Exploring latent regularisation bias in disentangled representation learning

School of Computer Science & Applied Mathematics
University of the Witwatersrand

Neelan Pather
2006601

Supervised by

Prof Benjamin Rosman

Dr Richard Klein

Dr Sonali Parbhoo

October 20, 2022

A research report submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in partial fulfilment of the requirements for the degree of Master of Science in Computer Science

Abstract

The field of **representation learning** involves **learning data representations** (or features) that capture the structure and relationships implicit in data rather than engineering them. **Latent (hidden) variables** are high-level data representations **inferred indirectly** from data via statistical models. Representation learning assumes data is generated via a process characterised by higher-order **generative factors**. When learning **disentangled latent representations**, the aim is individual latent variables sensitive to changes in **only one** generative factor. Disentangled representations are **interpretable** and provide insight into what a model has learnt.

The most popular disentangled representation learning model is Higgins *et al.* [2017]'s β VAE. It is an extension of Kingma and Welling [2013] and Rezende *et al.* [2014]'s VAE, which uses **variational inference** (VI), framing inference as optimisation of the **Evidence Lower BOund** (ELBO) objective. VAEs are also **Bayesian** models, with a prior belief of the latent structure before data is observed, as captured by the **prior latent distribution**. The framework fits a **latent posterior distribution** to data, **regularising** the resultant latent distribution by keeping it **"close"** to the prior.

The closeness between the latent prior and posterior is enforced by **penalising** the ELBO using distance metric called the **Kullback–Leibler (KL) divergence** between these distributions. The latent posterior distribution is thus encouraged not to deviate too much from the latent prior distribution. Notably, the KL divergence is **asymmetrical**, so swapping the argument distributions around results in a different distance metric. Furthermore, the chosen direction of the KL divergence results in different behaviour when the divergence is minimised. This, in turn, encourages different latent posterior solutions.

The **reverse KL divergence** is typically used in VI and encourages **"mode-seeking"** behaviour in the latent space, favouring **under-dispersed** solutions relative to the latent prior. Conversely, the **forward KL divergence** results in **"mean-matching"** behaviour in the latent space, favouring **over-dispersed** solutions relative to the latent prior. The β VAE includes a β hyperparameter factor to the reverse KL divergence in the ELBO, resulting in **hyperparameter constrained reverse KL latent regularisation**. It thus exhibits "mode-seeking" behaviour, favouring under-dispersed solutions relative to the latent prior.

Our study assesses the impact of the KL divergence direction on the resultant latent posterior. Since the goal of disentangled representation learning is latents that capture generative factors in an isotropic manner, we assess the impact multimodal generative spaces have on the resultant posterior solution when the KL divergence direction is varied. To facilitate this investigation, we extend Higgins *et al.* [2017]'s β VAE to include an **additional hyperparameter constrained forward KL** latent regulariser, deriving our model the $\beta\gamma$ **VAE**. Furthermore, we construct a collection of supervised datasets, each with a different number of generative space modes called **mSprites**. Finally, impacts in the study are assessed using information-theoretic disentangling metrics.

When using the reverse KL for latent regularisation, we find that multimodal generative spaces distort the overall information content captured by the learnt representation. This is related to the mode-seeking behaviour of the reverse KL, as evidenced by reduced fit from increased generative space modality. This distortion may be remedied by introducing an additional constrained forward KL for latent regularisation, as done in our $\beta\gamma$ VAE. The impact of multimodal generative spaces on disentangled representation learning, however, is less clear. Our study provides evidence that multimodal generative spaces negatively distort axis alignment between latent and generative dimensions. However, it is not clear that this necessarily hinders disentangled representation learning. Finally, we observe that while our $\beta\gamma$ VAE *can* improve some metrics in disentangled representation learning, eliminating the impact of multimodal generative spaces, it is not consistent when all disentangling metrics are considered, and thus results are less robust. Our $\beta\gamma$ VAE is thus suitable for representation learning with inconsistent evidence that it *may* also be useful in disentangled representation learning.

Our findings are summarised as follows (the related research question addressed is given in brackets):

- **6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β (3.2.1)**
- **6.3 Multimodal generative factors do *not consistently* distort disentangled representation learning, but improve axis alignment. (3.2.2)**
- **6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and eliminates negative impacts of multimodal generative space (3.2.3)**
- **6.5 $\beta\gamma$ VAE does *not consistently* learn better disentangled representations than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces (3.2.4)**

Our key contributions are *mSprites*, a series of supervised datasets designed to investigate the impact of multimodal generative spaces on disentangled representation learning, the $\beta\gamma$ VAE, a model with proven merit for representation learning that uses an additional constrained forward KL for latent regularisation, and empirical evidence validating the findings mentioned above.

Declaration

I, Neelan Pather, hereby declare the contents of this research report to be my own work. This research report is submitted for the degree of Master of Science in Computer Science at the University of the Witwatersrand. This work has not been submitted to any other university, or for any other degree.

Contents

Chapter 1

Introduction

Success in machine learning methods is typically driven by systems that learn something useful for the domain in which they are applied. In a natural image context, this means learning something about image data and the real world captured in these visual signals. Bengio *et al.* [2013] note in their seminal work that performance in machine learning is largely driven by choice of the **data representations** (or features). They argue that in order to make progress in artificial intelligence (AI), a machine learning system must fundamentally understand the world around us. They claim that this is only achievable if the system can learn to **identify underlying explanatory factors** hidden in low-level sensory data. **Representation learning** has become a field of machine learning research in its own right. This is where structures and relationships (or features) present in data are **learnt**, rather than engineered by domain experts. Such features extract information that is useful in downstream tasks such as classification and prediction.

Skrondal and Rabe-Hesketh [2007] describe **latent variables**, introduced by Lazarsfeld [1950], as **hidden** random variables. The properties of these hidden latents are **inferred indirectly** by connecting them to observed data via statistical (machine learning) models. These latents may be thought of as high-level data representations, and in representation learning, the observed data is assumed to be generated via a process characterised by higher-order ground-truth **generative factors**. A model's hidden, latent variables are fitted to data, assuming at least some of these ground truth factors are independent, with interpretable labels.

In **disentangled representation learning** the goal is to learn individual latent units that are sensitive to changes in only **one** generative factor. Individual disentangled latent dimensions are largely invariant to the changes in other factors. This allows for **identifiable** latents, allowing them to be matched to **corresponding generative factors**. Disentangled representations are thus **interpretable**, providing insight into what a model has learnt, a vital requirement for using such models in highly regulated domains such as finance and medicine.

Disentangled representation learning (**"disentangling"**) is synonymous with a **one-to-one relationship between generative and latent factors**[1]. Lake *et al.* [2017] note that disentangled representations could boost the performance of state-of-the-art AI where machines still struggle, but humans excel. Such scenarios include transfer learning and zero-shot inference[2] where reusing learnt representations leads to faster learning and enables reasoning through recombining previously learnt factors.

Kingma and Welling [2013] and Rezende *et al.* [2014]'s **variational auto-encoder (VAE)** learns a latent space via **deep models**. Kingma and Welling [2019] note that VAE's are extensively used in unsupervised representation learning, learning semantically meaningful, statistically independent, disentangled factors of variation from unlabelled data. **Inference networks** are flexible constructions that approximate distributions used during inference. The VAE framework provides a principled approach for the efficient learning of **both** the deep latent variable models and their corresponding inference networks.

VAEs use **variational inference** (VI), allowing for the estimation of an **intractable true posterior** with a **variational posterior**. In VI, inference is framed as optimisation and VAEs are fitted by optimising the **Evidence Lower BOund (ELBO)**. Furthermore, VAEs are **Bayesian** models, with beliefs about latent structure, held before ever observing data, as captured by a **latent prior distribution**. When data is observed, optimising the ELBO shapes the resultant learnt latent space by encouraging the latent posterior to be "close" to a prior latent distribution. This shaping of the latent space is known as **"latent regularisation"**.

The impact of latent regularisation in disentangling is a key focus of our work. In VAEs the notion of "closeness" between the (variational) latent posterior and the latent prior is defined by the (reverse) **Kullback–Leibler (KL) divergence**. The KL divergence is penalised during optimisation of the ELBO, encouraging the metric to be small, keeping the latent prior and posterior close. Importantly, the **KL divergence is asymmetrical**. Swapping around the order of the distributions used as arguments, results in a different distance metric. When minimising these metrics, either direction **biases the resultant posterior solution** in a different manner. Minimising the **reverse KL** has **"mode-seeking"** behaviour, favouring **under-dispersed** solutions, relative to the latent prior. By contrast, minimising the **forward KL** has **"mean-matching"** behaviour, favouring **over-dispersed** solutions, relative to the latent prior.

The most prominent disentangling model, Higgins *et al.* [2017]'s β VAE, uses a β hyperparameter weighted reverse KL between the (variational) latent posterior and the latent prior for latent regularisation. More recent work such as Kim and Mnih [2018]

[1] A one-to-many relationship is usually insufficient to match a latent to a factor.
[2] Inference on a new task without further training.

and Kim *et al.* [2019] have further factorised the forward KL into sub-components in an attempt to improve disentangling. Despite this, disentangling research has largely ignored latent regularisation for **divergences other than the reverse KL** between the (variational) latent posterior and the latent prior. The latent posteriors explored in disentangling literature are thus implicitly biased towards under-dispersion relative to the latent prior, with solutions exhibiting mode-seeking behaviour.

We hypothesise that the under-dispersion bias of the reverse KL results in latent posteriors that **capture local latent correlations**. We argue that the focus of disentangling is to learn a **global isotropic relationships** between latent factors and that the posterior's under-dispersion bias, inherited from the reverse KL latent regularisation, present in most prominent disentangling models, **obscures** this.

Our work explores the **latent regularisation bias induced by the reverse KL in typical disentangled representation learning**. The goal of representation learning is to learn latent spaces that capture the generative space well. With this in mind, we introduce a made-for-purpose collection of datasets ***mSprites*, with varying modalities in the generative space.** Given the mode-seeking behaviour induced by latent regularisation using the reverse KL, these datasets allow us to exacerbate latent bias and investigate the **impact of multimodal generative spaces on the resultant latent posterior**.

We also explore the potential merit of using an **additional constrained forward KL for latent regularisation**. To do this, we introduce the $\beta\gamma$ **VAE**, which uses **both a constrained reverse KL and a constrained forward KL**, measured between the (variational) latent posterior and the latent prior, for latent regularisation.

Our study observes that when using the reverse KL for latent regularisation, multimodal generative spaces largely distort the overall information content captured by the learnt representation. This is related to the mode-seeking behaviour of the reverse KL as evidenced by reduced fit from increased generative space modality in *mSprites*. This distortion may be remedied by introducing an additional constrained forward KL for latent regularisation, as done in our $\beta\gamma$ VAE. Despite this, the impact of multimodal generative spaces on disentangling is less clear. Our work provides evidence that multimodal generative spaces negatively distort alignment between latent and generative dimensions; however, it is not clear that this robustly hinders disentangling. Finally, we observe that while our $\beta\gamma$ VAE can improve certain disentangling metrics and eliminate the impact of multimodal generative spaces, however, this is not consistent across metrics, and thus results are less robust and clear. **Our $\beta\gamma$ VAE is thus suitable for representation learning. However, there is inconsistent evidence that our model is useful in disentangling.**

Our findings are summarised as follows (the related research question they address are given in brackets):

- **6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β (3.2.1)**
- **6.3 Multimodal generative factors does *not consistently* distort disentangling, but improve axis-alignment. (3.2.2)**
- **6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and eliminates negative impacts of multimodal generative space (3.2.3)**
- **6.5 $\beta\gamma$ VAE does *not consistently* disentangle better than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces (3.2.4)**

Our key contributions are *mSprites*, a series of supervised datasets designed to investigate the impact of multimodal generative spaces on disentangled representation learning; $\beta\gamma$ VAE, a model with proven merit for representation learning that uses an additional constrained forward KL for latent regularisation.

We begin by exploring the core background concepts underpinning our work in chapter 2. We then outline our hypotheses and research questions in chapter 3. Having outlined our hypotheses and research questions, in chapter 4 we detail our proposed methods, introducing *mSprites*, our collection of datasets with multimodal generative spaces, and deriving our $\beta\gamma$ VAE. Having outlined our proposed methods. Chapter 5 then outlines our experimental methods. Finally, we elaborate on our key findings in chapter 6, before outlining further research and concluding in chapter 7.

Chapter 2

Background

We now explore the core background concepts underpinning our work. Beginning with section 2.1, we explore variational inference (VI), a machine learning method for approximating probability distributions. We note the use of VI in Bayesian statistics, where inference amounts to estimating a difficult to estimate, potentially **intractable true posterior** with an **approximate variational posterior**. VI frames **inference as an optimisation** problem.

The variational auto-encoder (VAE), a popular Bayesian model that uses VI, is then introduced in section 2.2. The VAE's objective, the Evidence Lower BOund (ELBO), is then derived in section 2.3, and its optimisation (inference) particulars are outlined. Importantly, Bayesian models, like the VAE, assume data variability stems from **unobservable hidden, latent variables**. When the ELBO is optimised, the resultant latent space shape is encouraged to be similar to the latent prior distribution by **latent regularisation**. The fitted approximate (variational) latent posterior is thus encouraged to be "close" to some predefined latent prior distribution.

In a VAE, the notion of closeness is defined by the (backward) Kullback–Leibler (KL) divergence, outlined in section 2.4. Importantly, the KL divergence is asymmetrical, so swapping around the order of the distributions used as arguments results in a different distance metric. The **direction of the divergence used for latent regularisation impacts the constraints applied when fitting the approximate (variational) latent posterior.** This changes the approximate (variational) latent posterior solution. This is explored in section 2.4.3.

We then detail disentangled representation learning (disentangling) in section 2.5. Considering the real-world, ground-truth generative factors that generate the data, disentangling aims to learn latent dimensions that have a **one-to-one relationship** with these generative factor dimensions. We outline a key disentangling model, the β VAE, in section 2.5.1. Finally, we describe a plethora of disentangling assessments, outlining latent traversals in section 2.6, supervised metrics in section 2.7, along with unsupervised metrics in section 2.8.

2.1 Bayesian models and Variational inference (VI)

Bayesian models provide an elegantly simple conceptual framework for learning from data. The variability in data is assumed to stem from an unobservable, hidden latent variable $\mathbf{z}$. Consider the joint model of latents $\mathbf{z}$ and data observations $\mathbf{x}$ parameterised by θ, $p(\mathbf{z}, \mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})$. In a Bayesian model, latent $\mathbf{z}$ is drawn from a prior distribution $p_\theta(\mathbf{z})$, which captures belief about the latent structure before observing data. The latent is then related to the data via likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$. Inference amounts to approximating posterior $p_\theta(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z}) / \int p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z}) d\mathbf{z}$, updating the belief about the latent distribution, after observing data.

Such posteriors are often **computationally expensive** and typically difficult to compute or **intractable**. The approximation of difficult-to-compute distributions is a fundamental problem in modern statistics. **Variational techniques** attempt to address this by approximating a simpler variational latent posterior distribution rather than the true (intractable) latent posterior distribution. Salimans *et al.* [2015] note note that the two most popular variational Bayes methods are **variational inference** (VI) and **Markov Chain Monte Carlo** (MCMC). VI has the advantage of maximising an explicit objective and is typically faster. By contrast, MCMC is non-parametric and asymptotically exact.

