiste et al., 2019).

Over the years, multiple regularisation techniques have been developed to encourage disentanglement with a specific focus on enforcing the learned latent factors to be uncorrelated. As we will discuss in Section 2, while this regularisation is done on the sampled aggregated posterior,

the learned representation is generally taken to be the mean vector of the posterior distribution. However, Locatello et al. (2019b) reported an increased total correlation (TC) and averaged mutual information (MI) over the dimensions of the mean representation compared to the results obtained on its sampled counterpart. This finding raises questions on whether mean representations would still benefit from the appealing attributes of disentanglement since sampled representations were shown to be less correlated, and thus more disentangled.

Another line of research has shown that VAEs are behaving in a polarised regime, also known as selective posterior collapse (Dai and Wipf, 2018; Rolinek et al., 2019). In this regime, the relevant dimensions of the sampled representations (the active variables) are used by the decoder for reconstruction while the remaining dimensions (the passive variables) are 'shut down' to closely match the prior. However, the polarised regime has only been studied in the context of single data examples for sampled representations. Therefore, in Section 3, we extend the existing definition of the polarised regime to multiple data examples and explore the implications for mean and variance representations. Assuming that VAEs producing disentangled representations are behaving in a polarised regime, we show, based on this extended version, that active variables of mean representations should not be more correlated than the sampled ones. Thus, we argue that the higher correlation reported by Locatello et al. (2019b) is due to the impact of the passive variables on the metrics used. We verify this hypothesis empirically in Section 4, and provide further analytical justifications in Section 5.

Our contribution is three-fold: (1) we extend the definition of the polarised regime to mean and variance representations using multiple data examples. (2) we use this extended version to show that the discrepancies between mean and sampled representations observed by Locatello et al. (2019b) are mostly due to the impact of the polarised regime, and especially of the passive variables. (3) we explain why passive variables are leading to higher TC and averaged MI scores. The code of our experiments is available at `https://github.com/bonheml/tc_study`.

Notational considerations Throughout this paper, we use the superscript $^{(i)}$ to denote the values obtained for the ith sample $\mathbf{x}^{(i)}$ of the random variable $\mathbf{x}$, and represent the jth dimension of a vector representation using the subscript $_j$. For example, given a random variable $\boldsymbol{\epsilon}$ distributed according to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$, $\boldsymbol{\epsilon}_j^{(i)}$ is the jth dimension of the sample of $\boldsymbol{\epsilon}$ obtained for $\mathbf{x}^{(i)}$. While the mean and diagonal variance representations are functions of $\mathbf{x}$ and the network parameters ϕ, we use a shortened version when the meaning is clear from the context, such that $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\mathbf{x}; \phi)$, and $\boldsymbol{\sigma} \triangleq diag[\boldsymbol{\Sigma}(\mathbf{x}; \phi)]$. Similarly, for a specific sample $\mathbf{x}^{(i)}$, $\boldsymbol{\mu}^{(i)} \triangleq \boldsymbol{\mu}(\mathbf{x}^{(i)}; \phi)$, and $\boldsymbol{\sigma}^{(i)} \triangleq diag[\boldsymbol{\Sigma}(\mathbf{x}^{(i)}; \phi)]$. We adopt the same notation for the sampled representation, such that $\mathbf{z} \triangleq \boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}$, and $\mathbf{z}^{(i)} \triangleq \boldsymbol{\mu}^{(i)} + \boldsymbol{\epsilon}^{(i)}(\boldsymbol{\sigma}^{(i)})^{1/2}$.

2 Background

2.1 Variational Autoencoders

Variational Autoencoders (VAEs) (Kingma and Welling, 2014; Rezende and Mohamed, 2015) are deep probabilistic generative models based on variational inference. The encoder maps some input $\mathbf{x}^{(i)}$ to a latent representation $\mathbf{z}^{(i)}$, and the decoder uses these latent variables to generate an output

$\hat{\mathbf{x}}^{(i)}$ similar to $\mathbf{x}^{(i)}$. This can be optimised by maximising the evidence lower bound (ELBO):

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}) = \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big)}_{\text{regularisation term}}. \tag{1}$$

Generally, $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$ and $p(\mathbf{z})$ are modelled as multivariate Gaussian distributions to permit closed form computation of the regularisation term (Doersch, 2016).

As illustrated in Figure 1, given the mean $\boldsymbol{\mu}$ and diagonal covariance $\boldsymbol{\sigma}$ of a random variable $\mathbf{x}$, the sampled representation is obtained using the reparameterisation trick (Kingma and Welling, 2014) such that $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma}^{1/2}\boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a random variable with a Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

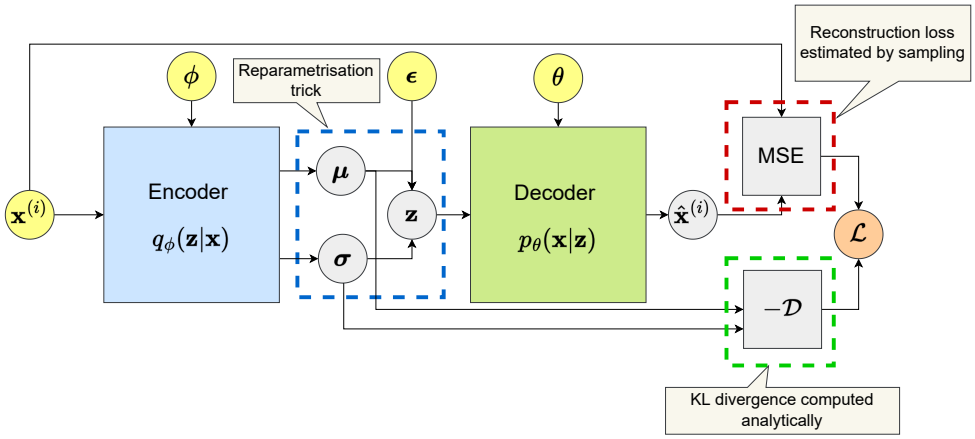


Figure 1: Illustration of a VAE during the training process. The distributions are assumed to be multivariate Gaussian, $\boldsymbol{\mu}$ is the mean layer and $\boldsymbol{\sigma}$ is the variance layer. $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are the parameters of the posterior over $\mathbf{z}$.

In this paper we are interested in investigating the discrepancies between the mean and sampled representations, that is $\boldsymbol{\mu}$ and $\mathbf{z}$ respectively. Specifically, our goal is to explain the higher correlation of mean representations reported by Locatello et al. (2019b). In the following sections, the representations learned by the mean layer will be referred to as mean representations, and those learned during the sampling stage as sampled representations, as per Figure 2.

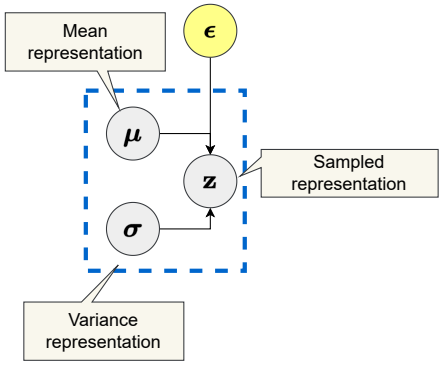


Figure 2: Mean, variance, and sampled representations of VAEs.

2.2 Disentangled representation learning with VAEs

While various learning objectives and architectures have been proposed for VAEs, we will focus on the family of methods increasing the weight on the regularisation term of Equation 1 to produce disentangled representations, similarly to Locatello et al. (2019b). We refer the reader to Tschannen et al. (2018) for a broader overview of the existing VAE architectures.

β-VAE Higgins et al. (2017) introduced a new learning objective whose goal was to bias the encoding-reconstruction trade-off by penalising the regularisation term more strongly. This is formulated as the following learning objective:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - \beta D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big). \tag{2}$$

Equation 2 is similar to the original VAE objective seen in Equation 1 with the addition of the β parameter which, when $\beta > 1$, increases the bias on the encoding optimisation. One of the downsides of penalising the encoding more strongly is that the reconstruction is of lower quality.

Annealed VAE Burgess et al. (2018) provided an analysis of β-VAE disentangled representations through the lens of information theory, based on the learning objective described by Alemi et al. (2017). They argue that because β-VAE is increasing the pressure on the encoding capacity of the network, the optimal way to encode information would be on separate dimensions, leading to disentanglement. They hypothesise that β-VAE will learn the latent variables having the most impact on the reconstruction first, then gradually optimise less critical variables. To ease the learning of these less important latent variables, they propose to gradually increase the encoding capacity during the training process, relaxing the initial constraint. This leads to the following objective, where C is a parameter that can be understood as a channel capacity and γ is a hyper-parameter penalising the divergence, similarly to β in β-VAE:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\big[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})\big] - \gamma\left|D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big) - \mathrm{C}\right|. \tag{3}$$

As the training progresses, the channel capacity C is increased, going from zero to its maximum channel capacity $\mathrm{C}_{\max}$. For example, given a maximum channel capacity of 100, during the first training step, any deviation from the KL divergence will be penalised similarly to β-VAE because $\mathrm{C} = 0$. After n steps, once the channel capacity will be annealed to its maximum value, the KL divergence will be penalised only when it is higher than 100. VAEs that use Equation 3 as a learning objective will be referred to as annealed VAEs in the rest of this paper.

Factor VAE In Factor VAE, Kim and Mnih (2018) addressed disentanglement learning by enforcing the independence of the latent factors learned by VAEs. This approach is slightly different from the β-VAE objective presented in Equation 2 as it takes into account the KL divergence $D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big)$ to further decompose the expected value over the data distribution of the second term of the loss function as:

$$\mathbb{E}_{p_{\boldsymbol{\theta}}(\mathbf{x})}\big[D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big)\big] = \mathrm{I}_e(\mathbf{x};\mathbf{z}) + D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big).$$

Kim and Mnih (2018) suggested that β-VAE's lower data generation quality is due to the penalisation of $\mathrm{I}_e(\mathbf{x};\mathbf{z})$, the mutual information between the observed and latent variables. They argued

that only the distance between the estimated latent factors and the prior should be penalised to encourage disentanglement, and they proposed a new objective to this end:

$$\mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}\big(\mathbf{x}|\mathbf{z}\big)\right] - D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}\big(\mathbf{z}|\mathbf{x}\big) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big)\right] - \gamma D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big). \tag{4}$$

Here, $D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}(\mathbf{z}) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big)$ is approximated by penalising the dependencies between the dimensions of $q_{\boldsymbol{\phi}}(\mathbf{z})$:

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}\big(\mathbf{x}^{(i)}|\mathbf{z}\big)\right] - D_{\mathrm{KL}}\big(q_{\boldsymbol{\phi}}\big(\mathbf{z}|\mathbf{x}^{(i)}\big) \parallel p_{\boldsymbol{\theta}}(\mathbf{z})\big)\right] - \underbrace{\gamma D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \,\middle\|\, \prod_{j=1}^{\mathrm{d}} q_{\boldsymbol{\phi}}(\mathbf{z}_j)\right)}_{\text{total correlation}}. \tag{5}$$

As the computation of the total correlation defined in Equation 5 is intractable, it is estimated by sampling a batch from $q_{\boldsymbol{\phi}}(\mathbf{z})$ and shuffling the values of each dimension of the latent variables to obtain the samples for $\prod_{j=1}^{\mathrm{d}} q_{\boldsymbol{\phi}}(\mathbf{z}_j)$. A binary classifier is then trained to recognise the samples belonging to $q_{\boldsymbol{\phi}}(\mathbf{z})$, and the density ratio is computed using the probability $p_{classif}(\mathbf{z})$ given by the classifier that the samples belong to $q_{\boldsymbol{\phi}}(\mathbf{z})$:

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}) \,\middle\|\, \prod_{j=1}^{\mathrm{d}} q_{\boldsymbol{\phi}}(\mathbf{z}_j)\right) \approx \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}\left[\log \frac{p_{classif}(\mathbf{z})}{1 - p_{classif}(\mathbf{z})}\right].$$

In Kim and Mnih (2018, Appendix F and I) the authors compared the results obtained with Equation 5 and Equation 4 (i.e., where $p_\theta(\mathbf{z})$ is not approximated by $\prod_{j=1}^{\mathrm{d}} q_{\boldsymbol{\phi}}(\mathbf{z}_j)$), but obtained lower disentanglement with the latter. They concluded that enforcing a factorised $q_\phi(\mathbf{z})$ was more beneficial for disentanglement than enforcing $q_\phi(\mathbf{z})$ to be as close as possible to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$.

β-TC VAE Similarly to Kim and Mnih (2018), Chen et al. (2018) proposed to optimise Equation 5. The main difference is that Chen et al. (2018) are approximating the total correlation using mini-batch weighted sampling. Here, the estimation is computed over a mini-batch of samples $\{\mathbf{x}^{(i)}\}_{i=1}^{m}$ as follows:

$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z})}[\log q_{\boldsymbol{\phi}}(\mathbf{z})] \approx \frac{1}{m}\sum_{i=1}^{m}\left(\log \frac{1}{nm}\sum_{k=1}^{m} q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}|\mathbf{x}^{(k)})\right), \tag{6}$$

where m is the number of samples in the mini-batch, and n total number of input examples. $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}_j)}[\log q_{\boldsymbol{\phi}}(\mathbf{z}_j)]$ can be computed in a similar way. We refer the reader to Chen et al. (2018, Appendix C.1) for the detailed derivation of Equation 6.

DIP-VAE Similarly to Kim and Mnih (2018) and Chen et al. (2018), Kumar et al. (2018) proposed to regularise the distance between $q_\phi(\mathbf{z})$ and $p(\mathbf{z})$ using Equation 4. The main difference is that here $D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}) \parallel p(\mathbf{z})\big)$ is measured by matching the moments of the learned distribution $q_\phi(\mathbf{z})$ and its prior $p(\mathbf{z})$. The second moment of the learned distribution is given by:

$$\mathrm{Cov}_{q_{\boldsymbol{\phi}}(\mathbf{z})}[\mathbf{z}] = \mathrm{Cov}_{p(\mathbf{x})}\left[\boldsymbol{\mu}\right] + \mathbb{E}_{p(\mathbf{x})}\left[\boldsymbol{\sigma}\right]. \tag{7}$$

Two divergences are then defined. The first, DIP-VAE I, penalises only the first term of Equation 7:

$$\lambda D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\big) = \lambda_{od}\sum_{i\neq j}\left(\mathrm{Cov}_{p(\mathbf{x})}\left[\boldsymbol{\mu}\right]\right)_{ij}^{2} + \lambda_d \sum_{i}\left(\mathrm{Cov}_{p(\mathbf{x})}\left[\boldsymbol{\mu}\right]_{ii} - 1\right)^2,$$

where λ_{od} and λ_d are the off-diagonal and diagonal regularisation terms, respectively. The second, DIP-VAE II, penalises both terms of Equation 7:

$$\lambda D_{\mathrm{KL}}\big(q_\phi(\mathbf{z}) \parallel p_\theta(\mathbf{z})\big) = \lambda_{od} \sum_{i \neq j} \Big(\mathrm{Cov}_{q_\phi(\mathbf{z})}\left[\mathbf{z}\right]\Big)^2_{ij} + \lambda_d \sum_i \Big(\mathrm{Cov}_{q_\phi(\mathbf{z})}\left[\mathbf{z}\right]_{ii} - 1\Big)^2.$$

Note that because DIP-VAE I directly encourages diagonal covariance matrices in the mean representation, it will have a low correlation in the mean representation, which, as observed by Locatello et al. (2019b), mirrors the correlations in the sampled representation. Moreover, the discrepancies between mean and sampled representations were observed by Locatello et al. (2019b) in the context of methods which explicitly regularise the disentanglement of $\mathbf{z}$, but not $\boldsymbol{\mu}$. In this study, we will thus consider DIP-VAE II, which enforces the covariance matrix of the sampled representation to be diagonal, but not DIP-VAE I as it explicitly regularises $\boldsymbol{\mu}$.

2.3 Benefits of disentanglement on downstream tasks

Disentangled representations have been shown to reduce the sample complexity of abstract reasoning tasks (van Steenkiste et al., 2019), improve the fairness of downstream task models (Locatello et al., 2019a; Creager et al., 2019) and their interpretability (Higgins et al., 2017; Adel et al., 2018). However, when the mean representations are more correlated than the sampled representations on which the disentanglement is measured, it will hamper the interpretability of downstream task models (Alin, 2010; Chan et al., 2022), and reduce their fairness (Locatello et al., 2019a; Träuble et al., 2021). Moreover, under a certain level of supervision, VAEs can provably provide identifiable representations (Khemakhem et al., 2020; Mita et al., 2021), and it is conjectured that under specific constraints, this is also possible in the unsupervised setting (Reizinger et al., 2022). It is thus important to investigate the origin of the discrepancies between mean and sampled representations to determine if one can still benefit from disentanglement when using mean representations on downstream tasks.

2.4 Related work

The discrepancy between the total correlation of mean and sampled representations observed by Locatello et al. (2019b) have recently been investigated by Cheng et al. (2021) who provide a theoretical justification of the higher total correlation scores of the mean representations. However, they did not consider the polarised regime which is a necessary condition for VAE to provide good reconstruction (Dai and Wipf, 2018; Rolinek et al., 2019; Dai et al., 2020). Thus, their work is complementary to ours in the case where VAEs are not learning in a polarised regime.

3 The polarised regime

The polarised regime, also known as selective posterior collapse, is the ability of VAEs to 'shut down' superfluous dimensions of their sampled latent representations while providing a high precision on the remaining ones (Rolinek et al., 2019; Dai et al., 2020). As a result, the sampled representation can be separated into two subsets of variables, active and passive. The active variables correspond to the subset of the sampled latent representation that is needed for the reconstruction. They have a low variance, and are close to the mean variables. The passive variables correspond to the superfluous dimensions that are discarded by the VAE. They follow a zero-mean unit-variance

Gaussian distribution to optimally match the prior and are ignored by the decoder, which only uses the variables that help to reconstruct the input.

The existence of active and passive variables has been shown to be a necessary condition for the VAEs to provide a good reconstruction (Dai and Wipf, 2018; Dai et al., 2020). However, when the weight on the regularisation term of the ELBO given in Equation 1 increases, VAEs are pruning more active variables to minimise the regularisation loss. When this weight becomes too large, the representations collapse to the prior, containing only passive variables (Lucas et al., 2019a; Dai et al., 2020).

3.1 The polarised regime for one data example

Based on the definition of Rolinek et al. (2019), we can characterise the active and passive variables of sampled representations as follows.

Definition 1 (Polarised regime) *When a VAE learns in a polarised regime, for a given data example* $\mathbf{x}^{(i)} \in \boldsymbol{X}$*, its mean, variance, and sampled representations,* $\boldsymbol{\mu}^{(i)} \triangleq \boldsymbol{\mu}(\mathbf{x}^{(i)}; \phi)$*,* $\boldsymbol{\sigma}^{(i)} \triangleq diag[\boldsymbol{\Sigma}(\mathbf{x}^{(i)}; \phi)]$*, and* $\mathbf{z}^{(i)} \triangleq \boldsymbol{\mu}^{(i)} + \boldsymbol{\epsilon}^{(i)}(\boldsymbol{\sigma}^{(i)})^{1/2}$*, respectively, are composed of a set of passive and active variables,* $\mathbb{V}_p^{(i)} \cup \mathbb{V}_a^{(i)}$ *such that, for each data example* $\mathbf{x}^{(i)}$*:*

(i) $|\boldsymbol{\mu}_j^{(i)}| \ll 1$*,* $\boldsymbol{\sigma}_j^{(i)} \approx 1$*, and* $\mathbf{z}_j^{(i)} \approx \boldsymbol{\epsilon}_j^{(i)} \quad \forall\, j \in \mathbb{V}_p^{(i)}$*,*

(ii) $\boldsymbol{\sigma}_j^{(i)} \ll 1$ *and* $\mathbf{z}_j^{(i)} \approx \boldsymbol{\mu}_j^{(i)} \quad \forall\, j \in \mathbb{V}_a^{(i)}$*,*

where $\boldsymbol{\epsilon}^{(i)} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$*,* j *is the* j^{th} *variable of a representation, and* $|\cdot|$ *denotes the absolute value.*

The polarised regime can also be seen as a sparsity-inducing mechanism which will prune the superfluous columns of the weights of the first layer of the decoder (Dai et al., 2017, 2018). As the corresponding dimensions of the mean and variance representations will not have any influence on the decoder (i.e., they are passive), they will only be optimised with respect to the KL divergence, thus becoming close to 0 and 1, respectively.

While Definition 1 provides an overview of the polarised regime for each data example, it is not readily usable to analyse discrepancies between mean and sampled representations as they are observed over the whole dataset. We will thus extend the definition of the polarised regime to multiple data examples in Section 3.2 and show that the discrepancies between mean and sampled representations can only originate from variables which are not active.

3.2 Generalisation of the polarised regime to multiple data examples

Now we will propose a new generalisation of Definition 1 to multiple data examples, which will serve as a basis for our analysis in Section 4. Given that a variable can either be active or passive for a given data example, when considering multiple data examples, three cases arise:

- A variable is passive for all the data examples.
- A variable is active for all the data examples.
- A variable is active for some data examples and passive otherwise.

These three types of variables are formalised in Definition 2 and illustrated in Figure 3.