VI frames inference as an **optimisation** problem, as opposed to MCMC, which is an **empirical estimation** problem. Ganguly and Earp [2021] note that it is precisely the fact that optimisation is used over empirical estimation that makes VI converge faster than MCMC. In VI, complexity is managed by the **variational family** $\mathcal{Q}$ chosen. The choice of this variational family is thus critical: the distributional family must be complex enough to capture $p_\theta(\mathbf{x} \mid \mathbf{z})$, yet simple enough to allow for efficient optimisation. However, Blei *et al.* [2017] notes that, compared to MCMC, VI has been less rigorously studied, with statistical properties that are less well understood. VI thus seeks fit variational latent posterior $q^*_\phi(\mathbf{z}) \in \mathcal{Q}$, by choosing variational parameters ϕ that minimise the reverse KL divergence between the latent variational posterior and true posterior $D_{\mathbb{KL}}\left(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right)$, as per equation 2.1, below.

$$q^*_\phi(\mathbf{z}) = \underset{q_\phi(\mathbf{z}) \in \mathcal{Q}}{\arg\min} D_{\mathbb{KL}}\left(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z} \mid \mathbf{x})\right) \tag{2.1}$$

Salimans *et al.* [2015] note that like VI, MCMC samples from $q(\mathbf{z} \mid \mathbf{x})$. However, unlike VI, MCMC does not optimise this distribution, rather applying a **stochastic transition operator** $\mathbf{z}_t \sim q\left(\mathbf{z}_t \mid \mathbf{z}_{t-1}, x\right)$ to generate successive samples t from samples at $t-1$. If the transition operator is judiciously chosen, random variable $\mathbf{z}_T$ converges to the exact posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$. However, practically one does not know how long such convergence will take, and thus T is unknown.

Our work focuses on VI and the explicit objective associated with this variational approximation technique. In section 2.2 that follows, we introduce arguably the most popular deep learning model estimated using VI, the variational auto-encoder (VAE).

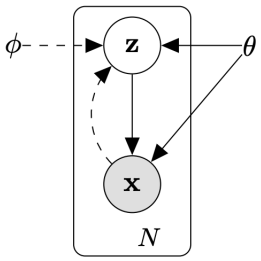


Figure 2.1: **VAE's directed graphical model:** Generative model $p_\theta(\mathbf{z}, \mathbf{x}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})$ (Solid line), with variational approximation $q_\phi(\mathbf{z} \mid \mathbf{x})$ to intractable posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ (Dashed line). Variational parameters ϕ are learned **jointly** with generative model parameters θ. Figure adapted from Kingma and Welling [2013]

2.2 Variational auto-encoders (VAEs)

Kingma and Welling [2013] and Rezende *et al.* [2014] introduced the **variational auto-encoder** (VAE), a powerful **generative latent model** for modelling **complex high-dimensional data** $\mathbf{x}$.

In Principal Component Analysis (PCA), complex variation in data $\mathbf{x}$ may be projected onto a straight line that combines latent factors in a linear fashion. By using the neural networks synonymous with deep learning, complex structures that combine latents factors $\mathbf{z}$ in **highly non-linear manner** are formed that capture far richer data variation.

Kingma and Welling [2019] note that a major division in machine learning is generative versus discriminative modelling. Discriminative models learn a predictor given the observations. By contrast, generative models learn a joint distribution over all variables, **simulating the real world data generative process**; a more general problem. Furthermore, Kingma and Welling [2019] note that generative models are highly intuitive and interpretable and may naturally express real-world causal relations. Bayes rule is needed to turn a generative model into a discriminator, which is computationally expensive, hence the need for VI, as outlined in section 2.1.

In VAEs, VI is used to **approximate** some **intractable** true posterior distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$ via optimisation. During training, model parameters θ and variational parameters ϕ are learnt concurrently via gradient decent, parameterising generative model $p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})$ and variational approximation $q_\phi(\mathbf{z} \mid \mathbf{x})$ to intractable posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$.

VAEs are **directed graphical models**, shown in figure 2.1. In directed graphical models, factor probabilities are a product of conditional probabilities, one for each node in the graph. Bengio *et al.* [2013] note that directed latent factor models, a subset of directed graphical models, separately parameterise the conditional likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$ and the prior $p_\theta(\mathbf{z})$ to construct the joint distribution, $p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})$. Continuous latents $\mathbf{z}$ are present for each i.i.d. datapoint $\mathbf{x}$, with maximum likelihood estimation (MLE) or maximum a posteriori (MAP) inference on global parameters and **variational inference** on the latent variables.

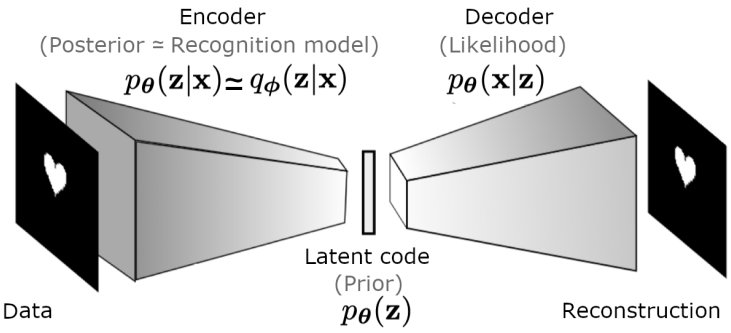


Figure 2.2: **Typical VAE:** Variational auto-encoder (VAE) **encodes** mapping from data $\mathbf{x}$ to latent space $\mathbf{z}$. Grey structures in the figure represent neural network architectures. Figure adapted from Kim and Mnih [2018].

Figure 2.2 shows a typical variational auto-encoder (VAE). The grey structures in the figure represent neural network architectures that allow for complex relationships between the latent factors and the data to be mapped.

In a VAE i.i.d data. $\mathbf{x} = \left\{\mathbf{x}^{(i)}\right\}_{i=1}^{N}$ is assumed to be generated by some unobservable process that maps from continuous latent variables $\mathbf{z}$. Typically $\texttt{Rank}(\mathbf{x}) > \texttt{Rank}(\mathbf{z})$, implying that $\mathbf{z}$ captures a compressed form of the variability observed in $\mathbf{x}$. What differentiates a VAE from a typical auto-encoder is that $\mathbf{z}$ is assumed to be stochastic rather than deterministic, and so the generation process may be generalised beyond the training data $\mathbf{x}$. The generation process is assumed to first sample $\mathbf{z}$ from prior distribution $p_\theta(\mathbf{z})$, before data is generated via likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$[1].

Simplifying assumptions, such as conjugacy, usually made about the marginal or posterior distributions, are **not** made in this approach, allowing the framework to be used when the posterior is **intractable**. It is precisely this assumption that makes the framework so practically useful, as intractable marginal likelihood integrals $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z})p_\theta(\mathbf{x} \mid \mathbf{z})d\mathbf{z}$ are common, even for moderately complicated likelihoods[2].

In VAEs, a **recognition model** $q_\phi(\mathbf{z} \mid \mathbf{x})$ is introduced as a variational approximation to intractable true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$. $q_\phi(\mathbf{z} \mid \mathbf{x})$ is chosen such that the bound and its gradient are computationally tractable or readily estimated by sampling.

Since unobserved latent variables $\mathbf{z}$ encode some data abstraction, the recognition model $q_\phi(\mathbf{z} \mid \mathbf{x})$ may be thought of as an **encoder**. Similarly $p_\theta(\mathbf{z} \mid \mathbf{x})$ may be thought of as a **decoder**. Kingma and Welling [2019] note that the recognition model is the approximate inverse of the generative model according to the computationally expensive Bayes rule. These are shown in figure 2.2.

[1]Prior $p_\theta(\mathbf{z})$ and likelihood $p_\theta(\mathbf{x} \mid \mathbf{z})$ are assumed to come from the same parametric family, with densities differentiable by θ and $\mathbf{z}$ almost everywhere.

[2]When marginal likelihoods are intractable so too are the true posteriors $p_\theta(\mathbf{z} \mid \mathbf{x}) = p_\theta(\mathbf{x} \mid \mathbf{z})p_\theta(\mathbf{z})/p_\theta(\mathbf{x})$.

2.3 The Evidence Lower BOund

A note on Evidence Lower BOund notion

Throughout this work, refer to the Evidence Lower BOund generally as $\mathcal{ELBO}(q)$. This is to emphasise that the expectation is with respect to tractable variational distribution q, drawn from variational family $\mathcal{Q}$ i.e. $q_\phi(\mathbf{z} \mid \mathbf{x}) \in \mathcal{Q}$. Importantly the exact formation of $\mathcal{ELBO}(q)$ differs between the VAE models explored in our work, which differ only by their hyperparameters. We use $\mathcal{L}_{\theta,\phi,\beta}(\mathbf{x}, \mathbf{z})$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta}}(\mathbf{x}, \mathbf{z})$ and $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta\gamma}}(\mathbf{x}, \mathbf{z})$ to refer to the $\mathcal{ELBO}(q)$ of section 2.2's VAE, section 2.5.1's $\boldsymbol{\beta}$ VAE and section 4.1's $\boldsymbol{\beta\gamma}$ VAE, respectively.

From sections 2.1 and 2.2, we know that variational auto-encoders (VAEs) use variational inference (VI), a variational approximation technique that optimises an **explicit objective**. For VAEs, the **Evidence Lower BOund** ($\mathcal{ELBO}(q)$), is precisely this explicit objective, which we derive and detail in this section.

Examining intractable true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ further, we note that by **Bayes rule** it may be split into its components, as in equation 2.2.

$$p_\theta(\mathbf{z} \mid \mathbf{x}) = \frac{p_\theta(\mathbf{z}, \mathbf{x})}{p_\theta(\mathbf{x})}. \tag{2.2}$$

Numerator $p_\theta(\mathbf{z}, \mathbf{x})$ is the **generative model** and is typically efficient to compute. The denominator $p_\theta(\mathbf{x})$, known as the marginal likelihood or **evidence**, however, is estimated via the following integral $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{z}, \mathbf{x}) d\mathbf{z}$, which is **intractable** in many cases. This integral has **exponential cost** in the number of latents to sum up all latent variable combinations.

The (reverse) KL divergence between variational latent posterior $q_\phi(\mathbf{z})$ and true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$, is be defined in equation 2.3.

$$D_{\mathbb{KL}}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z} \mid \mathbf{x})) = \int q_\phi(\mathbf{z}) \log \frac{q_\phi(\mathbf{z})}{p_\theta(\mathbf{z} \mid \mathbf{x})} d\mathbf{x} \tag{2.3}$$

In deriving the $\mathcal{ELBO}(q)$ we begin in equation 2.4a, with the definition of the (reverse) KL from equation 2.3. It should be noted that in this formulation, it is assumed that the approximate density $q_\phi(\mathbf{z})$ is **not** conditioned on the observed variable $\mathbf{x}$.

In equation 2.4b, we separate the numerator and denominator using *logarithm properties*. In equations 2.4c and 2.4d, we then split *one integral of two terms* into *two integrals of one term*. The resultant individual terms each form an *expectation with respect to variation posterior* $q_\phi(\mathbf{z})$. We rewrite these terms as expectations in equation 2.4e.

In equation 2.4f, we apply *Bayes rule* to the second term, before using the *logarithm properties* to split the second term further in equation 2.4g. We reorder the expectation terms in 2.4h. Finally, in equation 2.4i, we note that the second term is *not dependant on variation posterior* $q_\phi(\mathbf{z})$, and so the *expectation is constant with respect to* $q_\phi(\mathbf{z})$, and the *expectation notation is removed*.

$$
\begin{aligned}
D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x})) &= \int q_\phi(\mathbf{z})\log\frac{q_\phi(\mathbf{z})}{p_\theta(\mathbf{z}\mid\mathbf{x})}d\mathbf{z} && \text{(2.4a)}\\
&= \int q_\phi(\mathbf{z})(\log q_\phi(\mathbf{z}) - \log p_\theta(\mathbf{z}\mid\mathbf{x}))d\mathbf{z} && \text{(2.4b)}\\
&= \int (q_\phi(\mathbf{z})\log q_\phi(\mathbf{z}) - q_\phi(\mathbf{z})\log p_\theta(\mathbf{z}\mid\mathbf{x}))d\mathbf{z} && \text{(2.4c)}\\
&= \int q_\phi(\mathbf{z})\log q_\phi(\mathbf{z})d\mathbf{z} - \int q_\phi(\mathbf{z})\log p_\theta(\mathbf{z}\mid\mathbf{x})d\mathbf{z} && \text{(2.4d)}\\
&= \mathbb{E}_q[\log q_\phi(\mathbf{z})] - \mathbb{E}_q[\log p_\theta(\mathbf{z}\mid\mathbf{x})] && \text{(2.4e)}\\
&= \mathbb{E}_q[\log q_\phi(\mathbf{z})] - \mathbb{E}_q\left[\log\frac{p_\theta(\mathbf{x},\mathbf{z})}{p_\theta(\mathbf{x})}\right] && \text{(2.4f)}\\
&= \mathbb{E}_q[\log q_\phi(\mathbf{z})] - \mathbb{E}_q[\log p_\theta(\mathbf{x},\mathbf{z}) - \log p_\theta(\mathbf{x})] && \text{(2.4g)}\\
&= \mathbb{E}_q[\log q_\phi(\mathbf{z}) - \log p_\theta(\mathbf{x},\mathbf{z})] + \mathbb{E}_q[\log p_\theta(\mathbf{x})] && \text{(2.4h)}\\
&= \mathbb{E}_q[\log q_\phi(\mathbf{z}) - \log p_\theta(\mathbf{x},\mathbf{z})] + \log p_\theta(\mathbf{x}) && \text{(2.4i)}
\end{aligned}
$$

Equating the first and last equations of 2.4, we get equation 2.5.

$$
D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x})) = \mathbb{E}_q[\log q_\phi(\mathbf{z}) - \log p_\theta(\mathbf{x},\mathbf{z})] + \log p_\theta(\mathbf{x}). \tag{2.5}
$$

We know that both $p_\theta(\mathbf{z}\mid\mathbf{x})$ and $p_\theta(\mathbf{x})$ are intractable. As such both $D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x}))$ and $\log p_\theta(\mathbf{x})$ are **intractable**. In order to reach a **tractable bound for optimisation**, all intractable terms together are grouped together in equation 2.6.

$$
\mathbb{E}_q[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z})] = \log p_\theta(\mathbf{x}) - D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x})). \tag{2.6}
$$

We thus now arrive at the common form of the $\mathcal{ELBO}(q)$ in 2.7.

$$
\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z}) = \mathbb{E}_q[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z})] = \mathbb{E}_q\left[\log\frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z})}\right]. \tag{2.7}
$$

Our work focuses on **latent regularisation**, however, equation 2.6 still has an **intractable divergence**, due to the intractability of $p_\theta(\mathbf{z}\mid\mathbf{x})$. The $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$, is decomposed even further to reach a form with a **tractable divergence**, that will be the focus of our work.