Definition 2 (Variable types) *When considered over multiple data examples $\boldsymbol{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n}$, the latent representations are composed of a set of passive, active, and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$, which are defined as follows:*

(i) $\mathbb{V}_p \triangleq \bigcap_{i=1}^{n} \mathbb{V}_p^{(i)}$,

(ii) $\mathbb{V}_a \triangleq \bigcap_{i=1}^{n} \mathbb{V}_a^{(i)}$,

(iii) $\mathbb{V}_m \triangleq \Big(\bigcup_{i=1}^{n} \mathbb{V}_p^{(i)}\Big) \bigcap \Big(\bigcup_{i=1}^{n} \mathbb{V}_a^{(i)}\Big)$.

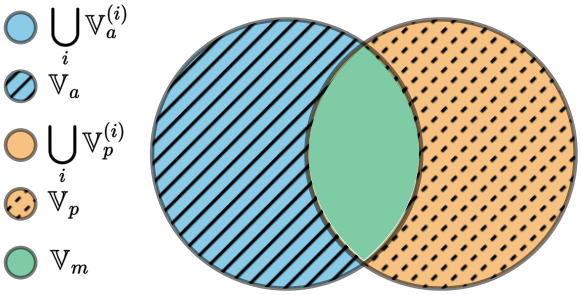


Figure 3: A graphical representation of Definition 2.

Based on Definitions 1 and 2, we will consider the properties of the mean and variance representations over multiple data examples in Propositions 3 and 4, then describe their implications for the sampled representations in Theorem 5.

Proposition 3 (Polarised regime of $\boldsymbol{\mu}$ over $\boldsymbol{X}$) *When a VAE learns in a polarised regime, its mean representation $\boldsymbol{\mu} \triangleq \boldsymbol{\mu}(\mathbf{x};\phi) \approx \boldsymbol{\mu}(\boldsymbol{X};\phi)$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$:*

(i) $|\bar{\boldsymbol{\mu}}_j| \ll 1$ *and* $Var(\boldsymbol{\mu}_j) \ll 1 \quad \forall\, j \in \mathbb{V}_p$,

(ii) $Var(\boldsymbol{\mu}_j) > Var(\boldsymbol{\mu}_k) \quad \forall\, j \in \mathbb{V}_a\ , \forall\, k \in \mathbb{V}_p$,

where $\bar{\boldsymbol{\mu}}_j \triangleq \mathbb{E}_{p(\mathbf{x})}[\boldsymbol{\mu}_j]$, and $Var(\cdot)$ denotes the variance.

The first point of Proposition 3 indicates that passive variables of mean representations are almost constant when considered over multiple data examples. Indeed, because they are passive for each data example, according to (i) of Definition 1 they will consistently take values close to 0. Thus, they will have a variance and expected value close to 0.

Because active variables of the mean representation encode some information about the input, their value will vary depending on the input and thus have a higher variance than their passive counterpart, as stated in (ii) of Proposition 3.

Now, let us analyse the effect of the polarised regime on sampled representations over multiple data examples.

Proposition 4 (Polarised regime of $\boldsymbol{\sigma}$ over $\boldsymbol{X}$) *When a VAE learns in a polarised regime, its variance representation $\boldsymbol{\sigma} \triangleq diag[\boldsymbol{\Sigma}(\mathbf{x};\phi)] \approx diag[\boldsymbol{\Sigma}(\boldsymbol{X};\phi)]$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$:*

(i) $\bar{\boldsymbol{\sigma}}_j \approx 1$ *and* $Var(\boldsymbol{\sigma}_j) \ll 1 \quad \forall j \in \mathbb{V}_p$,

(ii) $\bar{\boldsymbol{\sigma}}_j \ll 1$ *and* $Var(\boldsymbol{\sigma}_j) \ll 1 \quad \forall j \in \mathbb{V}_a$,

(iii) $Var(\boldsymbol{\sigma}_j) < Var(\boldsymbol{\sigma}_k) \quad \forall j \notin \mathbb{V}_m \, , \forall k \in \mathbb{V}_m$,

where $\bar{\boldsymbol{\sigma}}_j \triangleq \mathbb{E}_{p(\mathbf{x})}[\boldsymbol{\sigma}_j]$.

We know from Definition 1 that the variance representation is always close to 1 when the variables are passive and to 0 when they are active. Thus, variables that are passive (resp. active) over the whole dataset will be almost constant with an expected value close to 1 (resp. 0), as stated in (i) and (ii) of Proposition 4.

The variance representations of mixed variables will alternate between 1 and 0, depending on whether they are passive or active for the considered data examples. Thus, as described in (iii) of Proposition 4, they will vary more than active and passive variables.

Theorem 5 (Polarised regime of z over $\boldsymbol{X}$) *When a VAE learns in a polarised regime, its sampled representation $\mathbf{z} \triangleq \boldsymbol{\mu} + \boldsymbol{\epsilon}\boldsymbol{\sigma}^{1/2}$ is composed of a set of passive, active and mixed variables $\mathbb{V}_p \cup \mathbb{V}_a \cup \mathbb{V}_m$ such that, over $\boldsymbol{X}$:*

(i) $p(\mathbf{z}_j) \approx p(\boldsymbol{\epsilon}_j) \quad \forall \, j \in \mathbb{V}_p$,

(ii) $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j) \quad \forall \, j \in \mathbb{V}_a$,

(iii) $p(\mathbf{z}_j) = c \, p(\boldsymbol{\epsilon}_j) + (1-c) \, p(\boldsymbol{\mu}_j) \quad \forall \, j \in \mathbb{V}_m$,

where $0 < c < 1$.

The proof can be found in Appendix A.2.
Given that active variables are the only type of variables with approximately the same distributions in mean and sampled representations, Corollary 6 immediately follows.

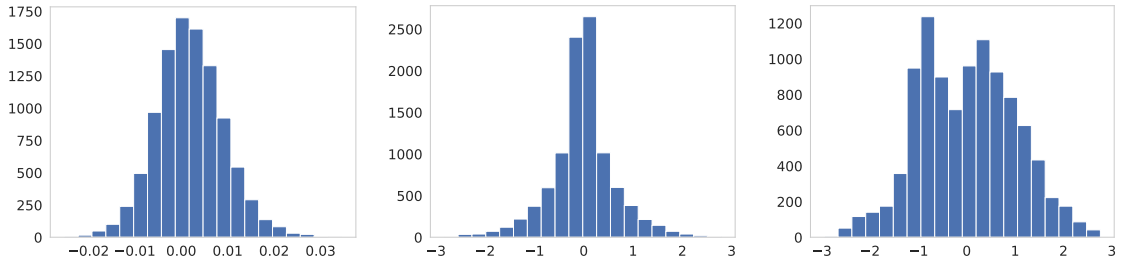
Corollary 6 *Any discrepancies between the mean and sampled representations can only come from the mixed and passive variables.*

3.3 Empirical demonstration of polarised regimes

We will now verify that Theorem 5 holds empirically using a β-VAE trained on the dSprites dataset (Higgins et al., 2017). By comparing the passive variable distribution of mean and sampled representations provided in Figures 4a and 5a, we can see that both have a mean of zero, and that the variance of the variable is close to zero in the mean representation, and to one in the sampled representation, consistently with statement (i) of Theorem 5 and Proposition 3. As described in statement (ii) of Theorem 5, the active variable of the mean representation observed in Figure 4c follows a similar distribution as its sampled counterpart in Figure 5c. Figures 4b and 5b show that mixed variables can also be identified. Their variance in the mean representation is larger than for

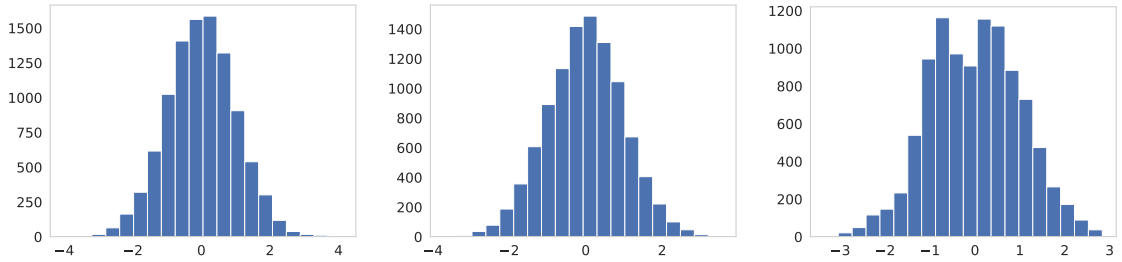
passive variables and increases in the sampled representation. Moreover, we can see a sharp peak around zero in the mean representation, which is smoothed out in the sampled one. This likely corresponds to the passive component of the mixed variables, which is close to zero with very low variance in the mean representation and to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ in the sampled representation. Overall, these observations are consistent with a mixture distribution, as per statement (iii) of Theorem 5.

Note that while the distribution of the passive variables of sampled representations is encouraged to be Gaussian by the KL divergence term of the ELBO, we can see in Figures 4 and 5 that the mixed and active variables are not guaranteed to be Gaussian in mean and sampled representations. Indeed, as they convey some information about the data, their distribution can be arbitrarily different from the prior and may or may not be Gaussian. The resulting increased KL divergence of active variables is then compensated by passive variables that match the prior exactly. Moreover, the distribution of the passive variables of mean representations may also not be Gaussian as it is only optimised to have low variance.



(a) Passive variable (b) Mixed variable (c) Active variable

Figure 4: Empirical distributions of the 2nd, 4th, and 10th latent dimensions of the mean representations of a β-VAE trained on dSprites with $\beta = 8$, which illustrate the typical appearance of passive, mixed, and active variables. The histograms are computed using 10000 input examples.



(a) Passive variable (b) Mixed variable (c) Active variable

Figure 5: Empirical distributions of the 2nd, 4th, and 10th latent dimensions of the sampled representations of a β-VAE trained on dSprites with $\beta = 8$, which illustrate the typical appearance of passive, mixed, and active variables. The histograms are computed using 10000 input examples.

These observations also provide empirical evidence of Corollary 6, that only the active variables are similar in mean and sampled representations. When the regularisation strength increases, the sampled and mean representations will contain more passive and mixed variables, hence *the*

higher the regularisation weight, the greater the difference between the mean and sampled representations. This is the main insight that will allow us to explain the discrepancies between the correlation scores of the mean and sampled representations observed by Locatello et al. (2019b).

4 Assessing the impact of mixed and passive variables on the differences between mean and sampled representations

In Theorem 5 and Section 3 we have seen that while active variables are equivalent in mean and sampled representations, it is not the case for mixed and passive variables. As the number of non-active variables increases with the regularisation strength, we hypothesise that they may be the source of the stronger correlation of mean representations observed by Locatello et al. (2019b).

To verify this hypothesis, we investigate if active variables alone are equivalently correlated in mean and sampled representations and how passive variables would impact the different metrics used. To do so, we divide our experiment into four steps:

(i) Using Proposition 4, identify the type of variable (active, mixed or passive) stored at each index of the representations considered.

(ii) Based on Proposition 3 and Theorem 5, provide a theoretical explanation of the the discrepancies between the averaged Mutual Information (MI) and total correlation (TC) scores of mean and sampled representations for passive variables components.

(iii) Verify that the correlation of mean representations increases with the number of passive and mixed variables. This will allow us to explore how the number of passive variables evolves with stronger regularisation and whether this number can explain the discrepancies in total correlation and averaged mutual information that were reported by Locatello et al. (2019b).

(iv) Empirically verify the results of step (ii) by comparing the impact of passive and mixed variables on averaged mutual information and total correlation scores for mean and sampled representations. By separately comparing the scores of every combination of variable type, we can further attribute the discrepancies to a specific type of variable, or a combination of variables.

These steps are implemented in Sections 4.1, 4.2, 4.3 and 4.4, respectively.

Empirical setup We based our implementation on `disentanglement_lib` (Locatello et al., 2019b), using the same datasets as the authors: dSprites (Higgins et al., 2017), smallNorb (LeCun et al., 2004), cars3D (Reed et al., 2015), and the alternative versions of dSprites (Locatello et al., 2019b) color-dSprites, Scream-dSprites and Noisy-dSprites. We relied on the 9000 pre-trained models released by Locatello et al. (2019b), corresponding to the 5 VAE architectures described in Section 2, over the 6 datasets mentioned above, with 6 different regularisation strengths for each method and 50 seeds per (dataset, method, regularisation) triplet.

4.1 Identifying variable types

Using Proposition 4 and defining a threshold $\alpha \geq 0$, one can easily classify the type of variables stored at each index $j \in d$ based on the variance representations:

- **Passive variable**: $\bar{\boldsymbol{\sigma}}_j = 1 \pm \alpha$ and $Var(\boldsymbol{\sigma}_j) = 0 \pm \alpha$,

- **Active variable**: $\bar{\boldsymbol{\sigma}}_j = 0 \pm \alpha$ and $Var(\boldsymbol{\sigma}_j) = 0 \pm \alpha$,
- **Mixed variable**: any variable not classified as active or passive.

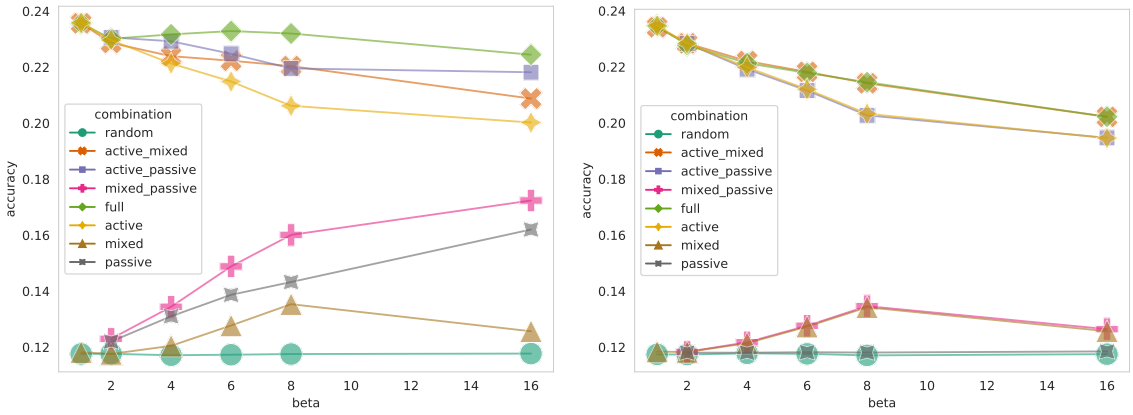
In all our experiments, α is set to 0.1. As stated above, any variable that does not satisfy both conditions required for active or passive variables becomes a mixed variable.

Note that the indexes obtained with this method are used in the same way to identify the variable types of mean and sampled representations. For example, if our procedure determines that index $j = 1$ of the variance representation corresponds to a passive variable, the variable at index $j = 1$ of the mean and sampled representations will be considered as passive.

Sanity check To verify that our thresholds are valid and that the variable types have been mapped correctly, we compare the scores of a classifier trained on the whole representations with those of classifiers trained on every combination of variable types. Similarly to Locatello et al. (2019b), the classification models are logistic regressions (LRs) that predict the labels of the dataset from which the representation was learned.

We trained the LRs for 300000 steps on 10000 data examples for each variable combination and computed the average accuracy over 5000 test examples. The LRs are cross-validated with 10 different regularisation strengths and 5 folds.

If our procedure to identify active, mixed and passive variables is correct, we should expect the following. Active variables should contribute the most to the predictions and give results close to those obtained by the full representation. Mixed variables should contribute less and have a much lower score but together with the active representation, they should provide the same score as the full representation. Passive variables should not contribute at all and ought to provide results close to a random classifier.



(a) Mean representation

(b) Sampled representation

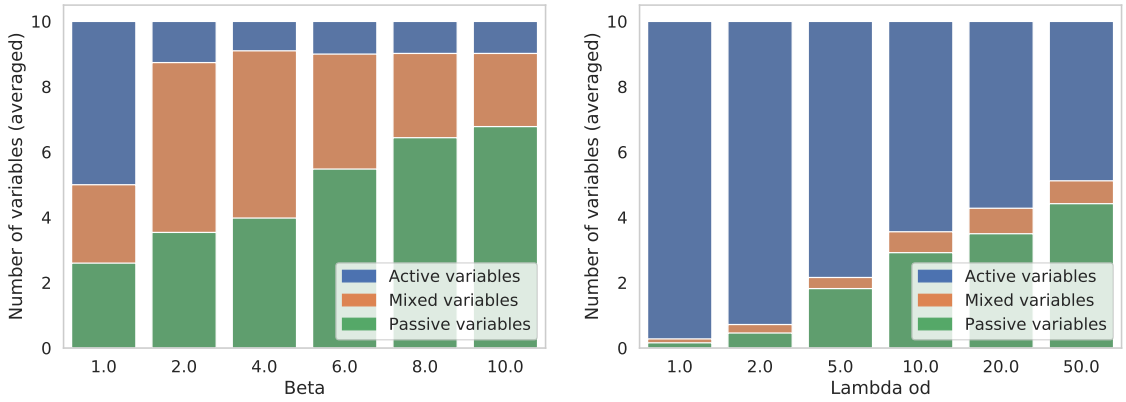
Figure 6: Average test accuracy of a logistic regression trained on the mean and sampled representations learned by a β-VAE trained on dSprites. Each figure shows the results obtained using the full representations and combinations of different variable types. This is also compared to a random classifier picking uniformly from the possible labels.

In Figure 6b, we can see that the results obtained for β-VAE trained on dSprites using sampled representations are exactly as expected. The passive variables gave equivalent results to a random

classifier, while mixed variables performed slightly better, which indicates that we correctly identify them. The score obtained with active and mixed representations is the same as the full score in sampled representations, while active variables alone or with passive variables performed a bit worse, which confirms that mixed and active variables have also been identified correctly.

Interestingly, the results obtained for mean representations in Figure 6a show that passive variables are becoming more informative as the regularisation strength increases. While the score obtained with active variables is still closer to the full representation score, we can see that in opposition to sampled representations, passive and active variables perform better than mixed and active variables. Thus, *despite being close to zero and having a very low variance, passive variables of the mean representation seem to capture some information about the data.* Note that this result does not indicate a problem in the detection of the passive variables: if we had incorrectly identified any active or mixed variables as passive, they would still convey some information in the sampled representation. As seen in Figure 6b, it is not the case here as their performance is equivalent to the performance of a random classifier.

As discussed in Section 3, we can also see in Figure 7 that, as expected, the number of passive and mixed variables increases with the regularisation strength. The number of passive variables at the lowest regularisation strength is generally close to zero for all datasets and all models, except annealed-VAE, whose special case will be discussed in Appendix B.



(a) β-TC VAE trained on smallNorb.

(b) DIP-VAE II trained on dSprites.

Figure 7: Number of passive, mixed and active variables with increased regularisation strength averaged over 50 runs for each regularisation value.

To summarise, given that the behaviour of the passive and active variables observed in Figures 6 and 7 is consistent with the findings of Rolinek et al. (2019) and Dai et al. (2020) regarding the polarised regime, we assume that our method to determine the type of variables is valid, and our thresholds properly set. We can thus proceed to the next steps of our experiment.

4.2 Metrics used to assess mean and sampled representations

To be consistent with the existing literature (Locatello et al., 2019b), we compared the mean and sampled representations using total correlation and averaged mutual information scores. We also

used effective rank (Roy and Vetterli, 2007) as a complementary measure. Those metrics are measuring slightly different things and have different limitations that we will detail below.

4.2.1 TOTAL CORRELATION

The total correlation (Watanabe, 1960) is a measure of the amount of information shared between multiple latent variables. It is measured as the KL divergence between the joint distribution and the product of its marginal distributions. More formally, given a latent representation $\mathbf{r} = \mathbf{r}(\mathbf{x}; \phi)$:

$$TC(\mathbf{r}) = D_{\mathrm{KL}}\Big(p(\mathbf{r}) \;\Big\|\; \prod_{j=1}^{d} p(\mathbf{r}_j)\Big). \tag{8}$$

In their experiment, Locatello et al. (2019b) assumed that $\mathbf{r}$ (either the mean or sampled representation) was multivariate Gaussian with a mean of zero. Thus, given the mean $\bar{\mathbf{r}}$ and covariance $\mathrm{Cov}[\mathbf{r}]$ of $\mathbf{r}$, Equation 8 takes the following form:

$$TC(\mathbf{r}) = D_{\mathrm{KL}}\Big(\mathcal{N}(\bar{\mathbf{r}}, \mathrm{Cov}[\mathbf{r}]) \;\Big\|\; \prod_{j=1}^{d} \mathcal{N}(\bar{\mathbf{r}}_j, \mathrm{Cov}[\mathbf{r}]_{jj})\Big), \tag{9}$$

which simplifies to

$$TC(\mathbf{r}) = \frac{1}{2}\Big(\sum_{j=1}^{d}(\log \mathrm{Cov}[\mathbf{r}]_{jj}) - \log\det(\mathrm{Cov}[\mathbf{r}])\Big). \tag{10}$$

The transition from Equation 9 to Equation 10 can be found in Appendix A.1. From Equation 10, we then obtain Theorem 7, as proved in Appendix A.3.

Theorem 7 (Impact of passive variables on $TC(\mathbf{z})$) *The total correlation of sampled representations is not modified by passive variables.*

Limitations According to Theorem 5, while the passive variables of the sampled representations follow standard Gaussian distributions, other variables of mean and sampled representations, especially mixed variables, may not be Gaussian. Moreover, the distributions of the latent variables may not be jointly Gaussian. As such, TC may provide inaccurate results. For this reason, it is important to use complementary metrics.