We begin by reproducing with equation 2.7 in equation 2.8a. In equation 2.8b, we spit a *single expectation of two terms* into *two expectations of single terms*. We then apply *Bayes rule* to the first term to get equation 2.8c, and split the first term using the *logarithm properties* in equation 2.8d. In equation 2.8e, we then *group* the last two terms together.

In equation 2.8f, we join the last two terms together, using *logarithm properties* and rewrite the *expectation with respect to the variational posterior* $q_\phi(\mathbf{z})$ as integrals. Finally, we notice that this equation may be rewritten as a (reverse) KL divergence, in the form noted in equation 2.3, to get equation 2.8g.

$$\begin{aligned}
\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z}) &= \mathbb{E}_q[\log p_\theta(\mathbf{x},\mathbf{z}) - \log q_\phi(\mathbf{z})] && \text{(2.8a)}\\
&= \mathbb{E}_q[\log p_\theta(\mathbf{z},\mathbf{x})] - \mathbb{E}_q[\log q_\phi(\mathbf{z})] && \text{(2.8b)}\\
&= \mathbb{E}_q[\log(p_\theta(\mathbf{x}\mid\mathbf{z})p_\theta(\mathbf{z}))] - \mathbb{E}_q[\log q_\phi(\mathbf{z})] && \text{(2.8c)}\\
&= \mathbb{E}_q[\log p_\theta(\mathbf{x}\mid\mathbf{z})] + \mathbb{E}_q[\log p_\theta(\mathbf{z})] - \mathbb{E}_q[\log q_\phi(\mathbf{z})] && \text{(2.8d)}\\
&= \mathbb{E}_q[\log p_\theta(\mathbf{x}\mid\mathbf{z})] + \mathbb{E}_q[\log p_\theta(\mathbf{z}) - \log q_\phi(\mathbf{z})] && \text{(2.8e)}\\
&= \mathbb{E}_q[\log p_\theta(\mathbf{x}\mid\mathbf{z})] + \int q_\phi(\mathbf{z}) \log \frac{p_\theta(\mathbf{z})}{q_\phi(\mathbf{z})} d\mathbf{z} && \text{(2.8f)}\\
&= \mathbb{E}_q[\log p_\theta(\mathbf{x}\mid\mathbf{z})] - D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z})). && \text{(2.8g)}
\end{aligned}$$

Optimising the VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$ from equation 2.9, is thus equivalent to minimising $D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x}))$, the (reverse) KL divergence between the variational latent posterior and latent prior, in the form noted in equation 2.9. Importantly, this **divergence is tractable**.

$$\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z}) = \mathbb{E}_q[\log p_\theta(\mathbf{x}\mid\mathbf{z})] - D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z})). \tag{2.9}$$

2.3.1 Understanding $\mathcal{ELBO}(q)$

Optimising the $\mathcal{ELBO}(q)$ with respect to **both** variational parameters ϕ and generative model parameters θ ratchets up the **likelihood** towards an optimum. Maximising $\mathcal{ELBO}(q)$ is thus synonymous with maximising log-likelihood $\log p_\theta(\mathbf{x})$ and may thus be seen as a form of MLE.

We explore the KL divergence further in section 2.4.1. For now, we note that $D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x}))$ is a **strictly positive** metric, as it is constructed from integrals of strictly positive probability distributions, as seen in equation 2.3. $\mathcal{ELBO}(q)$ thus forms a **lower bound** on the model's evidence $\log p_\theta(\mathbf{x})$ and allows for the **estimation of intractable terms in** $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$. The tightness of this bound is determined by $D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z}\mid\mathbf{x}))$. When the variational posterior and prior are identical, the bound is exactly equal to the evidence. As such, the variational parameters ϕ thus control the tightness of the bound rather than the overall level.

Kingma and Welling [2013] note that, when the $\mathcal{ELBO}(q)$ is seen in the form of 2.9, the connection to auto-encoders becomes clear. The second term is the KL divergence between the (variational) latent posterior and the latent prior and acts as a regulariser, penalising the latent posterior for differing from the latent prior.

Without any simplifying assumptions, the $\mathcal{ELBO}(q)$ objective may be intractable. Often conjugacy, where the prior and the posterior belong to the *same parametric family*, is assumed. It is common for the variational posterior to be **Gaussian with a diagonal covariance matrix** and for the prior to be an **isotropic unit Gaussian prior** $\mathcal{N}(0, 1)$. This simplified form will be the focus of our work. The fact that the Gaussian is unimodal means it is unable to readily adapt to the multimodal generative spaces we are interested in. We explore the impact such multimodal generative factors have when the posterior cannot readily adapt to this multimodality. More on this in chapter 4

Asperti and Trentin [2020] stress the tension between the two components of the $\mathcal{ELBO}(q)$. The terms have contrasting effects: the reconstruction loss tries to improve quality while neglecting the latent space's shape; by contrast, the KL divergence normalises the latent space.

2.3.2 Optimising $\mathcal{ELBO}(q)$

Kingma and Welling [2019] note that despite the recognition model's arbitrary complexity, it is still reasonably fast since, by construction, it is a single feed-forward pass from input to latent variables. However, a key issue for optimising the $\mathcal{ELBO}(q)$ is that **sampling induces noise**. The $\mathcal{ELBO}(q)$'s gradients, with respect to ϕ, thus need to be **stabilised to facilitate learning**. Kingma and Welling [2019] maintain that the greatest contribution of the VAE may be the realisation that the **reparameterisation trick** can be used to counteract this sampling noise.

The reparameterisation trick is a simple procedure that reorganises the $\mathcal{ELBO}(q)$'s gradient computation, reducing gradient variance and making it suitable for learning. For continuous random variable $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{x})$, it is possible to express $\mathbf{z}$ as as a **deterministic variable** $\mathbf{z} = g_\phi(\epsilon, \mathbf{x})$, where ϵ is an **auxiliary variable** with **independent marginal** $p(\epsilon)$, and $g_\phi(\cdot)$ is a **vector-valued function**, parameterised by variational parameters ϕ. This allows $\mathcal{ELBO}(q)$ to be **rewritten** such that the Monte Carlo estimate of the expectation is **differentiable with respect to** ϕ, facilitating optimisation.

Since VAEs provide a principled framework for learning deep latent-variable models and their corresponding inference models, they are a popular model for disentangled representation learning, discussed in the section 2.5. We now explore the role the divergence plays in determining the resultant posterior solution in section 2.4, which follows.

2.4 Divergence choice and latent bias

In section 2.3 we derive the VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x}, \mathbf{z})$ from equation 2.9, repeated in equation 2.10, below.

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}, \mathbf{z}) = \mathbb{E}_q[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - D_{\mathbb{KL}}(q_\phi(\mathbf{z}) \| p_\theta(\mathbf{z})) \tag{2.10}$$

In equation 2.10, above, the first term is the expected negative reconstruction loss and the second KL divergence term regularises the latent space, encouraging the approximate posterior to be close to the prior. The (reverse) KL divergence defines the notion of closeness. When minimising this divergence, it biases the result towards specific posterior solutions.

We now explore the KL divergence in more detail and outline its asymmetry. Since the divergence is asymmetrical, it has both a forward and reverse direction. In the following section, we detail the bias in the approximate posterior fitted that results from minimising either direction of the KL divergence.

2.4.1 Kullback–Leibler (KL) divergence

Shlens [2014] notes that the Kullback–Leibler (KL) divergence is a metric in information theory from Cover and Thomas [1991], quantifying the proximity of two probability distributions. The metric is non-negative, asymmetric, zero if distributions match exactly and may potentially be infinite. The KL divergence effectively measures the average likelihood of observing (infinite) data with the distribution $p_\theta(\mathbf{x})$ if the particular model $q_\phi(\mathbf{x})$ actually generated the data. It is worth noting that we follow the literature and describe the KL proprieties with support $\mathbf{x}$, even though our work is concerned with the latent space $\mathbf{z}$.

Entropy, introduced by Shannon [1948], is the uncertainty of a random variable $\mathbf{x}$, defined mathematically in equation 2.11.

$$\mathbb{H}(p_\theta(\mathbf{x})) = -\int p_\theta(\mathbf{x}) \log p_\theta(\mathbf{x}) d\mathbf{x} \tag{2.11}$$

Dembo *et al.* [1991] note that the KL divergence measures the relative entropy between the two distributions defined mathematically in equation 2.12.

$$D_{\mathbb{KL}}(p_\theta(\mathbf{x}) \| q_\phi(\mathbf{x})) = \int p_\theta(\mathbf{x}) \log \frac{p_\theta(\mathbf{x})}{q_\phi(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x} \sim p_\theta(\mathbf{x})} \left[\log \frac{p_\theta(\mathbf{x})}{q_\phi(\mathbf{x})} \right] \tag{2.12}$$

Ganguly and Earp [2021] note that the KL divergence may be related to the entropy and cross entropy, where the cross entropy[3]. This highlights Murphy [2012]'s view that KL-divergence is the average extra amount of information required to encode data using the candidate distribution $q_\phi(\mathbf{x})$, rather than actual distribution $p_\theta(\mathbf{x})$.

$$D_{\mathbb{KL}}(p_\theta(\mathbf{x})\|q_\phi(\mathbf{x})) = -\mathbb{H}(p_\theta(\mathbf{x})) + \mathbb{H}(p_\theta(\mathbf{x}), q_\phi(\mathbf{x})) \tag{2.13}$$

2.4.2 KL divergence asymmetry

The KL divergence is asymmetrical and the forward KL divergence $D_{\mathbb{KL}}(p_\theta(\mathbf{z} \mid \mathbf{x})\|q_\phi(\mathbf{z}))$ does **not** equal the reverse KL divergence $D_{\mathbb{KL}}(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z} \mid \mathbf{x}))$. Importantly, the direction of the KL divergence determines the distribution that defines the expectation. As noted in section 2.1, true posterior $p_\theta(\mathbf{z} \mid \mathbf{x})$ is intractable and thus the **forward KL** is also **intractable**, as it is an expectation with respect to intractable unknown true posterior. It is for this reason that the reverse KL is typically used as the starting point of the $\mathcal{ELBO}(q)$ derivation, in equation 2.4. The **reverse KL** takes expectations with respect to variational posterior $q_\phi(\mathbf{z})$ which is **tractable by construction**, being a member of variational family $\mathcal{Q}$.

Murphy [2012] note that the forward KL divergence is known as the **M-projection** or **moment projection**, while the reverse KL is known as the **I-projection** or **information projection**. In variational inference (VI), we choose ϕ and fit $q_\phi(\mathbf{z}))$ to minimise a (typically reverse) KL divergence. The result when minimising the reverse KL divergence $D_{\mathbb{KL}}(q_\phi(\mathbf{x})\|p_\theta(\mathbf{x}))$ is different to that when minimising the forward KL divergence $D_{\mathbb{KL}}(p_\theta(\mathbf{x})\|q_\phi(\mathbf{x}))$. Again, it is worth noting that we follow the literature and describe the KL proprieties with support $\mathbf{x}$, even though our work is concerned with the latent space $\mathbf{z}$.

The difference stems from the different limiting behaviour of the respective divergences. Repeating the expectation form from 2.12, for forward KL divergence

$$D_{\mathbb{KL}}(p_\theta(\mathbf{x})\|q_\phi(\mathbf{x})) = \mathbb{E}_{\mathbf{x}\sim p_\theta(\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x})}{q_\phi(\mathbf{x})}\right], \tag{2.14}$$

we note

$$\lim_{q_\phi(\mathbf{x})\to 0} \frac{p_\theta(\mathbf{x})}{q_\phi(\mathbf{x})} \to \infty, \text{ where } p_\theta(\mathbf{x}) > 0. \tag{2.15}$$

[3] $\mathbb{H}(p_\theta(\mathbf{x}), q_\phi(\mathbf{x})) = -\int p_\theta(\mathbf{x}) \log q_\phi(\mathbf{x}) d\mathbf{x}$

Murphy [2012] notes that the limit in equation 2.15 is infinite if $q_\phi(\mathbf{x}) = 0$ and $p_\theta(\mathbf{x}) > 0$. Thus order to minimise the divergence we must choose ϕ s.t $q_\phi(\mathbf{x}) > 0$ when $p_\theta(\mathbf{x}) > 0$. In this **"zero-avoiding"** behaviour, $q_\phi(\mathbf{x})$ over-estimates $p_\theta(\mathbf{x})$.

Conversely, for the expectation form of the reverse KL divergence,

$$D_{\mathbb{KL}}(q_\phi(\mathbf{x})\|(p_\theta(\mathbf{x})) = \mathbb{E}_{\mathbf{x}\sim q_\phi(\mathbf{x})}\left[\log \frac{q_\phi(\mathbf{x})}{p_\theta(\mathbf{x})}\right], \tag{2.16}$$

.

we note

$$\lim_{q_\phi(\mathbf{x})\to 0} \frac{q_\phi(\mathbf{x})}{p_\theta(\mathbf{x})} \to \infty, \text{ where } q_\phi(\mathbf{x}) > 0. \tag{2.17}$$

Murphy [2012] notes the limit in equation 2.17 is infinite if $p_\theta(\mathbf{x}) = 0$ and $q_\phi(\mathbf{x}) > 0$. Thus order to minimise the divergence we must choose ϕ s.t $q_\phi(\mathbf{x}) = 0$ whenever $p_\theta(\mathbf{x}) = 0$. In this **"zero-forcing"** behaviour, $q_\phi(\mathbf{x})$ under-estimates $p_\theta(\mathbf{x})$.

2.4.3 Divergence choice, data structure and solution bias

A note on the space of interest

In this section, we **follow the literature** and outline the variational posterior solution bias induced by either direction of the KL divergence **in data space** $\mathbf{x}$. However, it is important to note that our work focuses on the regularisation in the latent space $\mathbf{z}$. The behaviour and solution bias remains the same, but the space of interest in our work is the **latent space** $\mathbf{z}$ **rather than the data space** $\mathbf{x}$.

A note on "mode" terminology

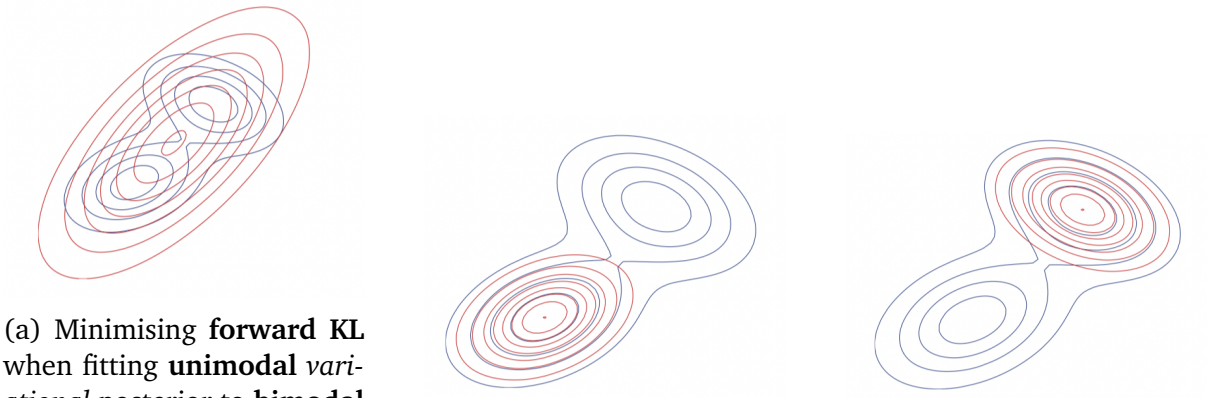
Throughout this work, we refer to a region of high probability, relative to the rest of the distribution, as ***"mode"***. This is in keeping with VAE literature, where preventing ***"mode collapse"*** is an active area of research. Furthermore, we refer to ***"multimodal"*** distributions as a distribution having more than one mode. It should be noted that we ***do not explore "multimodal" inputs***, as in other machine perception work, where inputs may be many classes of data such as images and text.