4.2.2 AVERAGED MUTUAL INFORMATION

Mutual information is a measure of the information shared by two latent variables (Cover, 1999). Similarly to TC, which is a generalisation of mutual information to the multivariate case, it is measured as the KL divergence between the joint distribution and the product of its marginal distributions. That is, given two latent factors $\mathbf{r}_1$ and $\mathbf{r}_2$:

$$MI(\mathbf{r}_1, \mathbf{r}_2) = D_{\mathrm{KL}}\big(q(\mathbf{r}_1, \mathbf{r}_2) \parallel q(\mathbf{r}_1)q(\mathbf{r}_2)\big). \tag{11}$$

In their experiment, Locatello et al. (2019b) calculated the mutual information over discretised values using:

$$MI(\mathbf{r}_1, \mathbf{r}_2) = \sum_{i=1}^{|\boldsymbol{U}|}\sum_{j=1}^{|\boldsymbol{V}|} \frac{|\boldsymbol{U}_i \cap \boldsymbol{V}_j|}{n} \log \frac{n|\boldsymbol{U}_i \cap \boldsymbol{V}_j|}{|\boldsymbol{U}_i||\boldsymbol{V}_j|}, \tag{12}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are the bins of $\mathbf{r}_1$ and $\mathbf{r}_2$ respectively, n is the number of samples, and $|\cdot|$ denotes the cardinality. They then used the averaged mutual information over all the latent factors:

$$MI_{avg}(\mathbf{r}) = \frac{1}{k^2 - k} \sum_{i=1}^{k} \sum_{j \neq i} MI(\mathbf{r}_i, \mathbf{r}_j), \tag{13}$$

where k is the dimensionality of the latent representation $\mathbf{r}$. Theorem 8 follows, as proved in Appendix A.4.

Theorem 8 (Impact of passive variables on $MI_{avg}(\mathbf{z})$) *The averaged mutual information of sampled representations decreases with the number of passive variables.*

Limitations As it is using discretised values, averaged MI does not have the downside of TC regarding the Gaussian assumption. However, because MI is averaged, it may diminish the strong relationships between two variables if the other MI scores are close to zero, as seen in Theorem 8.

4.2.3 EFFECTIVE RANK

The effective rank of a matrix (Roy and Vetterli, 2007) is a real-valued generalisation of the integer-valued rank of a matrix. More formally, let us consider a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Its singular value decomposition is $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ and $\boldsymbol{V} \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is a rectangular diagonal matrix containing the l singular values $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_l \geqslant 0$ where $l = \min(m, n)$. The singular value distribution is given by:

$$p_k = \frac{s_k}{\sum_{i=1}^{l} s_i} \text{ for } k = 1, 2, \cdots, l.$$

The effective rank is then defined as:

$$\text{erank}(\boldsymbol{A}) = \exp\big(\text{H}(p_1, p_2, \cdots, p_l)\big),$$

where $\text{H}(p_1, p_2, \cdots, p_l)$ is the spectral entropy (Campbell, 1960; Yang et al., 2005):

$$\text{H}(p_1, p_2, \cdots, p_l) = -\sum_{k=1}^{l} p_k \log p_k.$$

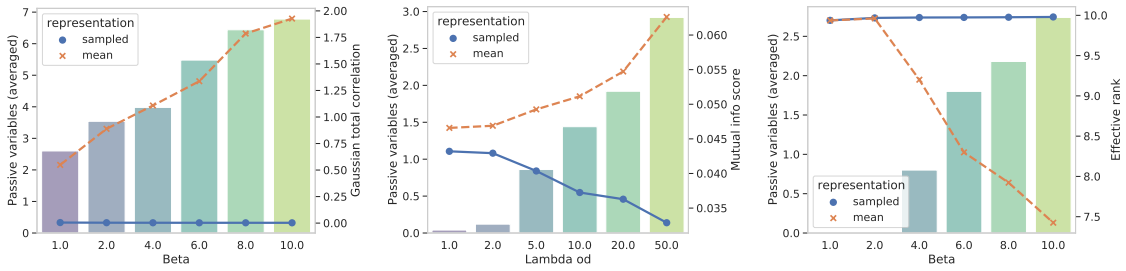
The effective rank is generally more informative than the rank as it can take all possible values in the interval $[1, l]$, whereas the rank is limited to integer values in the set $\{1, 2, ..., l\}$. Consider, for example, a Gaussian distribution of dimension two whose variables are highly correlated. The first singular value will dominate while the second will be close to zero. As both singular values are higher than zero, the matrix rank will be two. However, because the second singular value is very low, its effective rank will be only slightly above one. Overall, the effective rank tells us more about the data than the integer-valued rank.

Limitations While the effective rank does not have the weaknesses of TC regarding the Gaussian assumption nor averaged MI regarding the dilution of a strong relationship when dimensionality is high, it would not make sense to measure it separately for each type of variables as the significance of the dimensions would be relative to the subset considered. Thus, effective rank is only included in Section 4.3, where we use it to analyse all variables jointly.

4.3 Preliminary observations

As we have established in Section 3, the active variables of mean and sampled representations should be similarly correlated. Hence, while Locatello et al. (2019b) concluded that uncorrelated sampled representations did not guarantee uncorrelated mean representations, we argue that active variables, which encode the most information, should have similar correlation in mean and sampled representations. We thus suggest that the increased correlation of mean representations may be due to mixed and passive variables. If the passive variables are, as hypothesised, responsible for the higher correlation of mean representations, one should expect the effective rank to be close to the total number of latent variables minus the passive ones.

Our preliminary observations are consistent with this hypothesis. Specifically, in sampled representations, Figures 8a and 8b show that TC does not change when we increased the number of passive variables and MI decreases, as described in Theorems 7 and 8. Moreover, Figure 8a shows that the discrepancies between the total correlation scores of the mean and sampled representations are increasing with the number of passive variables, and we can observe the same trend for mutual information in Figure 8b. The effective rank of mean representations is, as shown in Figure 8c, close to the total number of latent variables minus the passive ones. One can also notice that in sampled representations, the effective rank is close to the total number of variables as the passive variables are replaced by uncorrelated samples from $\mathcal{N}(0, 1)$, and those uncorrelated variables cannot reduce the effective rank any more. Interestingly, in addition to showing an increased correlation of passive variables, the effective rank is also providing further confirmation that passive variables are correctly identified in Proposition 4 and Section 4.1.



(a) Total correlation (b) Mutual information (c) Effective rank

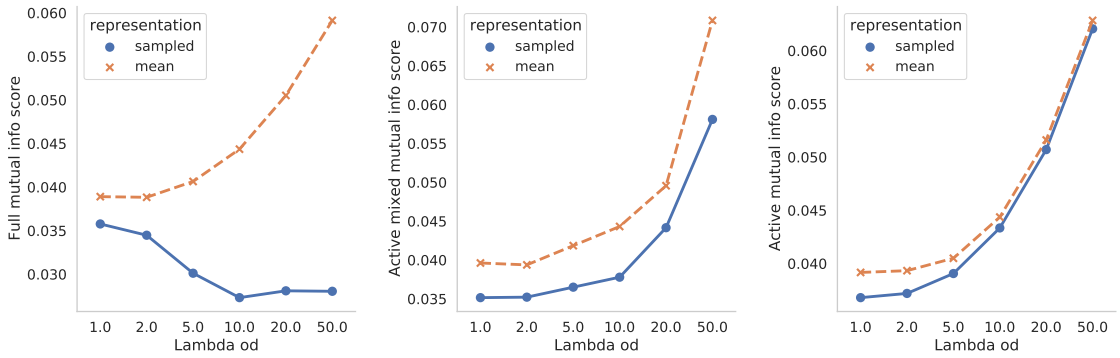
Figure 8: Relationship between the number of passive variables and the total correlation, mutual information, and effective rank scores of mean and sampled representations. Figures (a), (b) and (c) use the representations learned by β-TC VAE trained on smallNorb, DIP-VAE II trained on Cars3D, and β-TC VAE trained on colour dSprites, respectively. Lines indicate the metric scores of the representation and the bars the average number of passive variables.

4.4 Impact of passive and mixed variables on averaged mutual information and total correlation

To further validate our hypothesis that only mixed and passive variables are responsible for the increased correlation of mean representations, we will compare TC and averaged MI scores with and without mixed and passive variables. Thus, we will be able to determine whether we can have an increased correlation between any dimensions of the mean representations, as initially inferred

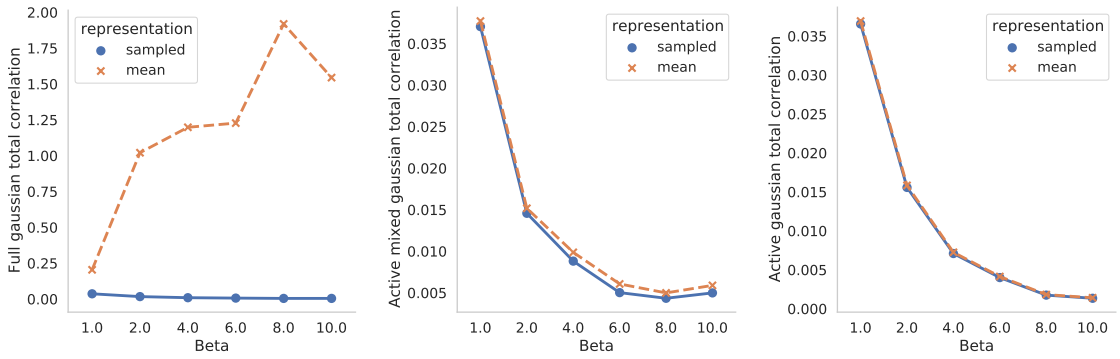
by Locatello et al. (2019b), or only between a specific subset corresponding to mixed and passive variables.

As we have mapped each index of the mean and sampled representation to a specific variable type in Section 4.1, we can now assess the impact of mixed and passive variables on TC and averaged MI by comparing the scores with and without them. We first calculate TC and averaged MI using the full representation and compare these scores with those obtained without passive variables and with only active variables.



(a) Full representation (b) Mixed and active variables (c) Active variables

Figure 9: Comparison of the averaged mutual information scores of the mean and sampled representations of DIP-VAE II trained on dSprites. In Figure (a), the number of passive variables increases with λ leading to lower averaged MI scores for sampled representations. The mean and sampled representation scores become more similar once passive variables are removed in Figures (b) and (c).



(a) Full representation (b) Mixed and active variables (c) Active variables

Figure 10: Comparison of the total correlation scores of the mean and sampled representations of β-TC VAE trained on noisy dSprites. In Figure (a), the number of passive variables increases with β leading to lower TC scores for sampled representations. The mean and sampled representation scores become more similar once passive variables are removed in Figures (b) and (c).