Our work focuses on unimodal variational posteriors and multimodal generative spaces, outlined in chapter 4. Murphy [2012] details that when the *true* posterior distribution $p_\theta(\mathbf{z} \mid \mathbf{x})$ is **multimodal** and *variational* posterior approximation $q_\phi(\mathbf{z})$ is **unimodal**, the forward KL places probability mass in regions of low true probability. This is seen in 2.3a, where the *true* posterior distribution (contours in blue) is **bimodal** and *variational* approximation (contours in red) is **unimodal**. This results in a *variational* posterior mode and mean in an area of low probability between the two peaks, with $q_\phi(\mathbf{z} \mid \mathbf{x})$ consuming the support of $p_\theta(\mathbf{z} \mid \mathbf{x})$.

Zhang *et al.* [2019] note that this behaviour may be termed as **"mean-matching"**. Dieng *et al.* [2016] note that minimising the forward KL favours **over-dispersed approximations**, relative to the *true* posterior distribution, with the *variational* posteriors probability mass distributed far wider than true posterior's support. This performs well on models with light-tailed true posteriors.

Conversely, in figures 2.3b and 2.3c we see that when using the reverse KL the **unimodal** *variational* approximation locks onto one of two modes. Murphy [2012] notes that this only makes more statistical sense but, importantly, is more tractable since the expectation is with respect to a member from the known, tractable variational family $\mathcal{Q}$.

Zhang *et al.* [2019] note that behaviour may be termed **"mode-seeking"**. As a result of this behaviour, key modes from the *true* posterior distribution are dropped, and the unimodal variational approximation locks onto a particular mode. Dieng *et al.* [2016] note minimising the reverse KL favours **under-dispersed approximations**, relative to the *true* posterior distribution, with the *variational* posterior's mass being distributed far wider than the resultant posterior support. Difficulties arise when heavier tail variational distributions are used with light-tailed exact posteriors, which can lead to unstable optimisation and poor approximation.



(a) Minimising **forward KL** when fitting **unimodal** *variational* posterior to **bimodal** *true* posterior distribution, results in variational approximation covers true posterior. Note that regions with low true probability are erroneously assigned high probabilities.

(b) Minimising **reverse KL** when fitting **unimodal** *variational* posterior to **bimodal** *true* posterior distribution, results in variational approximation locking onto a single mode.

(c) Minimising **reverse KL** when fitting **unimodal** *variational* posterior to **bimodal** *true* posterior distribution, results in variational approximation locking onto a single mode.

Figure 2.3: Fitting a **unimodal** *variational* approximation (contours in red) to a **bimodal** *true* posterior distribution (contours in blue). Minimising either the forward or reverse KL results in different solutions. Figures from Murphy [2012].

2.4.4 Importance of KL divergence

Prokhorov *et al.* [2019] note how the KL divergence controls information transmitted in the VAE channel and highlight the **trade-off between information encoded and model generative capacity**.

The (reverse) KL divergence between prior $p_\theta(\mathbf{z})$ and latent posterior $D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p(\mathbf{z})\right)$, the key focus of our work, acts as a regulariser preventing the inference network $q_\phi(\mathbf{z} \mid \mathbf{x})$ from over-fitting and memorising data $\mathbf{x}$. When both prior and posterior are Gaussian, the KL has a closed-form solution. The **encoder channel's capacity** may be measured via the mutual information between $\mathbf{x}$ and $\mathbf{z}$: $\mathbb{I}(\mathbf{x}; \mathbf{z})$, defined in section 2.8.1, that follows.

Prokhorov *et al.* [2019] go on to note the VAE's encoder-decoder structure, seen in figure 2.2, is a noisy sender-receiver communication network. Input communication $\mathbf{x}$ is sender generated and encoded noisily into message $\mathbf{z}$. The receiver then tries to decode $\mathbf{z}$ to match $\mathbf{x}$ with the highest fidelity. The terms of $\mathbf{z}$'s compression rate $\mathbb{R}$ and distortion $\mathbb{D}$ bound the encoder channel's capacity. Seen in this manner the reconstruction loss term naturally measures distortion ($\mathbb{D} := -\log p_\theta(\mathbf{x} \mid \mathbf{z})$); conversely the KL divergence is the compression rate ($\mathbb{R} := D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right)$), measuring the divergence between $p_\theta(\mathbf{z})$, a channel which transmits no information about $\mathbf{x}$, and VAE encoder channel $q_\theta(\mathbf{z} \mid \mathbf{x})$.

Alemi *et al.* [2018] bounded the mutual information $\mathbb{I}(\mathbf{x}; \mathbf{z})$ using the $\mathbb{R}$, $\mathbb{D}$ and constant empirical data entropy $\mathbb{H}(\mathbf{x})$

$$\mathbb{H}(\mathbf{x}) - \mathbb{D} \leq \mathbb{I}(\mathbf{x}; \mathbf{z}) \leq \mathbb{R}, \tag{2.18}$$

or using the equivalent mathematical quantities

$$\mathbb{H}(\mathbf{x}) + \log p_\theta(\mathbf{x} \mid \mathbf{z})) \leq \mathbb{I}(\mathbf{x}; \mathbf{z}) \leq D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right)). \tag{2.19}$$

Using inequality 2.18, the trade-off between reconstruction and the KL term becomes clearer.

Increasing $\mathbb{I}(\mathbf{x}; \mathbf{z})$ can be encouraged by increasing both bounds of $\mathbb{H} - \mathbb{D} \leq \mathbb{I}(\mathbf{x}; \mathbf{z}) \leq \mathbb{R}$. Increasing the upper bound ($\mathbb{R} := D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right)$) may be seen as a means to control the maximum capacity of the encoder channel. Conversely reducing the distortion ($\mathbb{D} := -\log p_\theta(\mathbf{x} \mid \mathbf{z})$) tightens the bound, pushing the bound towards its limit $\mathbb{H}$. A stronger decoder may achieve a similar effect on the lower bound by potentially decreasing the reconstruction loss. Alternatively the KL divergence may be explicitly controlled via the β VAE formulation discussed in section 2.5.1.

2.5 Disentangled representation learning

VAEs excel at representation learning, where latents $\mathbf{z}$ learn something useful about the data's ground truth. Data $\mathbf{x}$ is assumed to be generated from a simulator process characterised by higher-level, ground-truth generative factors that describe the data. It is assumed that the data contains a balanced distribution of ground-truth generative factors $(\mathbf{v},\ \mathbf{w})$, where at least some of these ground-truth generative factors $\mathbf{v}$ are independent with **given** interpretable labels, and the other $\mathbf{w}$ ground-truth factors are non-interpretable. We have $\mathbf{x} \sim \mathrm{Sim}(\mathbf{v}, \mathbf{w})$, for simulation operator Sim.

Burgess *et al.* [2018] note that **disentangled representations** go beyond independence, with **particular** latent units sensitive to changes in **particular** generative factors, but invariant other factors. Such representations are factorised and often interpretable, with specific independent latent units encoding particular independent ground-truth data generative factors of variation. This one-to-one relationship allows for axis-alignment between the generative factors and latents and provides interpretability since at least some generative factors $\mathbf{v}$ are assumed to have semantic meaning.

Our baseline model, the β VAE, is introduced in section 2.5.1, outlining some mechanics of why the model is good for disentangling. We then outline supervised, and unsupervised metrics used to quantify disentangling in section 2.6.

2.5.1 Disentangling using $\boldsymbol{\beta}$ VAE

$\boldsymbol{\beta}$ VAE is a generalisation of the VAE developed specifically by Higgins *et al.* [2017] to induce disentangled representation learning.

$$\mathcal{L}_{\theta,\phi,\boldsymbol{\beta}}(\mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x} \mid \mathbf{z})\right] - \boldsymbol{\beta} D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right), \tag{2.20}$$

where $\boldsymbol{\beta} > 0$.

The vanilla VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x}, \mathbf{z})$ from equation 2.9, is modified to include the hyperparameter $\boldsymbol{\beta}$, resulting in the $\boldsymbol{\beta}$ VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi\boldsymbol{\beta}}(\mathbf{x}, \mathbf{z})$ seen in equation 2.20. Typically this hyperparameter $\boldsymbol{\beta} > 1$; this increases the impact of latent regularisation in the $\mathcal{ELBO}(q)$ and provides **additional pressure**, when compared to the typical VAE. This encourages for the resultant posterior approximation $q_\phi(\mathbf{z} \mid \mathbf{x})$ to be closer to independent Gaussian prior $p_\theta(\mathbf{z})$ and orthogonal, when compared to the vanilla VAE. However, the exact value of $\boldsymbol{\beta}$ that is needed to induce disentangling changes depending on the network architectures selected and the data itself. Selection of hyperparameter $\boldsymbol{\beta}$ is thus not straightforward, requiring timely tuning for each network-dataset combination.

2.5.2 β for bottle neck

A high β in the β VAE may create a greater trade-off between reconstruction fidelity and disentanglement. For disentangled representations to emerge, **information preservation** and **latent channel capacity restriction** must be appropriately balanced. $\beta > 1$ often results in good disentanglement with blurry reconstructions as high-frequency details are lost as gradients pass through a constrained latent bottleneck. Burgess *et al.* [2018] associates poor reconstruction with a constant latent capacity constriction (β value) applied over training.

Higgins *et al.* [2017] hypothesised that **constrained optimisation** is important to encourage the unsupervised learning of disentangled representations, hence the inclusion of the β hyperparameter. Furthermore Burgess *et al.* [2018] note the similarity between the $\mathcal{ELBO}(q)$ in equation 2.20, and the information bottleneck principle, given for mutual information $\mathbb{I}(\cdot;\cdot)$ Lagrange multiplier β below:

$$\max[\mathbb{I}(\mathbf{z};\mathbf{y}) - \beta\mathbb{I}(\mathbf{x};\mathbf{z})]. \tag{2.21}$$

An information bottleneck is concerned with the constrained optimisation objective in equation 2.21, maximising mutual information between latent bottleneck $\mathbf{z}$ and some auto-encoder task $\mathbf{y}$. Irrelevant information about $\mathbf{y}$ that is present in the input $\mathbf{x}$ is discarded. Posterior $q_\phi(\mathbf{z} \mid \mathbf{x})$ may be seen as an information bottleneck for reconstruction task $\max \mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x} \mid \mathbf{z})]$. The $\mathcal{ELBO}(q)$ in equation 2.20 thus encourages $q_\phi(\mathbf{z} \mid \mathbf{x})$ to efficiently transmit information about data $\mathbf{x}$. This is done by minimising the β weighted KL term, while **jointly** maximising the data log-likelihood.

Role of bottleneck in representation learning

As long as posterior and prior are factorised (diagonal covariance), posterior samples may be obtained using the reparameteriseation trick described in section 2.3.2. The trick allows posterior samples to be drawn by adding scaled independent Gaussian noise $\sigma_i\epsilon_i$ to a deterministic encoder mean μ_i for each latent unit $\mathbf{z}_i = \mu_i + \sigma_i\epsilon$. The posterior may thus be thought of as a set of **independent additive Gaussian white noise channels** $\mathbf{z}_i$, each noisily transmitting information about the data.

The KL divergence term in the $\mathcal{ELBO}(q)$ may thus be seen as an **upper bound** of information transmission via latent channels **per sample**. The latent channel capacity is increased by dispersing posterior means across data and decreasing posterior variances, both of which increase the KL divergence.

A bottleneck reduces latent capacity, encouraging representations with less mean dispersion and increased variance, reducing the KL divergence. Each posterior sampled is thus squeezed into a shared coding space. The more posterior samples that overlap, the smaller the KL divergence with log-likelihood costs since each sample's posterior has reduced average discriminability. For any overlapping posteriors, there is a chance

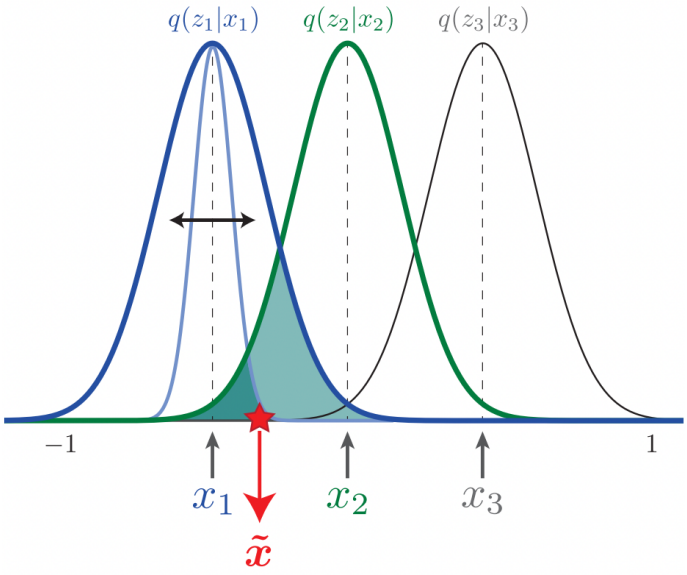


Figure 2.4: **Posterior overlap increased through bottleneck:** Sample's KL divergence of between prior & posterior reduced by broadening posteriors and bringing means closer together, increasing posterior overlap.

$\tilde{x}$ sampled from distribution $q(z_2 \mid x_2)$ will be confused as being from $q(z_1 \mid x_1)$ as posterior overlap over datapoints increases. When neighbouring points in data space are also close in latent space, the log-likelihood cost of this confusion is reduced. Figure from Burgess *et al.* [2018]. We abuse our notion slightly to align with this figure.

a sample may have a higher probability under another sample's fitted posterior. **For the smallest log-likelihood cost, nearby points in data space are arranged close together in the latent space**. Overlapping posteriors thus encourage locality in the manifold resulting in smooth latent transitions, shown in figure 2.4.

Burgess *et al.* [2018] hypothesised that β VAE finds latents that make a **different** contribution to the $\mathcal{ELBO}(q)$'s log-likelihood (reconstruction) term, suspecting that this is due to the diagonal posterior approximation. They hypothesise that this is why the β VAE can recover representational axes that are aligned with the data's generative factors.

2.5.3 Diagonal posterior promotes disentangling

Rolinek *et al.* [2019] claim disentanglement requires a latent space **aligned** with the coordinate axes. They note that the VAE's design does not **explicitly** suggest any such mechanism. They claim the **diagonal posterior approximation**, together with the inherent stochasticity, force **local orthogonality** of the decoder.

Latent space rotations may be equally suitable for reconstruction, but disentanglement is **sensitive to axis-alignment**. Enforcing a diagonal posterior in the encoder disrupts any rotational asymmetry inherited from the prior. This encourages disentangling.

2.5.4 Inductive biases and (implicit) supervision

Work from Locatello *et al.* [2019] challenges whether truly unsupervised disentangling is possible. They theoretically show that truly unsupervised disentangling is fundamentally impossible and that common methods inject inductive bias via the chosen models and their appropriateness for the dataset they are applied on.

In their extensive study, they train over 12000 disentangling models. They observe that the various disentangling models do successfully enforce properties via their loss objective, but supervision is required to measure disentangling and assess their merit.

Importantly they claim **"while all considered methods prove effective at ensuring that the individual dimensions of the aggregated posterior (which is sampled) are not correlated, we observe that the dimensions of the representation (which is taken to be the mean) are correlated."** and note that practitioners should be explicit about inductive biases and (implicit) supervision in their experiments.

When seen in this light, disentangling is not truly unsupervised, with the inductive bias being injected via the model selected. Furthermore, the residual correlation may remain between latents, despite which objective is enforced.

2.6 Assessing disentangling with latent traversals

The goal of disentangling is to learn latent dimensions that are aligned with generative factors in a one-to-one fashion. Knowing these generative factors' ground-truth is thus particularly useful when quantifying and measuring disentangling. Synthetic datasets, with known ground-truths, are often used in disentangling research, discussed in section 4.2 where our proposed group of datasets *mSprites* is introduced.

When measuring disentangling, literature often uses **latent traversals as a subjective qualitative assessment** of how well a latent identifies a single generative factor. This involves generating data by changing only a single latent dimension while keeping other dimensions constant. Disentangling is assessed by inspection of the data produced by the generative model and evidenced when only one generative factor changes and the other factors remain unchanged. For example, in figure 2.5, we can see that when the latent unit associated with wall colour is traversed, we see that only the wall colour of the generative image changes and other factors are not impacted. Subjectively creeps in, particularly when results are less clear and the degree of entanglement is difficult to assess via inspection, i.e. latent traversal only indicates if a latent unit is clearly disentangled.

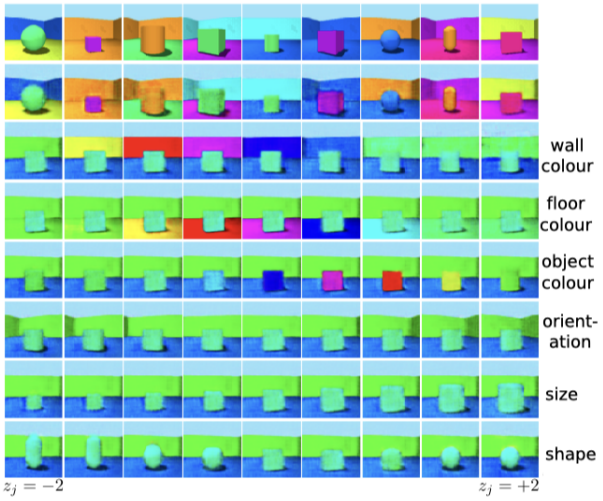


Figure 2.5: **Example of latent traversals**, from Kim and Mnih [2018], used to subjectively assess disentangling via inspection of generated data. Interpretation of factors traversed are given on the right.

Kim and Mnih [2018] note that the qualitative nature of latent traversals makes it **unsuitable for comparing algorithms reliably**. Furthermore, they note that latent traversal assessment **requires a human in the loop** and is **time-consuming**.

A better understanding of the merit of new models and methods may be provided through quantitative measures. To this end, we outline a host of complementary disentangling metrics. Kim and Mnih [2018] note that unsupervised disentangling metrics that do not need ground-truth generative factor labels are important since such labels are often not accessible. However, they still believe that **supervised disentangling metrics still serve as a gold standard**.

The in sections 2.8 and 2.7, that follow, the**outlined metrics are both supervised and unsupervised**. The selection of metrics seeks to assess disentanglement based on axis alignment and overall fit (global informational content shared) between the generative factors and the learnt latents. Furthermore, inter-dimensional relationships between individual latents, as well as that between latents and generative factors, are considered. This provides a **holistic, multi-faceted view of the representation learnt and whether it is disentangled**. All but the Higgins' metric[4]. are used in the experiments outlined in chapter 5.

2.7 Assessing disentangling with supervised metrics

For simulation operator Sim, data is generated from true world ground-truth simulation process $\mathbf{x} \sim \mathrm{Sim}(\mathbf{v}, \mathbf{w})$. In this process at least some generative factors, $\mathbf{v}$, are

[4]Factor score and mutual information gap are used over the Higgins' metrics for their ability to detect axis alignment.

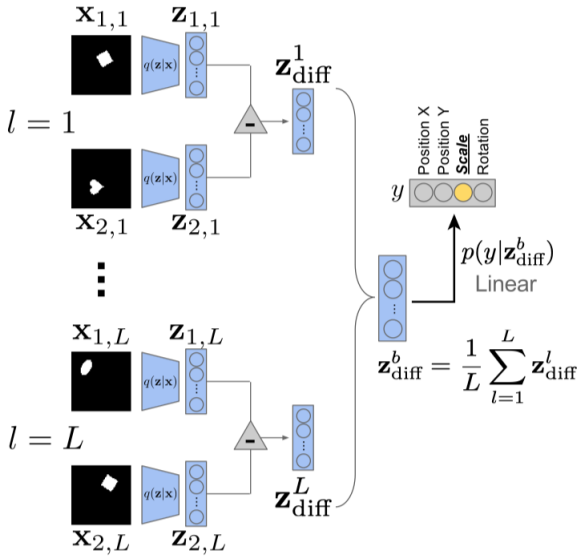


Figure 2.6: **Schematic for Higgins' metric:** For batch of L samples, each pair of images has fixed target generative factor y value (y = scale in this figure), differing for other generative factor in $\mathbf{v}$. Linear classifier trained to identify y, using average pairwise difference $\mathbf{z}^b_{\text{diff}}$. Figure from Higgins *et al.* [2017].

both conditionally **independent** and **interpretable**, and the remaining factors, $\mathbf{w}$, are **not** independent nor interpretable. A representation that is disentangled with respect to these generative factors encodes each factor as a separate latent. As such there is a single generative factor is associated with a single latent unit. This allows for robust classification, hence providing **interpretability**.

This notion of alignment between the generative factors and separate latents is paramount for interpretability and goes beyond simple independence. Simple cross correlation between inferred latents will **not** suffice as the interpretability desiderata would not be captured.

2.7.1 Higgins' metric

Higgins *et al.* [2017] introduced a metric to quantify the level of disentanglement, measuring both the **independence** and **interpretability** of the resultant representation. This evaluation metric is often referred to as the **Higgins' metric**.

It is important to note that in order to use this metric, the true underlying interpretable generative factors $\mathbf{v}$ should be **known**, allowing for the creation of a supervised evaluation metric. While this is acceptable for toy examples, where the data is generated by the researcher, knowing ground-truth factors is a particularly onerous condition and is generally not known for real-world data. This is particularly true since VAEs are unsupervised learning models.

The process for calculating the metrics is presented in figure 2.6. The metric quantifies the accuracy of a linear classifier identifying the fixed ground-truth factors. Recall,

from section 2.5, that data $\mathbf{x}$ is produced by ground-truth factor $\mathbf{x} \sim \mathrm{Sim}(\mathbf{v}, \mathbf{w})$, with at least some factors $\mathbf{v}$ that are conditionally independent and interpretable.

In the Higgins' metric, batches of L images are generated by fixing one of the K generative data factors $\mathbf{y}$, chosen uniformly, i.e. $\mathbf{y} \sim \mathrm{uniform}[1 \dots K]$. Other randomly sampling from the other factors. For each index l, in the batch of size L, the metric is calculated over a batch of L samples, each containing two sets of ground-truth factors $\mathbf{v}_{1,l}$ and $\mathbf{v}_{2,l}$, with the value of factor $k = y$ kept fixed.

For each of the L pairs of ground-truth generative factors, simulate image data $\mathbf{x}_{1,l} \sim \mathrm{Sim}\left(\mathbf{v}_{1,1}\right)$ and $\mathbf{x}_{2,l} \sim \mathrm{Sim}\left(\mathbf{v}_{2,1}\right)$. $\mathbf{z}_{1,l} = \mu\left(\mathbf{x}_{1,l}\right)$ and $\mathbf{z}_{2,l} = \mu\left(\mathbf{x}_{2,l}\right)$ are then inferred using the encoder $q_\phi(\mathbf{z} \mid \mathbf{x}) \sim N(\mu(\mathbf{x}), \sigma(\mathbf{x}))$. The absolute linear difference between the inferred latent representations is then computed $\mathbf{z}^l_{\text{diff}} = \left|\mathbf{z}_{1,l} - \mathbf{z}_{2,l}\right|$. The average $\mathbf{z}^b_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} \mathbf{z}^l_{\text{diff}}$ is then used to predict $p\left(y \mid \mathbf{z}^b_{\text{diff}}\right)$. The Higgins' metric is the classifier's accuracy in predicting y over multiple batches.

A linear classifier is chosen so that it cannot aid in non-linear disentangling. The differences of two inferred latents $\mathbf{z}^l_{\text{diff}}$ are used to reduce the classifier input variance and reduce the conditional dependence on the inputs. The key insight is that the inferred latents should have less variance when corresponding to the fixed generative factor y. Within the batch, this low variance latent is identified with the linear classifier. The classifier's accuracy in predicting y over multiple batches is the Higgins' metric.

2.7.2 Axis alignment detection via factor score

A note on notation abuse in the section that follows

In outlining the factor score in this section we abuse our notion to align with that of Kim and Mnih [2018] and figure 2.7.

Kim and Mnih [2018] note that the Higgins' metric has several weaknesses. Firstly, it could be sensitive to hyperparameters of the linear classifier optimisation, such as optimiser choice, its hyperparameters, the weight initialisation and the number of training iterations. They also note that having a linear classifier is not intuitive and that this could result in latents where each dimension is a linear combination of other dimensions. Lastly, and most importantly, they note that the Higgins' metric has a failure mode: 100% accuracy is assigned when K-1 of K total factors are disentangled.

An axis-aligned representation has one latent factor that relates to one ground-truth generative factor. A criticism of Higgins' metric is that it lacks axis alignment detection, a criterion that is needed for a truly disentangled representation.

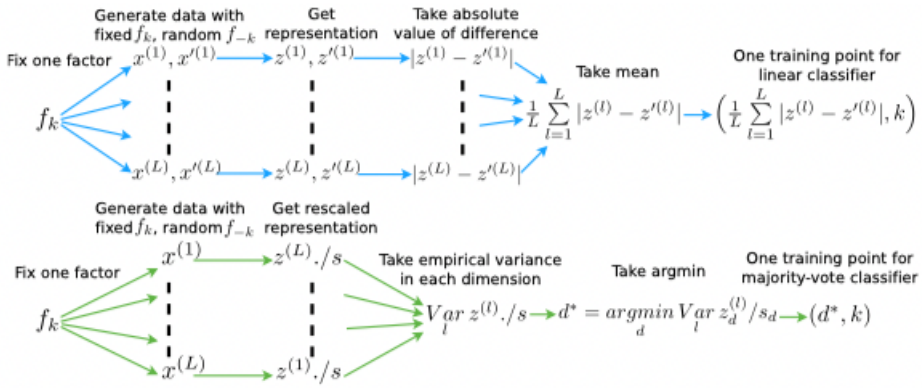


Figure 2.7: **Schematic for factor score** (below), contrasted against Higgins' metric schematic (above). Figure from Kim and Mnih [2018]. **Note the notation abuse in this figure.**

The process for calculating the factor score, contrasted against the Higgins' metric, is given in figure 2.7. Calculating an input/output pair starts by choosing factor dimension k, f_k, fixing it, and allowing all other factor dimensions f_{-k} to vary. For each of the L data dimensions generated, $x^{(1)} \dots x^{(L)}$, get their rescaled representations of each sample $z^{(1)} \cdot /s \dots z^{(L)} \cdot /s$. Here the representation $z^{(\cdot)}$ are normalised by their empirical standard deviation s over all data (or large enough subset). The empirical variance in each dimension $\mathrm{Var}\, z^{(l)} \cdot /s$. The index of the dimension with the lowest variance d^* is then taken, $d^* = \underset{d}{\mathrm{argmin}}\, \mathrm{Var}_l\, z_d^{(l)}/s_d$, and the input-out pair is then (d^*, k).

The inputs and outputs are both in the discrete space Kim and Mnih [2018] note that the optimal classifier is a majority-vote classifier. **The factor score is defined as the error rate of the classifier. If the representation is perfectly disentangled the fixed factor's empirical variance will be 0. Thus a lower factor score is better for disentangling.** Kim and Mnih [2018] note that the resulting classifier is a deterministic function of the training data and thus has no optimisation hyperparameters to tune. Importantly, it does not have the aforementioned failure mode of the Higgins' metric.

Despite the popularity of the Higgins' metric in disentangling literature, the **factor score is used over the Higgins' metric in this work to include axis alignment detection**.

2.8 Assessing disentangling with unsupervised metrics

We now outline the unsupervised metrics that may be used to assess disentangling. Such unsupervised disentangling metrics do not need ground-truth generative factor labels, and often such labels are not accessible. Recall that Kim and Mnih [2018] believe that **supervised disentangling metrics serve as a gold standard** for assessing disentangling.

2.8.1 Mutual information (MI)

Tschannen *et al.* [2019] notes that the mutual information (MI) between random variables $\mathbf{x}$ and $\mathbf{y}$, that have joint density $p(\mathbf{x}, \mathbf{y})$ and marginal densities $p(\mathbf{x})$ and $p(\mathbf{y})$, is defined in equation 2.22. as the Kullback–Leibler (KL) divergence, from section 2.4.1, between the joint and the product of the marginals.

$$\mathbb{I}(\mathbf{x}; \mathbf{y}) = D_{\mathbb{KL}}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) = \mathbb{E}_{p(\mathbf{x},\mathbf{y})}\left[\log \frac{p(\mathbf{x}, \mathbf{y})}{p(\mathbf{x})p(\mathbf{y})}\right] \tag{2.22}$$

Tschannen *et al.* [2019] observe that many recent unsupervised representation learning feature extractors are trained by maximising an estimate of mutual information on different views of the data. They note that **estimating** the MI in high-dimensional spaces is notoriously difficult, and, in practice, tractable lower bounds are often maximised instead. However, despite this, such methods often excel in practice.

An estimate of empirical mutual information may also be used as a metric rather than an objective component, measuring the total informational content shared between latents $\mathbf{z}$ and interpretable generative ground-truth factors $\mathbf{v}$. Seen in this light, MI is a metric of **global overlap** between the latents and generative factors distributions. MI is not strictly a disentangling metric; however, a **higher MI means more information about the factors is captured over all latents and is thus desirable in representation learning**. For disentangling, a **particular** latent should uniquely capture information about a **particular** generative factor; increased global information, indicated by the MI, thus seems broadly aligned with this goal.