Mutual information and total correlation of mean representations In Figures 9a and 10a, we can see that the mixed and passive variables are raising the TC and averaged MI scores of the mean representations (the dashed orange curves). Indeed, when these variables are removed, the TC and averaged MI scores of the mean and sampled representations are quite similar, as observed in Figures 9b and 9c, and Figures 10b and 10c. While the mixed variables impact the score to a small extent, the passive variables lead to a dramatic score increase, especially for TC. These observations, thus, show that *active variables are as disentangled in mean as in sampled representations, and passive variables of the mean representations should have strong correlations with other variables.* We, therefore, examine the correlation of passive variables in Section 5 to better understand with which type of variable they are correlated and how this correlation emerges. Even before we present this analysis, we should recall that passive variables of the mean representations performed better on downstream tasks than their sampled counterpart, as observed in Section 4.1. Thus, to convey useful information, the passive variables have to be correlated with known informative variables. As a result, one should expect the passive variables in the mean representations to be correlated with active ones. This is the subject of our investigation in Section 5.

5 Where does the correlation of passive variables in the mean representation come from?

As the passive variables of mean representations lead to higher TC and averaged MI scores, they should exhibit some correlation with other variables. This can be seen empirically in Figure 11 where the correlation between each latent variable of a β-VAE trained with $\beta = 16$ on colour dSprites shows that, indeed, the passive variables tend to have strong correlation scores with one or more variables, which is generally not the case for other variable types. This result should not be surprising as VAEs optimise passive variables to be close to zero with low variance, but not to be uncorrelated. However, one can ponder on the origin of these correlations. Are they present from the beginning of the learning process and lead these variables to become passive, or are the correlations the consequence of the variables being passive?

To gain some insights into this question, we trained a β-VAE with $\beta = 8$ on dSprites for 300K steps and saved a snapshot of the model parameters every 1000 steps and observed the evolution of the latent representations. Using the same technique as in Section 4, after the model has been trained (i.e. after 300K training steps), we determined that the variables 1, 4, and 6 were passive, the variables 0, 3 and 8 mixed, and the remaining ones active. In Figure 12, we can see that the correlation score of the passive variables is more often above 0.2 than the scores of the active variables across the 300 snapshots recorded during the training process. This highlights the important correlation scores of the passive variables in most of the training steps. Figure 13 shows that the correlation between passive variables and other variables varies significantly during the training process and can be relatively high, while the correlation scores of active variables remain very low. These high correlations are consistent with the observations of Section 4.1 where the logistic regression had better accuracy with passive variables than with mixed ones. While a more in-depth study of the learning dynamics of VAEs would be needed to provide a complete explanation of this phenomenon, the frequent changes of the correlation scores makes it likely to be an inherent property of the neural network training process.

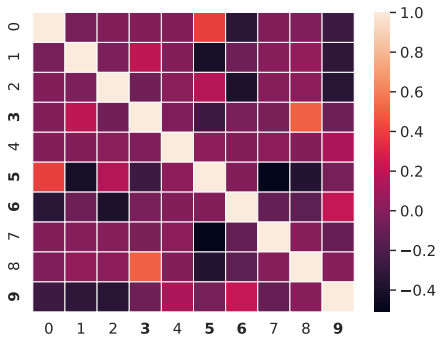


Figure 11: Correlation of the passive variables of β-VAE trained on colour dSprites with $\beta = 16$ with other variables. The passive variables are at indexes 3, 5, 6, and 9. Several correlations of the passive variables clearly stand out.

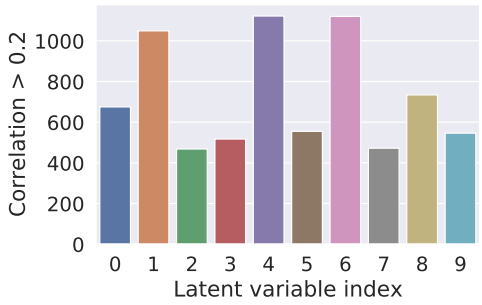


Figure 12: Number of times where the absolute value of the correlation of each latent variable with another variable in the mean representation was above 0.2 in the 300 snapshots recorded during the training of a β-VAE with $\beta = 8$ on dSprites. The passive variables are at indexes 1, 4, and 6, and are the most often correlated variables.

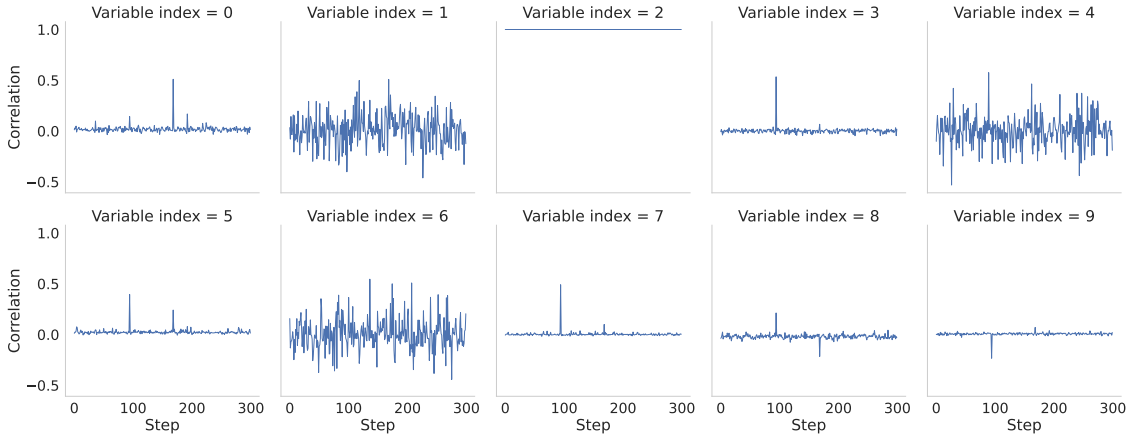


Figure 13: The correlation scores of the active variable at index 2 of the mean representation with all the other variables during the 300K training steps of a β-VAE with $\beta = 8$ trained on dSprites. We can see an increased correlation with all the passive variables (indexes 1, 4, and 6).

6 Conclusion

In their study, Locatello et al. (2019b) have reported that mean representations seemed to be more correlated than sampled ones in a large number of experiments. They concluded that enforcing uncorrelated sampled representations may not be sufficient to obtain uncorrelated mean representations. By extending the definition of the polarised regime to mean representations, we have shown that while aiming to optimise uncorrelated sampled representations is not sufficient to guarantee completely uncorrelated mean representations, it is sufficient to obtain uncorrelated *active variables*

in mean representations. We thus hypothesised that the increasing discrepancies between the two representations should only be attributed to a subset of variables: the passive ones. This hypothesis was consistent across different levels of regularisation and has been further confirmed by empirical observations showing an increased correlation of passive variables. By considering the latent representations over the whole dataset, we have also introduced mixed variables, a type of variable that can either be active or passive depending on the input example provided. Our empirical results confirmed the existence and importance of such variables.

Should we use mean representations for downstream tasks? One of the concerns that was raised by the findings of Locatello et al. (2019b) was that mean representations, which are generally used for downstream tasks, would not benefit from the disentanglement that was exhibited by sampled representations. However, we showed that active variables, the relevant part of the representations, are quite similar in both representations. Thus, we can expect mean representations to be as useful as sampled ones for downstream tasks. We established that the passive variables of mean representations are near zero with low variance and seem to have an arbitrary high correlation with other variables. Thus, one may want to remove them, especially when feeding the mean representations into algorithms that are sensitive to near-zero or highly correlated features.

Generalisation of our results to other types of VAEs Because our paper explains the reason of the discrepancies between the mean and sampled representations observed by Locatello et al. (2019b), we chose to remain as close as possible to their experimental protocol, and used the same models to obtain consistent results. However, the authors proved that unsupervised disentanglement learning was not possible without further inductive biases. Thus one should keep in mind that disentangled representations obtained from the models used in this paper can only be selected post-hoc based on disentanglement metrics, which may not always be practical for downstream task applications. Despite this, as described in Section 3, our explanation mainly relies on the polarised regime, which has been shown to occur in any well-behaved VAE as long as their prior and posterior distributions are Gaussian with diagonal covariances (Dai and Wipf, 2018; Rolinek et al., 2019; Dai et al., 2020; Bonheme and Grzes, 2023). This makes our findings to readily applicable to identifiable VAEs which can directly provide disentangled representations (Khemakhem et al., 2020; Mita et al., 2021).

Other applications of this study While our main focus was to explain the higher correlation observed in mean representations by Locatello et al. (2019b), our new definitions of the polarised regime may also be useful to monitor the number of passive variables and prevent posterior collapse due to over-regularisation (Lucas et al., 2019b; Dai et al., 2020).

Future work The surprising behaviour of annealed VAEs discussed in Appendix B and the correlation of passive variables reported in Section 5 show that we could gain a deeper understanding of the representations learned by VAEs by studying their learning dynamics more in depth. While this was not the topic of this paper, we plan to address it in our future work. Indeed, we believe that this could give more insight into how VAEs learn and how we can improve the learned representations.

Acknowledgments and Disclosure of Funding

We thank our action editor and reviewers for their detailed and constructive feedback, which helped us improve the quality of the present paper.

Appendix A. Derivations and proofs

A.1 Details of the simplification of Equation 9 to Equation 10

Equation 9 has the following form

$$TC(\mathbf{r}) = D_{\mathrm{KL}}\left(\mathcal{N}(\bar{\mathbf{r}}, \mathrm{Cov}[\mathbf{r}]) \;\middle\|\; \prod_{j=1}^{d} \mathcal{N}(\bar{\mathbf{r}}_j, \mathrm{Cov}[\mathbf{r}]_{jj})\right).$$

We assume that all the distribution have a mean of zero. Thus,

$$TC(\mathbf{r}) = D_{\mathrm{KL}}\left(\mathcal{N}(\mathbf{0}, \mathrm{Cov}[\mathbf{r}]) \;\middle\|\; \prod_{j=1}^{d} \mathcal{N}(0, \mathrm{Cov}[\mathbf{r}]_{jj})\right).$$

Moreover, the second term of the KL divergence represents the case where $\mathbf{r}$ is composed of j independent and normally distributed random variables. We can thus reexpress it as a multivariate Gaussian distribution with diagonal covariance,

$$TC(\mathbf{r}) = D_{\mathrm{KL}}\Big(\mathcal{N}(\mathbf{0}, \mathrm{Cov}[\mathbf{r}]) \;\big\|\; \mathcal{N}(\mathbf{0}, diag[\mathrm{Var}(\mathbf{r})])\Big).$$

Now, let us define $\mathbf{\Sigma} \triangleq \mathrm{Cov}[\mathbf{r}]$ and $\bar{\mathbf{\Sigma}} \triangleq diag[\mathrm{Var}(\mathbf{r})]$. Using the analytical solution of the KL divergence between two multivariate Gaussian distributions, we have:

$$TC(\mathbf{r}) = \frac{1}{2}\left(\log\frac{\det(\bar{\mathbf{\Sigma}})}{\det(\mathbf{\Sigma})} + \mathrm{Tr}(\bar{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - n\right).$$

Because $\mathbf{\Sigma}$ and $\bar{\mathbf{\Sigma}}$ have the same diagonal values, $\bar{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma} = \boldsymbol{I}$ and $\mathrm{Tr}(\bar{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) = n$. Thus,

$$TC(\mathbf{r}) = \frac{1}{2}\left(\log\det(\bar{\mathbf{\Sigma}}) - \log\det(\mathbf{\Sigma})\right).$$

As $\bar{\mathbf{\Sigma}}$ is diagonal, this further simplifies to

$$\begin{aligned}
TC(\mathbf{r}) &= \frac{1}{2}\left(\log\prod_{j=1}^{d}\bar{\mathbf{\Sigma}}_{jj} - \log\det(\mathbf{\Sigma})\right),\\
&= \frac{1}{2}\left(\sum_{j=1}^{d}\log\bar{\mathbf{\Sigma}}_{jj} - \log\det(\mathbf{\Sigma})\right),\\
&= \frac{1}{2}\left(\sum_{j=1}^{d}\log\mathrm{Var}(\mathbf{r}_j) - \log\det(\mathrm{Cov}[\mathbf{r}])\right),
\end{aligned}$$

as expected.