2.8.2 Mutual information gap (MIG)

Chen *et al.* [2018] propose an estimate for the **empirical mutual information** between latent $\mathbf{z}_j$ and ground-truth generative factor $\mathbf{v}_k$. They use this empirical estimate to calculate in their proposed metric the Mutual Information Gap (MIG), outlined in section 2.8.2 that follows. Chen *et al.* [2018] note that the empirical mutual information may be estimated using the joint distribution $q(\mathbf{z}_j, \mathbf{v}_k) = \sum_{n=1}^{N} p(\mathbf{v}_k)\, p(n \mid \mathbf{v}_k)\, q(\mathbf{z}_j \mid n)$. Assuming the underlying factors $p(\mathbf{v}_k)$, along with the generating process, is known for empirical data samples $p(n \mid \mathbf{v}_k)$, then

$$\mathbb{I}_n(\mathbf{z}_j; \mathbf{v}_k) = \mathbb{E}_{q(\mathbf{z}_j,\mathbf{v}_k)}\left[\log \sum_{n \in \mathcal{X}_{\mathbf{v}_k}} q(\mathbf{z}_j \mid n)\, p(n \mid \mathbf{v}_k)\right] + \mathbb{H}(\mathbf{z}_j), \tag{2.23}$$

where $\mathcal{X}_{\mathbf{v}_k}$ is the support of $p(n \mid \mathbf{v}_k)$.

Again, a higher mutual information indicates $\mathbf{z}_j$ contains a lot of information about $\mathbf{v}_k$, with maximal mutual information when there is a **deterministic** relationship between $\mathbf{z}_j$ and $\mathbf{v}_k$. When $\mathbf{v}_k$ is discrete, $0 \leq \mathbb{I}(\mathbf{z}_j; \mathbf{v}_k) \leq \mathbb{H}(\mathbf{v}_k)$, for entropy of $\mathbf{v}_k$ $\mathbb{H}(\mathbf{v}_k) = \mathbb{E}_{p(\mathbf{v}_k)}[-\log p(\mathbf{v}_k)]$, from equation 2.11 in section 2.4.1.

The **encoder channel's capacity** may be measured via the mutual information between $\mathbf{x}$ and $\mathbf{z}$, $\mathbb{I}(\mathbf{x};\mathbf{z})$, as outlined in section 2.4.2.

Chen *et al.* [2018] argue that **in order to improve disentangled representation, learning evaluation metrics that are more sensitive to subtle detail are needed to reveal deficiencies in existing methods**. They noted that the Higgins' metric lacks axis alignment detection, where only one latent factor relates to one generative factor. In addition, they claim that **classifier-based disentangled metrics tend to be ad hoc and sensitive to hyperparameters**. Such metrics may be **loosely** interpreted as the reduction in entropy of latent $\mathbf{z}$ if ground-truth factor $\mathbf{v}$ is observed.

Ground-truth factors can have high empirical mutual information with many latents. Normalising the empirical mutual information $\frac{\mathbb{I}(\mathbf{z}_j;\mathbf{v}_k)}{\mathbb{H}(\mathbf{v}_k)}$ axis-alignment is measured by the difference in MI between the two latents with the highest MI for each factor. This is called the mutual information gap (MIG). It is bounded by 0 and 1, estimated with an entire pass through the data, given in equation 2.24.

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{\mathbb{H}\left(\mathbf{v}_k\right)}\left(\mathbb{I}_n\left(\mathbf{z}_{j^{(k)}};\mathbf{v}_k\right)-\max_{j\neq j^{(k)}}\mathbb{I}_n\left(\mathbf{z}_j;\mathbf{v}_k\right)\right), \text{ where } j^{(k)}=\text{argmax}_j\mathbb{I}_n\left(\mathbf{z}_j;\mathbf{v}_k\right) \tag{2.24}$$

The MIG **heavily penalises latent factors with axes unaligned with the ground-truth factors**, where latents can contain information about two or more factors. It also **encourages compactness** in the latent representations since if one latent adequately captures a ground-truth factor, it is unnecessary for any other latent to be informative about the factor. Table 2.1 from Chen *et al.* [2018] shows merits of MIG when compared to to other common unsupervised disentangling metrics. The section where the metrics are covered is shown in brackets.

Metric	Axis	Unbiased
Higgins' metric (2.7.1)	No	No
Factor score (2.7.2)	Yes	No
MIG (2.8.2)	Yes	Yes

Table 2.1: Table from Chen *et al.* [2018] showing the merits of MIG when compared to other common disentangling metrics.

Chen *et al.* [2018] argue that MIG not only improves upon the Higgins' metric by **detecting axis alignment but is also generally applicable**, being limited only by whether mutual information can be estimated. They argued in favour of MIG, providing a comparison to alternatives, outlined in table 2.1.

2.8.3 Separated Attribute Predictability (SAP) score

Kumar *et al.* [2017] introduced the Separated Attribute Predictability (SAP) score as an alternative to the Higgins' metric. They observed that good performance in their score aligns better with qualitative disentanglement they observe during latent traversals when compared with the Higgins' metric. Furthermore, the SAP score does not involve training a classifier.

To calculate the SAP score for d latents and k generative factors, construct $d \times k$ score matrix $\mathcal{S}$, with entry $\mathcal{S}_{ij}$ being a score when the i^{th} factor is used to predict the jth latent. The nature of the score depends on whether the latent being predicted is discrete or continuous.

For continuous latent variables, the $\mathbb{R}^2$ score is given by equation 2.25.

$$\mathbb{R}^2 = \left(\frac{\text{Cov}\left(\left[\boldsymbol{\mu}_\phi(\mathbf{x}) \right]_i, \mathbf{y}_j \right)}{\sigma_{\left[\mu_\phi(\mathbf{x}) \right]_i} \sigma_{\mathbf{y}_j}} \right)^2 \tag{2.25}$$

The SAP score ranges from 0 to 1. A maximal score of 1 implies a linear function of the i^{th} latent explains **all** variation of the j^{th} generative factor.

When the target latents being predicted are discrete, $\mathcal{S}_{ij}$ is the balanced classification accuracy of the j^{th} generative factor. The classification score is obtained by fitting one or more real number thresholds directly on the i^{th} inferred latent minimising balanced classification error.

Each column in $\mathcal{S}$ is associated with a corresponding generative factor. The top two entries represent the top two most predictive latent dimensions. The difference between these top two entries is then computed. The mean of these differences then becomes the final SAP score. Note that this method only considers the top two entries. It is possible that other factors are captured by latents not considered by the score.

A high SAP score implies individual generative factors are primarily captured by only one latent dimension. This is the key focus of disentangling. Note that a high SAP score does not rule out the possibility of one latent dimension being captured by two or more generative factors well. This occurs when the factors themselves are correlated. A low SAP score does not rule out good disentanglement. This can occur when several latent dimensions are strongly correlated with a particular generative factor but poorly correlated with other factors. The SAP score is thus mostly readily interpreted when there is some correlation between latents that can be verified empirically using known ground-truth generative factors.

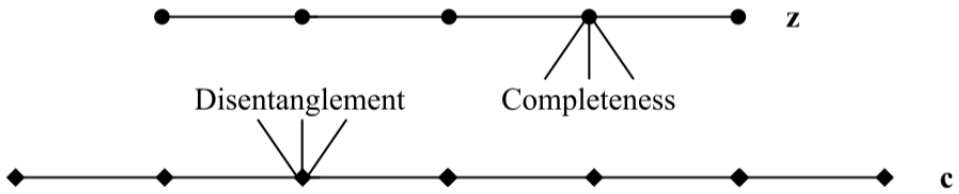


Figure 2.8: **Disentanglement score (DCI-D)** quantifies how many generative factors are captured by a **single latent**. Conversely, the **Completeness score (DCI-C)** quantifies how many latent variables are captured by a single generative factor. Both metrics quantify distortion of the ideal one-to-one mapping synonymous with disentangling in different directions. Figure from Eastwood and Williams [2018]. **Note the notation abuse in this figure.**

2.8.4 Disentanglement, completeness and informativeness (DCI)

A note on notation abuse in the section that follows

In outlining DCI in section 2.8.4, that follows, we abuse our notion to align with that of Eastwood and Williams [2018] and figure 2.8. Somewhat confusingly, latent codes (previously $\mathbf{z}$) are shown as $\mathbf{c}$ and generative factors (previously $\mathbf{v}$) are shown as $\mathbf{z}$.

Eastwood and Williams [2018] proposed a supervised framework for quantifying disentangled representations. They attempt to remove the need for visual inspection, which introduces subjectivity into the assessment. In order to do this, they introduce three scores:

- **Disentanglement score (DCI-D)**
- **Completeness score (DCI-C)**
- **Informativeness score (DCI-I)**

The **Disentanglement score (DCI-D)** in section 2.8.4 below, captures the extent to which the **learnt latent** $\mathbf{z}$ factorises (disentangles) the **underlying factors** of variation, with each variable capturing **at most** one ground-truth factor. DCI-D looks at **one underlying factor** and checks **how many latents** capture it.

The **Completeness score (DCI-C)** in section 2.8.4 below, on the other hand, captures the extent to which **each underlying factor** is captured by a **single latent variable**. DCI-C looks at **one latent** and checks **how many underlying factors** it captures. This is the opposite of the direction of the DCI-D. DCI-D considers if one latent representation $\mathbf{c}$ is captured by many generative factors $\mathbf{z}$; conversely, DCI-C considers if one generative factor $\mathbf{z}$ is captured by many latent representations $\mathbf{c}$.

DCI-D and DCI-C are shown in figure 2.8. Both scores quantify deviation from the ideal one-to-one mapping synonymous with disentangling but operate in different directions.

The **Informativeness score (DCI-I)** in section 2.8.4 below is the amount of information about the underlying factors of variation captured by the representation. Capturing information about these factors is needed to be useful for natural tasks in the data domain.

Eastwood and Williams [2018] detail the following steps to construct the **relative importance matrix**, from which DCI is calculated.

1. Train given model $\mathcal{M}$ on a synthetic dataset with generative factors $\mathbf{z}$

2. Retrieve latent $\mathbf{c}$ for each sample x in the dataset $(\mathbf{c} = \mathcal{M}(\mathbf{x}))$

3. Train regressor f to predict generative factor $\mathbf{z}$ given latent representation $\mathbf{c}(\hat{\mathbf{z}} = f(\mathbf{c}))$

4. Quantify f's deviation from the ideal mapping and the prediction error

In the above, K regressors are trained to predict each of the K generative factors. Regressor f_j predicts $\mathbf{z}_j$ given $\mathbf{c}$, such that $f_j(\mathbf{c}) : \mathbb{R}^D \to \mathbb{R}^1$. The regressors are used to construct a relative importance matrix $\mathcal{R}$ with $\mathcal{R}_{ij}$ denoting the relative importance of $\mathbf{c}_i$ when predicting $\mathbf{z}_j$. Using the relative importance matrix $\mathcal{R}$, Eastwood and Williams [2018] define each DCI criteria, outlined below.

Disentanglement score (DCI-D)

The score captures the extent to which the learnt latent $\mathbf{z}$ factorises (disentangles) the underlying ground-truth factors of variation. Since disentangling measurement is the goal, the score captures the extent that each variable captures only one ground-truth factor.

The disentanglement score of latent $\mathbf{c}_i$, $\mathcal{D}_i$, i given in equation 2.26.

$$\mathcal{D}_i = \left(1 - \mathbb{H}_K\left(P_i\right)\right), \tag{2.26}$$

where entropy $\mathbb{H}_K$, is defined in equation 2.27,

$$\mathbb{H}_K\left(P_{i.}\right) = -\sum_{k=0}^{K-1} P_{ik} \log_K P_{ik}, \tag{2.27}$$

and probability of $\mathbf{c}_i$ being important for predicting $\mathbf{z}_j$, $P_{i.}$, is given by equation 2.28.

$$P_{ij} = R_{ij} / \sum_{k=0}^{K-1} R_{ik} \tag{2.28}$$

When c_i is important in predicting only a single generative factor $\mathcal{D}_i$ is 1 . Conversely $\mathcal{D}_i$ is 0 when $\mathbf{c}_i$ is equally important in predicting **all** generative factors.

Completeness score (DCI-C)

The score captures the extent to which each of the **underlying generative factor** $\mathbf{z}$ are **captured by an individual latent** variable $\mathbf{c}$. The completeness score $\mathcal{C}_j$ for factor $\mathbf{z}_j$ is give by equation 2.29.

$$\mathcal{C}_j = \left(1 - \mathbb{H}_D\left(\tilde{P}_{.j}\right)\right), \tag{2.29}$$

where $\mathbb{H}_D$, is the entropy of the $\tilde{P}_{.j}$ distribution, is define by equation 2.30.

$$\mathbb{H}_D\left(\tilde{P}_{.j}\right) = -\sum_{d=0}^{D-1} \tilde{P}_{dj} \log_D \tilde{P}_{ij}. \tag{2.30}$$

When a single latent variable contributes to $\mathbf{z}_j$'s prediction, $\mathcal{C}_j$ is complete with a score of 1. Conversely when all codes equally contribute to $\mathbf{z}_j$'s prediction then $\mathcal{C}_j$ is maximally overcomplete with a score of 0.

Informativeness score (DCI-I)

This score captures the amount of **information about the underlying factors of variation captured by the representation**. Capturing information about these factors is needed for the representation to be useful for natural tasks in the data domain.

Prediction error $\mathbb{E}(\mathbf{z}_j, \hat{\mathbf{z}}_j)$ with appropriate error function $\mathbb{E}$ and $\hat{\mathbf{z}}_j = f_j(c)$ is averaged over the dataset and used to quantify how informative code c about generative factor $\mathbf{z}_j$. The prediction error, in turn, is dependent on the capacity of f.

The score also depends on the model's ability to explicitly represent information about $\mathbf{z}$ in c, which in turn is dependent on the model's ability to disentangle the underlying factors of variation ($\mathbf{z}$). There is thus overlap between the disentanglement and informativeness scores that depend upon f's capacity (no overlap when capacity is infinite).

2.8.5 Correlation between disentangling metrics

Locatello *et al.* [2019] examine the extent to which common disentangling metrics agree with one another. Figure 2.9 shows an adaptation of their pairwise scatter plots on a suite of disentangling metrics. All metrics are run on *dSprites*, a seminal disentangling dataset outlined in section 4.2.1, using a variety of disentangling models (some of which are not explored in our work).

They find that disentangling metrics are largely correlated, indicated by diagonal lines on the pairwise scatter plots, with MIG and DCI-D exhibiting the strongest covariation. They note that the Higgins' metric from section 2.7.1 and the factor score from section 2.7.2 measure similar notions of disentangling, motivating our decision to drop the Higgins' metric in our experiments outlined in chapter 5. Importantly, Locatello *et al.* [2019] note that **all** disentangling metrics we consider in this work appear to be correlated. The level, however, depends on the data sets.

2.8.6 Influence of random seeds in disentangling metrics

From section 2.8.5, disentangling metrics are largely correlated, and the level of correlation depends on the dataset. Locatello *et al.* [2019] note that **variation in disentangling scores** is primarily due to different **hyperparameters** and **random seeds**. They observe that variation from random seeds can matter more than the hyperparameter value, and a poor hyperparameter with a good random seed can beat a good hyperparameter with a bad random seed. They **qualitatively** conclude that, since scores heavily overlap for different models, the hyperparameter and random seed matter more than the objective function choice. They note that, while some models **do** achieve higher maximum scores than others, they observe **no consistent pattern of one model universally performing better than another**.

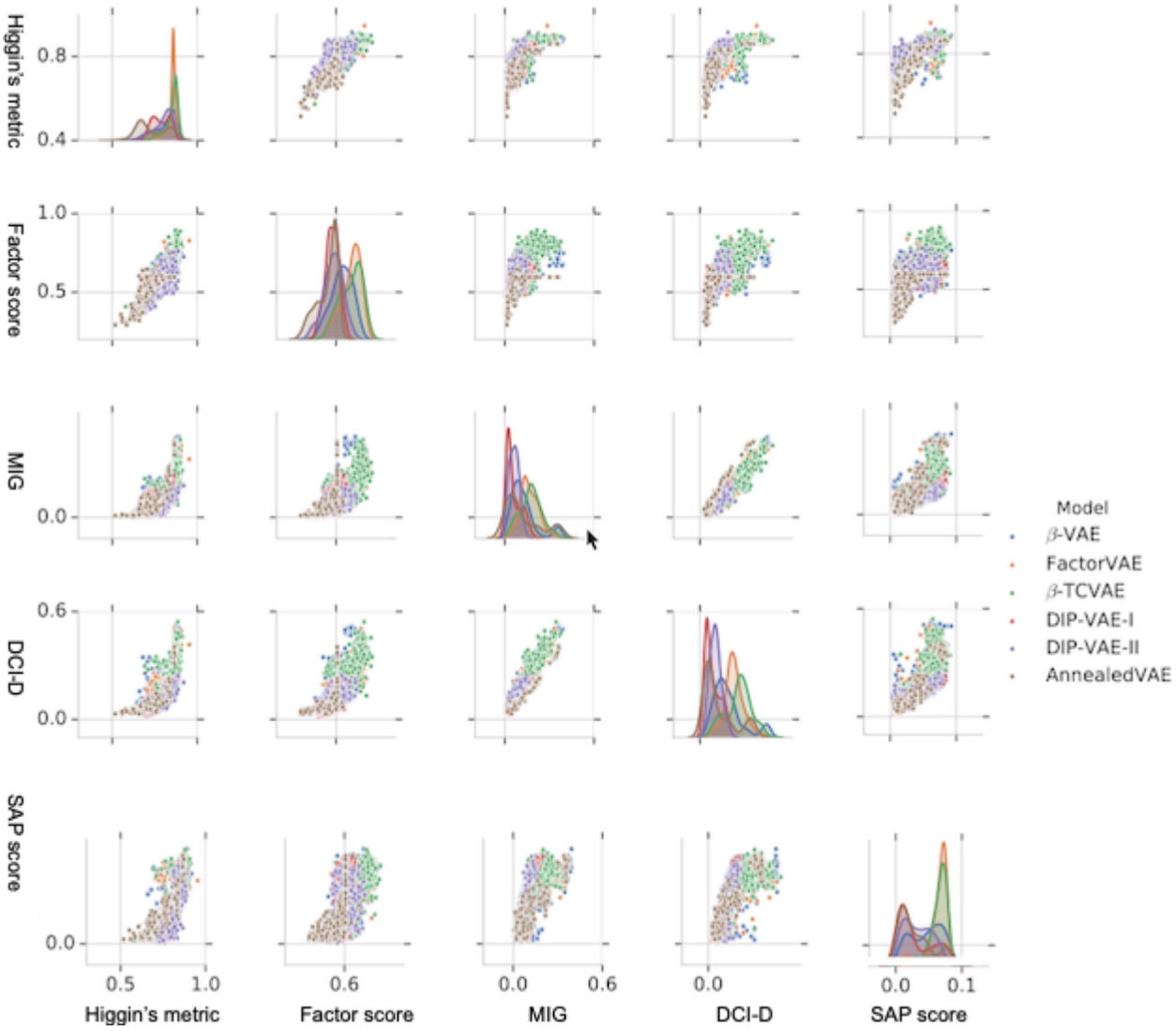


Figure 2.9: **Correlation between disentangling metrics:** Pairwise scatter plots of disentangling metrics on dSprites dataset from section 4.2.1, with the diagonals showing score variation. The figure shows a variety of disentangling models, each represented by different colours given on the right. It should be noted that some of these are not explored in our work. Locatello *et al.* [2019] show that disentangling metrics are largely correlated, with MIG and DCI-D exhibiting the strongest correlation. Figure adapted from Locatello *et al.* [2019].

2.8.7 Model selection with disentangling metrics

Notably, Locatello *et al.* [2019] advocate that model selection should not be driven by disentangling metric scores. This is because the disentangling scores rely on supervision, and tuning hyperparameters based on these scores would bias the model and defeat the purpose of purely unsupervised learning. In addition, they argue that fixing a hyperparameter sweep **a priori** and showing that one of those hyperparameters, with a specific random seed, achieves a good disentanglement score does not help with model selection.

In this chapter, we have outlined the background literature underpinning our work. In section 2.1, we outlined Bayesian models and variational inference (VI), a technique used to approximate difficult-to-estimate posteriors that frames inference as optimisation of an objective. In section 2.2, we then outlined variational auto-encoders (VAE), a popular framework for unsupervised representation learning that uses VI. In section 2.3 we outline the Evidence Lower BOund (ELBO), the VAE objective that is optimised for inference. In section 2.4 we then outlined the impact of the divergence choice in latent regularisation performed by the ELBO, detailing the Kullback–Leibler (KL) divergence. We note KL divergences asymmetry and the associated mode-seeking and mean-matching behaviour of the reverse and forward KL, respectively. In section 2.5, we outlined disentangling. Finally, we detailed the disentangling assessments, outlining latent traversals in section 2.6, supervised metrics in section 2.7, and finally unsupervised metrics in section 2.8.

With these necessary background concepts in mind, we proceed to describe our work's hypotheses, along with related research questions, in chapter 3 that follows.

Chapter 3

Hypotheses and research questions

A focus of our work is to examine the latent bias induced by using the constrained **reverse KL** $\boldsymbol{\beta} D_{\mathbb{KL}}\left(q_\phi(\mathbf{z} \mid \mathbf{x}) \| p_\theta(\mathbf{z})\right)$ as a latent regulariser. From section 2.4.1 we know that the reverse KL exhibits **mode-seeking** behaviour, resulting in an **under-dispersed** fitted variational posterior relative to the latent prior. From section 2.5.1, we know this constrained KL divergence is used as a latent regulariser in a key disentangling model, the $\boldsymbol{\beta}$ VAE. Specifically, we seek to assess whether this bias induces variational posteriors that inherit **local latent correlation**, failing to capture global latent correlation and whether this **negatively impacts disentangling.**

We introduce our *mSprites* collection of datasets in section 4.2. *mSprites* is designed to have varying levels of **generative space modality**. In representation learning, the aim is to have latent spaces that capture the generative space well, with a high degree of overlap. We believe that multimodality in the generative space impacts the learnt **latents' modality**. The resultant latents would thus be **more sensitive to mode-seeking behaviour** from the reverse KL latent regulariser, when compared to other datasets, an enhancement that is well suited to our study.

Furthermore, to test the latent bias induced in the $\boldsymbol{\beta}$ VAE, we introduce an alternative model, the $\boldsymbol{\beta\gamma}$ VAE in section 4.1. The $\boldsymbol{\beta\gamma}$ VAE includes an additional constrained forward KL latent regulariser $\boldsymbol{\gamma} D_{\mathbb{KL}}\left(p_\theta(\mathbf{z}) \| q_\phi(\mathbf{z} \mid \mathbf{x})\right)$. From section 2.4.1 we know that the **forward KL** exhibits **mean-seeking** behaviour, resulting in an **over-dispersed** fitted variational posterior, relative to the latent prior. We seek to assess if the introduction of a forward KL latent regulariser biases learning towards capturing **global latent correlations** and whether this **improves disentangling**.

We describe our work's hypotheses, along with related research questions, in sections in section 3.1 and 3.2 below.

3.1 Hypotheses

We hypothesise that using the **reverse KL** as a latent regulariser induces a bias towards under-dispersed, mode-seeking latent posterior solutions. Furthermore, we hypothesise that this induces variational posteriors in the β VAE and captures local latent correlations. In addition, we hypothesise that the focus of disentangling is to learn a **global isotropic relationships between latent factors** and that the solution's **under-dispersion bias obscures disentangling**.

In section 4.2 we introduce our *mSprite* collection of datasets, designed to have a multimodal generative space. Disentangling attempts to learn isotropic latents that capture this generative space. We hypothesise that this multimodal generative space impacts learnt latents' modality. We seek to assess whether multimodal generative space enhances or diminishes mode-seeking and mean-seeking latent bias and how this impacts disentangling.

With this in mind, we propose the following hypotheses regarding the disentangled representations learnt when fitting a variational posterior using the β VAE.

3.1.1 Multimodal generative spaces degrade β VAE's global overlap between the generative and latent space. This negatively impacts disentangling.

We hypothesise that because of the mode-seeking behaviour of the reverse KL in the β VAE outlined in section 2.4.2, the presence of multimodal generative spaces results in under-dispersed latents that focus on particular modes in the latent space. This results in latent correlation that **fails to capture latent correlation holistically, representative of all data** thus degrading global overlap between the latents and factors. This **poor global overlap** between latent and generative factor distributions **negatively impacts disentangling**.

We believe that this poor global overlap between latent and generative factors is thus a **source of latent entanglement**. This means that distinct latent dimensions fail to uniquely capture distinct generative factors in a one-to-one manner synonymous with disentangling.

In an attempt to address the aforementioned hypothesised pathology, our work explores the potential merit of augmenting $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta}}(\mathbf{x},\mathbf{z})$ to derive an alternative $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\gamma}(\mathbf{x},\mathbf{z})$, underpinning our $\boldsymbol{\beta\gamma}$ VAE, proposed in section 4.1. By including an additional constrained **forward KL** latent regularisation $\boldsymbol{\gamma}\mathcal{D}_{\mathbb{KL}}(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x}))$ we hypothesise that the learnt variational posterior **better captures global correlations** from the entire latent space, resulting a distribution is **more isotropic on a global basis and thus better disentangled**.

We hypothesise that such overlap means latents better capture information about the factors. This means that distinct latent dimensions better capture distinct generative factors in the one-to-one manner synonymous with disentangling. Furthermore, improved disentangling means improved correspondence between distinct latents and factors, improving interpretability.

3.1.2 $\beta\gamma$ VAE provides robust global generative and latent space overlap. This improves disentangling.

We hypothesise that our proposed $\beta\gamma$ VAE, outlined in section 4.1, may (at least partially) remedy the pathologies hypothesised in section 3.1.1, above.

We hypothesise that the γ weighted forward KL term provides a **mean-matching** behavioural bias that **counteracts** the **mode-seeking** behaviour of the reverse KL, particularly in the presence of multimodal generative space. This means that global overlap between the latent and factor distributions is encouraged more by the $\beta\gamma$ VAE despite data drawn from a multimodal generative space. We believe that this improved global overlap provides for more robust disentangling in the $\beta\gamma$ VAE.

We hypothesise that the γ constraint improves global overlap and disentangling, despite the fact that data is drawn from a multimodal generative space.

3.2 Research questions

The following research questions have been developed in an attempt to test the hypotheses outlined in the previous section 3.1.

3.2.1 Do multimodal generative spaces distort global latent fit in the β VAE?

Given our made-for-purpose group of datasets *mSprites* from section 4.2, what impact does increased modality in the data's generative space have on the global overlap between generative factors and latents? Recall from section 2.8.1 that mutual information (MI) measures the shared informational content between the latent and the generative factors. We wish to measure the impact on shared informational content, as measured by MI, in the presence of generative modality, i.e. when moving from *mSprites: uniform* to *mSprites: 2-mode*. Furthermore we will to assess the impact on MI of additional generative space modality, i.e. when moving from *mSprites: 2-mode* to *mSprites: 4-mode*, and finally to *mSprites: 8-mode*.

3.2.2 Do multimodal generative spaces distort disentangling in the β VAE?

When applying our host of disentangling metrics, outlined in section 2.6, we wish to measure the impact of the presence of generative modality, i.e. when moving from *mSprites* to *mSprites: uniform*. Furthermore we will to assess the impact of additional generative space modality, i.e. when moving from *mSprites: 2-mode* to *mSprites: 4-mode*, and finally to *mSprites: 8-mode*.

3.2.3 Does our proposed $\beta\gamma$ VAE provide robust global latent fit when compared to the β VAE? Is this impacted by multimodal generative spaces?

Does our proposed $\beta\gamma$ VAE improve mutual information (MI) from section 2.8.1, when compared to the β VAE? This would indicate that the overall overlap between the latent and generative factor distributions has improved.

3.2.4 Does our proposed $\beta\gamma$ VAE provide robust disentangling when compared to the β VAE. Is this impacted by multimodal generative spaces?

When applying our host of disentangling metrics, outlined in section 2.6, does our $\beta\gamma$ VAE disentangle more robustly, improving metrics, when compared to the traditional β VAE?

We now propose methods designed to investigate our research questions in the following section.

Chapter 4

Proposed methods

Having outlined our hypotheses and research questions in chapter 3, we now online the methods we propose to answer these questions.

In section 2.2 we derive the VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$. Importantly, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$ enforces shape in the latent space by minimising latent regulariser $D_{\mathbb{KL}}\left(q_\phi(\mathbf{z})\|p_\theta(\mathbf{z})\right)$, from equation 2.9: reverse KL divergence between variational posterior $q_\phi(\mathbf{z}\mid\mathbf{x})$ and prior distribution $p_\theta(\mathbf{z})$.

We detail the asymmetry of the KL divergence in section 2.4.2, and note the forward and reverse KL. Optimising either (forward or reverse) KL divergence induces different behaviour and results in markedly different variational posterior solutions. Importantly, optimising the reverse KL induces mode-seeking behaviour, while optimising its converse, the forward KL, induces mean-matching behaviour. Furthermore, optimising the reverse KL results in solutions that are biased towards under-disperse distributions, while optimising the forward KL biases the solution towards over-dispersed distributions.

The VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$, uses the **reverse KL** latent regulariser, $D_{\mathbb{KL}}\left(q_\phi(\mathbf{z}\mid\mathbf{x})\|p_\theta(\mathbf{z})\right)$. We hypothesise that this biases posterior $q_\phi(\mathbf{z}\mid\mathbf{x})$ towards underdispersed distributions in the latent space. Furthermore we hypothesise that this solution is biased to capture **localised correlations** associated with a **specific regions in the latent space**.

In section 2.5 we introduce the concept of disentangling, outlining a prominent disentangling model, the $\boldsymbol{\beta}$ VAE. The $\boldsymbol{\beta}$ VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta}}(\mathbf{x},\mathbf{z})$, enhances the latent regularisation by introducing a $\boldsymbol{\beta}>0$ factor, applied to the reverse KL latent regulariser. The resultant constrained latent regulariser is thus $\boldsymbol{\beta}D_{\mathbb{KL}}\left(q_\phi(\mathbf{z}\mid\mathbf{x})\|p_\theta(\mathbf{z})\right)$, from equation 2.20. This has been shown to encourage disentangling. Importantly the **reverse** KL divergence is still used and the solution continues to be biased towards localised correlations in the latent space, perhaps even more so.