A.2 Proof of Theorem 5

Proof Let $\mathbf{z}_j$ be the sampled representation variable at index j. There are three cases:

(i) If $j \in \mathbb{V}_p$, then, from statement (i) of Proposition 3, $\boldsymbol{\mu}_j$ is almost constant with a value close to 0. Thus, $\mathbf{z}_j \approx \boldsymbol{\epsilon}_j \boldsymbol{\sigma}_j^{1/2}$. Using statement (i) of Proposition 4, we also know that $\boldsymbol{\sigma}_j$ is almost constant with a value close to 1, thus $\boldsymbol{\sigma}_j^{1/2} \approx 1$ and $\mathbf{z}_j \approx \boldsymbol{\epsilon}_j$. It follows that $p(\mathbf{z}_j) \approx p(\boldsymbol{\epsilon}_j)$, which proves statement (i).

(ii) If $j \in \mathbb{V}_a$, then, from statement (ii) of Proposition 4, $\boldsymbol{\sigma}_j$ is almost constant with a value close to 0. Thus, $\mathbf{z}_j \approx \boldsymbol{\mu}_j$. It follows that $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j)$, which proves statement (ii).

(iii) If $j \in \mathbb{V}_m$, then from statement (iii) of Definition 2 we know that $\mathbf{z}_j$ is composed of a subset of active components and a subset of passive components. Thus, $\mathbf{z}_j$ is distributed according to a mixture distribution. Using step (i) and (ii) of the proof, we know that $p(\mathbf{z}_j) \approx p(\boldsymbol{\epsilon}_j)$ for passive variables and $p(\mathbf{z}_j) \approx p(\boldsymbol{\mu}_j)$ for active variables. It follows that for mixed variables $p(\mathbf{z}_j) = c\, p(\boldsymbol{\epsilon}_j) + (1-c)\, p(\boldsymbol{\mu}_j)$ where $0 < c < 1$. This concludes the proof.

■

A.3 Proof of Theorem 7

Proof Let us consider a sampled representation $\mathbf{z}$ with n latent variables having a covariance matrix $\mathrm{Cov}[\mathbf{z}] \in \mathbb{R}^{n \times n}$. Now, let us create a second sampled representation $\hat{\mathbf{z}}$ by concatenating the latent variables of $\mathbf{z}$ with m passive variables. The resulting covariance matrix $\mathrm{Cov}[\hat{\mathbf{z}}]$ can be partitioned as

$$\mathrm{Cov}[\hat{\mathbf{z}}] = \begin{bmatrix} \mathrm{Cov}[\mathbf{z}] & \mathbf{0}_{n,m} \\ \mathbf{0}_{m,n} & \boldsymbol{I}_{m,m} \end{bmatrix}.$$

From this, we can immediately see that

$$\begin{aligned} \sum_{i=1}^{n+m} (\log \mathrm{Cov}[\hat{\mathbf{z}}]_{ii}) &= \sum_{i=1}^{n} (\log \mathrm{Cov}[\mathbf{z}]_{ii}) + \sum_{i=1}^{m} (\log \boldsymbol{I}_{ii}) \\ &= \sum_{i=1}^{n} (\log \mathrm{Cov}[\mathbf{z}]_{ii}). \end{aligned} \tag{14}$$

Moreover, as $\mathrm{Cov}[\mathbf{z}]$ is invertible, using Schur's identity (Brualdi and Schneider, 1983):

$$\begin{aligned} \det(\mathrm{Cov}[\hat{\mathbf{z}}]) &= \det(\mathrm{Cov}[\mathbf{z}]) \det(\boldsymbol{I} - \mathbf{0}_{m,n} \mathrm{Cov}[\mathbf{z}]^{-1} \mathbf{0}_{n,m}) \\ &= \det(\mathrm{Cov}[\mathbf{z}]) \det(\boldsymbol{I}) \\ &= \det(\mathrm{Cov}[\mathbf{z}]). \end{aligned}$$

Thus,

$$\log \det(\mathrm{Cov}[\hat{\mathbf{z}}]) = \log \det(\mathrm{Cov}[\mathbf{z}]). \tag{15}$$

Recall from Equation 10 that

$$TC(\hat{\mathbf{z}}) = \frac{1}{2} \left(\sum_{i=1}^{n+m} (\log \mathrm{Cov}[\hat{\mathbf{z}}]_{ii}) - \log \det(\mathrm{Cov}[\hat{\mathbf{z}}]) \right). \tag{16}$$

Using the result of Equation 14, we can replace the first term of Equation 16, so that

$$TC(\hat{\mathbf{z}}) = \frac{1}{2}\left(\sum_{i=1}^{n}(\log \mathrm{Cov}[\mathbf{z}]_{ii}) - \log\det(\mathrm{Cov}[\hat{\mathbf{z}}])\right). \tag{17}$$

Finally, we can replace the second term of Equation 17 using Equation 15 to obtain

$$\begin{aligned} TC(\hat{\mathbf{z}}) &= \frac{1}{2}\left(\sum_{i=1}^{n}(\log \mathrm{Cov}[\mathbf{z}]_{ii}) - \log\det(\mathrm{Cov}[\mathbf{z}])\right), \\ &= TC(\mathbf{z}), \end{aligned}$$

as required. ■

A.4 Proof of Theorem 8

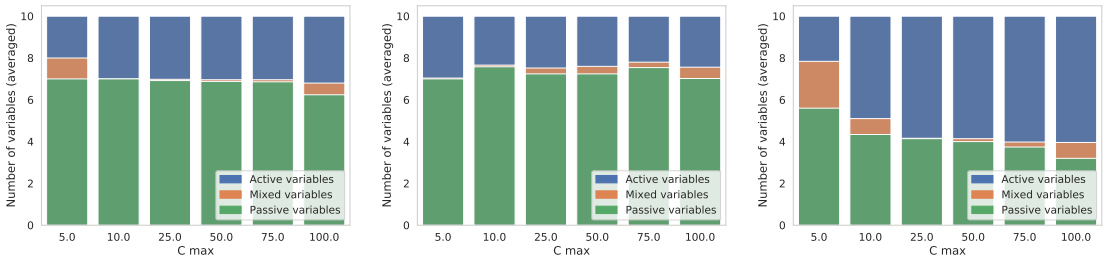
Proof Let us consider a sampled representation $\mathbf{z} \in \mathbb{R}^2$ composed of two active variables $\mathbf{z}_1$ and $\mathbf{z}_2$ such that $MI(\mathbf{z}_1, \mathbf{z}_2) = c$ with $c > 0$. As MI is symmetric, we have $MI_{avg}(\mathbf{z}) = \frac{1}{2}2c = c$.
Now, let us consider a sampled representation $\hat{\mathbf{z}} \in \mathbb{R}^n$ composed of the two actives variables of $\mathbf{z}$ and $n-2$ additional passive variables $\{\mathbf{z}_j\}_{j=3}^{n}$.
Because passive variables do not contain any information about the input, the mutual information between an active variable i and a passive variable j will be zero (i.e., $MI(\mathbf{z}_i, \mathbf{z}_j) = 0$). Passive variables are also independent and normally distributed, hence the mutual information between two different passive variables i and j will also be zero. We thus have

$$\begin{aligned} MI_{avg}(\hat{\mathbf{z}}) &= \frac{1}{n^2-n}\left(\sum_{i\in\mathbb{V}_a}\sum_{\substack{j\in\mathbb{V}_a \\ i\neq j}} MI(\mathbf{z}_i,\mathbf{z}_j) + 2\sum_{i\in\mathbb{V}_a}\sum_{j\in\mathbb{V}_p} MI(\mathbf{z}_i,\mathbf{z}_j) + \sum_{i\in\mathbb{V}_p}\sum_{\substack{j\in\mathbb{V}_p \\ i\neq j}} MI(\mathbf{z}_i,\mathbf{z}_j)\right), \\ &= \frac{1}{n^2-n}2c < MI_{avg}(\mathbf{z}), \end{aligned}$$

as required. ■

Appendix B. The curious case of Annealed VAE

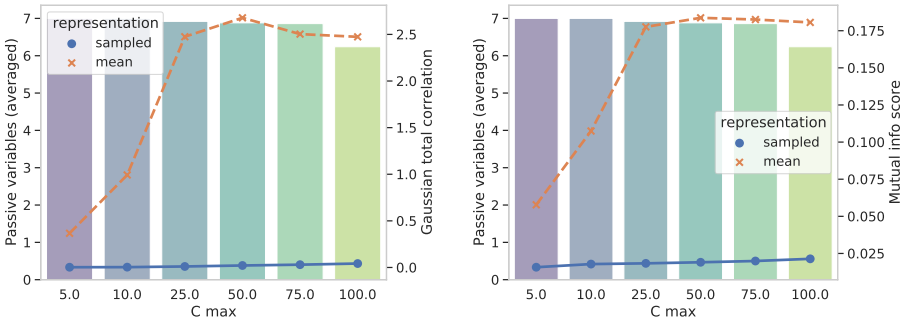
In opposition to the other models that we studied in this paper, annealed VAE surprisingly exhibits a high number of passive variables regardless of the regularisation strength on most datasets, which can be seen in Figure 14. Note that in contrast to the remaining architectures that we study, a higher value of the hyperparameter C means that the regularisation strength decreases, whereas higher β in β-VAEs implies stronger regularisation. As Burgess et al. (2018) originally argued that a higher channel capacity, C, should help the model to learn more latent factors as the training progresses, one would assume that the number of active variables should increase with a higher value of C, but it is generally not the case in Figure 14.



(a) DSprites (b) Scream dSprites (c) Cars3D

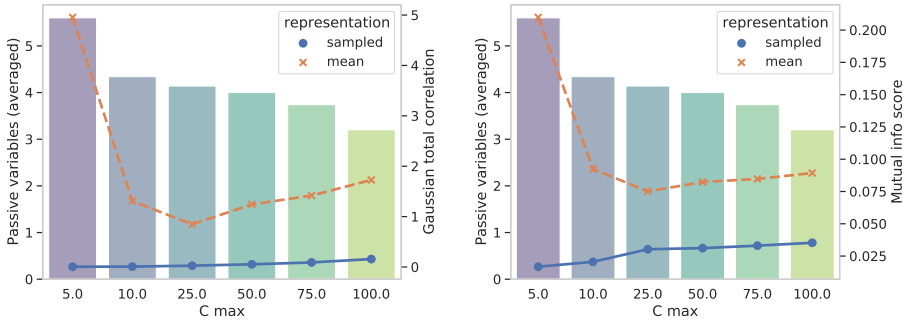
Figure 14: Number of passive, mixed and active variables of annealed VAE trained on dSprites, scream dSprites and cars3D with decreased regularisation strength. The results are averaged over 50 runs for each regularisation value.

Given the near constant number of passive variables observed across all the bars in Figure 14, one could expect similarly constant TC and averaged MI scores. However, we can see in Figures 15 and 16 that the TC and averaged MI generally decrease with higher regularisation strengths. Moreover, Figure 18 shows that removing the passive variables effectively reduces the overall TC and averaged MI scores, suggesting that a higher channel capacity, C, (i.e., a lower regularisation strength at the end of training) may encourage the passive variables of mean representations to be more correlated. This is further confirmed in Figure 17, where we can see that the effective rank obtained with a higher channel capacity and less passive variables is close to the one obtained with a lower channel capacity and more passive variables. For example, in cars3D, the effective rank of the mean representation for a channel capacity of 25 and 4 passive variables is the same as the one obtained for a channel capacity of 100 and 3 passive variables. In conclusion, despite their near-constant number of passive variables, annealed VAE's results are consistent with our other findings: the passive variables of the mean representations are still responsible for the higher TC and averaged MI scores, but their correlation seems to increase with the channel capacity, C.



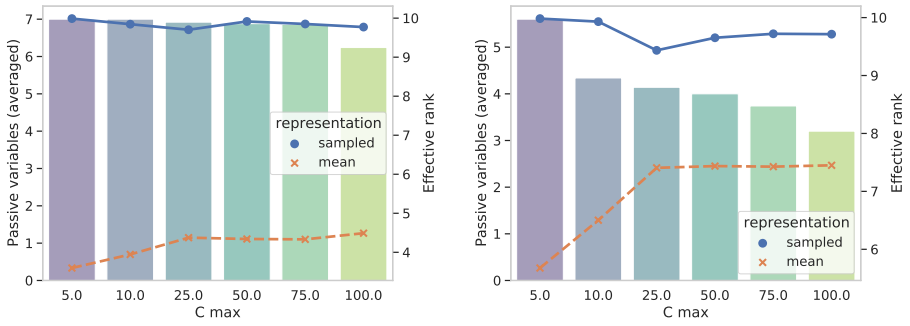
(a) (b)

Figure 15: Comparison of the total correlation and averaged mutual information with the number of passive variables of mean and sampled representations of annealed VAE trained on dSprites. Figure (a) is the total correlation and Figure (b) the averaged mutual information. The lines indicate the metric scores of the two representations, and the bars the average number of passive variables.



(a) (b)

Figure 16: Comparison of the total correlation and averaged mutual information with the number of passive variables of mean and sampled representations of annealed VAE trained on Cars3D. Figure (a) is the total correlation and Figure (b) the averaged mutual information. The lines indicate the metric scores of the two representations, and the bars the average number of passive variables.



(a) (b)

Figure 17: Comparison of the effective rank with the number of passive variables of mean and sampled representations of annealed VAE trained on dSprites and Cars3D. The lines indicate the metric scores of the two representations, and the bars the average number of passive variables.

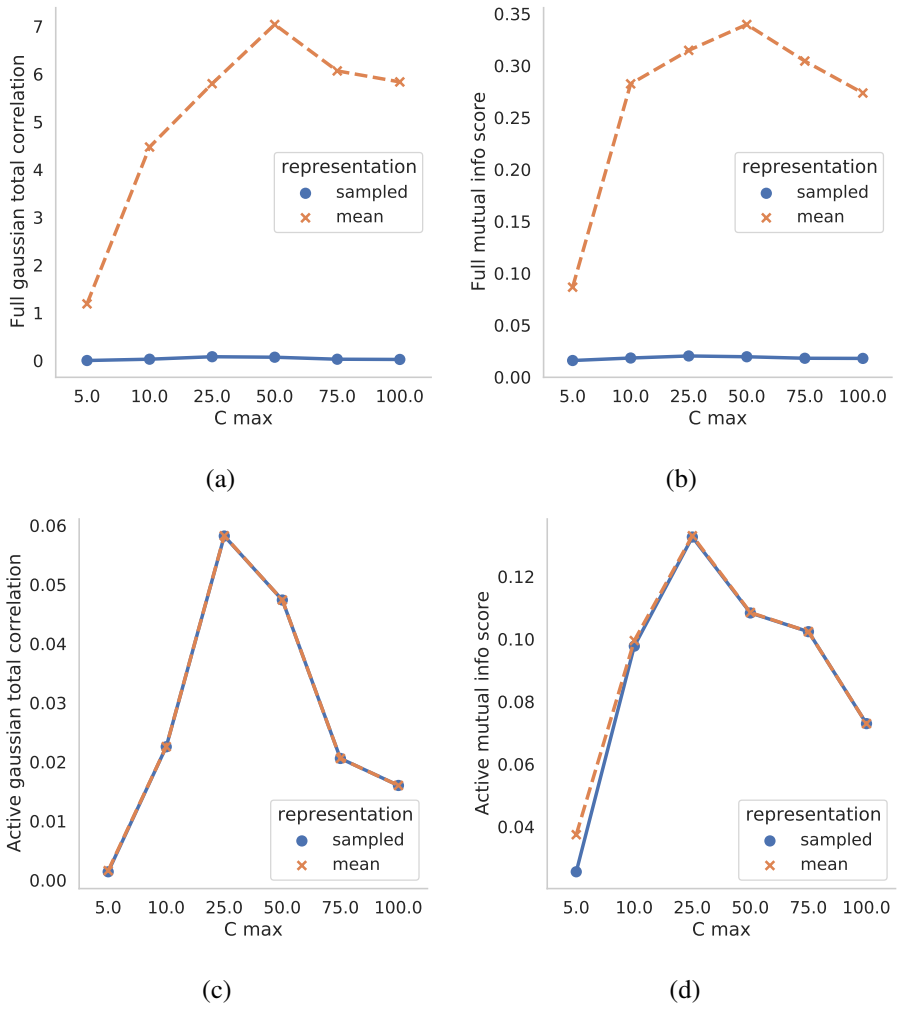


Figure 18: Comparison of the total correlation and averaged mutual information scores of the mean representation of annealed VAE trained on scream dSprites. Figures (a) and (b) are the results of the total correlation and averaged mutual information score using the full representation, and figures (c) and (d) are the results using active variables only.

References

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 50–59, 10–15 Jul 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*, volume 5, 2017.

Aylin Alin. Multicollinearity. *WIREs Computational Statistics*, 2(3):370–374, 2010.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), 2013.

Lisa Bonheme and Marek Grzes. The polarised regime of identifiable variational autoencoders. *ICLR TinyPapers*, 2023. URL `https://openreview.net/forum?id=iSkcAjBqUHU`.

Richard A. Brualdi and Hans Schneider. Determinantal identities: Gauss, Schur, Cauchy, Sylvester, Kronecker, Jacobi, Binet, Laplace, Muir, and Cayley. *Linear Algebra and its Applications*, 52-53, 1983. ISSN 0024-3795. doi: 10.1016/0024-3795(83)80049-4.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding Disentangling in β-VAE. *arXiv e-prints*, 2018.

L. Lorne Campbell. Minimum coefficient rate for stationary random processes. *Information and Control*, 3(4), 1960. ISSN 0019-9958. doi: 10.1016/S0019-9958(60)90949-9.

Jireh Yi-Le Chan, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. Mitigating the multicollinearity problem and its machine learning approach: A review. *Mathematics*, 10(8), 2022.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Ze Cheng, Juncheng Li, Chenxu Wang, Jixuan Gu, Hao Xu, Xinjian Li, and Florian Metze. Revisiting factorizing aggregated posterior in learning disentangled representations. *arXiv e-prints*, 2021.

Thomas M. Cover. *Elements of information theory*. John Wiley & Sons, 1999.

Elliot Creager, David Madras, Joern-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445, 2019.

Bin Dai and David Wipf. Diagnosing and Enhancing VAE Models. In *International Conference on Learning Representations*, volume 6, 2018.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Hidden Talents of the Variational Autoencoder. *arXiv e-prints*, 2017.

Bin Dai, Yu Wang, John Aston, Gang Hua, and David Wipf. Connections with Robust PCA and the Role of Emergent Sparsity in Variational Autoencoder Models. *Journal of Machine Learning Research*, 19(41):1–42, 2018.

Bin Dai, Ziyu Wang, and David Wipf. The Usual Suspects? Reassessing Blame for VAE Posterior Collapse. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Carl Doersch. Tutorial on Variational Autoencoders. *arXiv e-prints*, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Mohamed Shakir, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*, volume 5, 2017.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2020.

Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2018.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, volume 2, 2014.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations. In *International Conference on Learning Representations*, volume 6, 2018.

Yann LeCun, Fu Jie Huang, and Léon Bottou. Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, 2004.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the Fairness of Disentangled Representations. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 2019b.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Understanding Posterior Collapse in Generative Latent Variable Models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop*, 2019a.

James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Don't Blame the ELBO! A linear VAE Perspective on Posterior Collapse. In *Advances in Neural Information Processing Systems*, volume 32, 2019b.

Graziano Mita, Maurizio Filippone, and Pietro Michiardi. An Identifiable Double VAE For Disentangled Representations. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 7769–7779, 18–24 Jul 2021.

Scott Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep Visual Analogy-Making. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: VAEs perform independent mechanism analysis. In *Advances in Neural Information Processing Systems*, volume 36, 2022.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 2015.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Olivier Roy and Martin Vetterli. The Effective Rank: a Measure of Effective Dimensionality. In *2007 15th European Signal Processing Conference*, 2007.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10401–10412, 2021.

Michael Tschannen, Olivier Frederic Bachem, and Mario Lučić. Recent Advances in Autoencoder-Based Representation Learning. In *Third workshop on Bayesian Deep Learning, (NeurIPS 2018)*, 2018.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are Disentangled Representations Helpful for Abstract Visual Reasoning? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Satosi Watanabe. Information Theoretical Analysis of Multivariate Correlation. *IBM Journal of Research and Development*, 4(1), 1960.

Wenye Yang, J.D. Gibson, and Tao He. Coefficient rate and lossy source coding. *IEEE Transactions on Information Theory*, 51(1), 2005. doi: 10.1109/TIT.2004.839531.