Our work explores the potential merit of augmenting $\mathcal{L}_{\theta,\phi,\beta}(\mathbf{x},\mathbf{z})$ to derive an alternative $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\gamma}(\mathbf{x},\mathbf{z})$. This is achieved by including the addition of constrained

forward KL latent regularisation from $\gamma\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big)$. The forward KL is expected to counteract the mode-seeking behaviour of the reverse KL, altering the latent posterior learnt.

We hypothesise that the **under-dispersion bias of the β VAE's variational posterior biases solutions toward local latent correlations**. We argue that the focus of disentangling is to learn global isotropic relationships between latent factors and that the solution's under-dispersion bias, inherited from the reverse KL latent regularisation, obscures this. By including a **constrained forward KL latent regularisation** we hypothesise that the learnt variational posterior **better captures correlations from the entire latent space**, resulting a distribution is more isotropic on a global basis and thus better disentangled.

This chapter details our key contributions beginning with our proposed $\beta\gamma$ VAE in section 4.1. Our work also creates a series of datasets with multimodal generative spaces to better test our hypotheses. We propose a **collection** of supervised synthetic datasets called *mSprites* in section 4.2. Each dataset in the *mSprites* collection of datasets contains a different number of spatial modes, allowing for latent mode-seeking behaviour at different calibrations to be examined. The collection of datasets is supervised with known ground truths, allowing for the easy quantification of the disentangling metrics outlined in sections 2.7 and 2.8.

The remainder of this chapter details our proposed methods, namely the $\beta\gamma$ VAE in section 4.1 and the *mSprites* dataset in section 4.2.

4.1 Proposed $\beta\gamma$ VAE

In this section we derive our proposed $\beta\gamma$ VAE. The model augments the β VAE's $\mathcal{L}_{\theta,\phi,\beta}(\mathbf{x},\mathbf{z})$ to derive an alternative $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\gamma}(\mathbf{x},\mathbf{z})$. To achieved this, we include an additional constrained **forward KL** latent regularisation from $\gamma\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big)$. The forward KL is expected to counteract the mode-seeking behaviour of the reverse KL. We hypothesise that this induces a latent space that is beneficial in disentangled representation learning. The derivation of $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\gamma}(\mathbf{x},\mathbf{z})$ is outlined below.

4.1.1 Extending β VAE's $\mathcal{ELBO}(q)$

Sason and Verdu [2016] note that the $\mathcal{ELBO}(q)$, the lower bound of the model evidence, it is an upper bound on the forward divergence.

We derive the $\beta\gamma$ VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\gamma}(\mathbf{x},\mathbf{z})$, by extending the β VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta}}(\mathbf{x},\mathbf{z})$ from equation 2.20. We include the addition of the term $-\gamma\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big)$.

$$\mathcal{L}_{\theta,\phi,\beta,\gamma}(\mathbf{x},\mathbf{z}) = \mathcal{L}_{\theta,\phi,\beta}(\mathbf{x},\mathbf{z}) - \boldsymbol{\gamma}\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big) \tag{4.1}$$

i.e.

$$\mathcal{L}_{\theta,\phi,\beta,\gamma}(\mathbf{x},\mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\big[\log p_\theta(\mathbf{x}\mid\mathbf{z})\big] - \Big(\boldsymbol{\beta}\mathcal{D}_{\mathbb{KL}}\big(q_\phi(\mathbf{z}\mid x)\|p_\theta(\mathbf{z})\big) + \boldsymbol{\gamma}\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big)\Big), \tag{4.2}$$

where $\boldsymbol{\beta}, \boldsymbol{\gamma} > 0$.

For a more detailed derivation of the VAE's $\mathcal{ELBO}(q)$, $\mathcal{L}_{\theta,\phi}(\mathbf{x},\mathbf{z})$ we refer the reader back to section 2.3.

From section 2.4.1 we know that, the KL divergence is strictly positive, $\mathcal{D}_{\mathbb{KL}}\big(p_\theta(\mathbf{z})\|q_\phi(\mathbf{z}\mid\mathbf{x})\big) > 0$. Since $\boldsymbol{\gamma} > 0$ we thus have

$$\mathbb{E}_q\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right] > \mathcal{L}_{\theta,\phi,\beta}(\mathbf{x},\mathbf{z}) > \mathcal{L}_{\theta,\phi,\beta,\gamma}(\mathbf{x},\mathbf{z}) \tag{4.3}$$

$\mathcal{L}_{\theta,\phi,\beta,\gamma}(\mathbf{x},\mathbf{z})$ thus forms a lower bound on $\mathcal{L}_{\theta,\phi,\beta}(\mathbf{x},\mathbf{z})$, which is itself a lower bound on the model evidence $\mathbb{E}_q\left[\log p_\theta(\mathbf{x}\mid\mathbf{z})\right]$. $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x},\mathbf{z})$ thus bounds all quantities of interest from below and, for our work, may be thought of as an Evidence Low**est** BOund. We outline a schematic for the computation flow and algorithm when training our proposed $\boldsymbol{\beta\gamma}$ VAE in section 4.1.2 that follows.

4.1.2 Proposed training loop

The training loop for the proposed $\boldsymbol{\beta\gamma}$ VAE is detailed in algorithm 1. The algorithm details a doubly-stochastic optimisation of the $\mathcal{ELBO}(q)$ proposed in equation 4.2 $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x},\mathbf{z})$. The two sources of stochasticity come from noise inherited from the mini-batch sampling process of stochastic gradient decent and from sampling $p(\cdot)$.

The corresponding schematic of the computational flow for algorithm 1 is adapted from Kingma and Welling [2019] and presented in figure 4.1. [1]

4.1.3 Related work

Work that examines the impact of mismatch between the generative and latent spaces in VAEs is related to our work. Davidson *et al.* [2018] note that the default Gaussian prior and posterior in VAEs fail to model data with a hyperspherical latent structure.

[1] In reality mini-batches of data are used in training, rather than a single datum presented for illustration purposes

Algorithm 1 Stochastic optimisation of $\mathcal{ELBO}(q)$ is the vanilla VAE training loop from Kingma and Welling [2019], with **our proposed** $\mathcal{L}_{\theta,\phi,\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{x},\mathbf{z})$. Noise originates from both mini-batch sampling and sampling of $p(\cdot)$, doubly-stochastic optimisation procedure

Data:
$\mathcal{D}$: Dataset
$q_\phi(\mathbf{z} \mid \mathbf{x})$: Inference model (tractable variational posterior)
$p_\theta(\mathbf{x},\mathbf{z})$: Generative model
Result:
θ, ϕ: Learned parameters
θ, $\phi \leftarrow$ Initialise parameters
while SGD **not converged do**
$\mathcal{M} \sim \mathcal{D}$ (Random mini-batch of data)
$\epsilon \sim p(\epsilon)$ (Random noise for every datapoint in $\mathcal{M}$)
Compute $\tilde{\mathcal{L}}_{\theta,\phi,\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathcal{M},\epsilon)$ and its gradients $\nabla_{\theta,\phi}\tilde{\mathcal{L}}_{\theta,\phi,\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathcal{M},\epsilon)$
Update θ and ϕ using SGD optimiser
end while

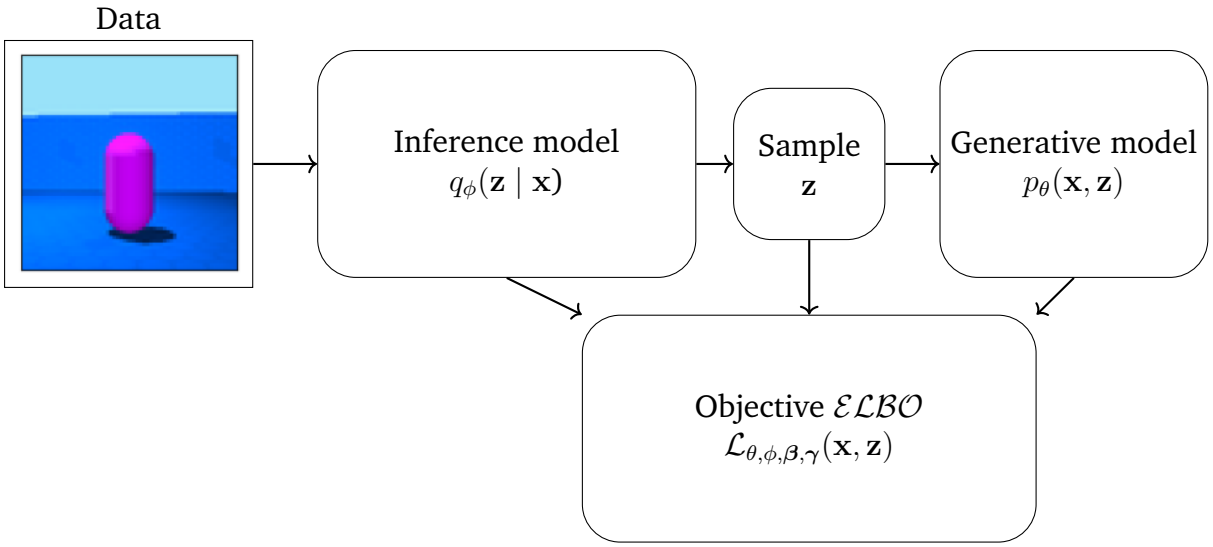


Figure 4.1: Schematic of computational flow of algorithm 1

This is problematic for data types better explained through spherical representations, examples of which include directional data, protein structure and observed wind directions. They address this with their proposed a hyperspherical VAE, that uses *von Mises-Fisher* distributions for their VAE's prior and posterior. They show that this leads to hyperspherical, rather than a hyperplanar, latent spaces. This is more suitable for capturing data with a hyperspherical latent structure, outperforming Gaussian VAEs. This work indicates how changing the underlying variational distribution may induce latent structures that conforms better to the true data generative structures, a focus of our work.

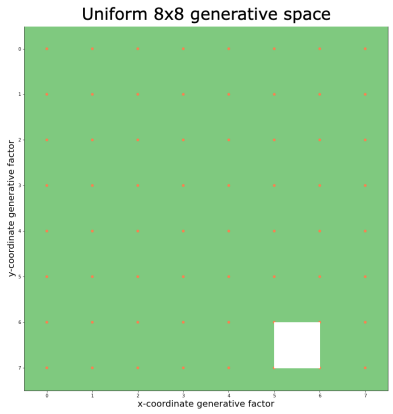
Choi and Min [2022] examine VAE-based trajectory forecasting models and observe that, problematically, the generated trajectory is located between adjacent lanes. They liken this to generated samples being blurry in image reconstruction and synthesis tasks. This is addressed in their work by introducing a hierarchical latent structure. Their work assumed that trajectory distribution are approximated as a mixture of unimodal distributions. They introduce **low-level latents** to model each trajectory distribution and **high-level latents** to model the mixtures of these trajectories. This work relates to ours, in that it explores better matching between the generative and latent space. However, it is worth noting that the definition of disentanglement we use in section 2.5, where a **particular latent unit** is sensitive to changes in particular generative factor, but largely invariant changes in other factors, while simple, does not allow for correlations between latent and generative factors $\mathbf{v}$, nor hierarchies amongst them. However, an alternative definition of disentangling in the presence of hierarchies would have to be defined. Without such a definition, hierarchical latent structures thus seem inappropriate when exploring disentangling, a focus of our work.

Connor *et al.* [2021] note that when the prior structure does not match that of the true data generative space it can lead to data being modelled less accurately. Their VAE with Learned Latent Structure has a learnable manifold model in the latent space, allowing for nonlinear manifold structure from the data to be learnt and use that structure to define a latent prior. This may be an solution to the mismatch between generative and latent spaces explored in our work, but is not explored in our work.

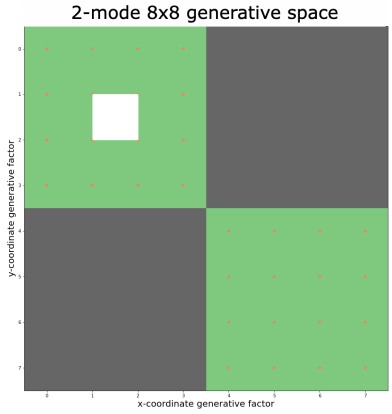
4.2 Proposed *mSprites* group of datasets

Recall from section 2.5 that we assume data $\mathbf{x}$ is generated from a simulator process containing ground truth factors $(\mathbf{v}, \mathbf{w})$. Additionally we assume at least some of these ground truth factors $\mathbf{v}$ are independent with given interpretable labels, $\mathbf{w}$ being the other non-interpretable factors. The data generation process is thus defined as $\mathbf{x} \sim \mathrm{Sim}(\mathbf{v}, \mathbf{w})$, for simulation operator Sim.

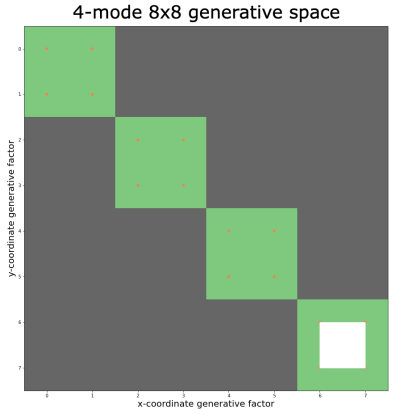
Currently, there appear to be no public datasets present in literature that are specifically designed to explore **disentangling** when the (known) generative factors are dis-



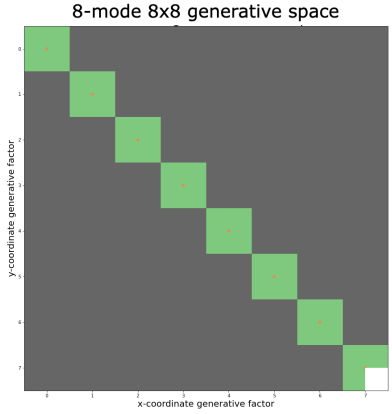
(a) *mSprites uniform* generative factor dataset



(b) *mSprites 2-mode* generative factor dataset



(c) *mSprites 4-mode* generative factor dataset



(d) *mSprites 8-mode* generative factor dataset

Figure 4.2: ***mSprites* collection of multimodal generative factor disentangling datasets:** White square sprites moves around the image. Its location is limited by modality in the spatial generative space. Figures 4.2a, 4.2b, 4.2c and 4.2d show the *uniform*, *2-mode*, *4-mode* and *8-mode* generative factor datasets respectively. The green regions are multimodal generative factors are uniformly sampled to create the final $\mathbf{x}$ and $\mathbf{y}$ spatial positions. The white region shows examples of where the sprite location may be sampled from. The red dots are potential spatial coordinates that the sprite may be on.

tinctly **multimodal**. The typical disentangling dataset is *dSprites*, explored in section 4.2.1. This involves sprites with spatial motion, moving around an image, with **all possible pairs** of $\mathbf{x}$-locations and $\mathbf{y}$-locations. The dataset is generative and the spatial locations in *dSprites* are sampled from generative factors. Since the sprites move through all possible $\mathbf{x}$-locations and $\mathbf{y}$-locations, the generative factors that give rise to these locations are thus fully traversed and **uniform**.

With this in mind we propose a group of four datasets collectively called *mSprites*[2], illustrated in figure 4.2. Our generative datasets have a square sprite moving around an image, however we limit the movement by creating modality in the generative factor space that is sampled for spatial positioning. Future work could improve on these datasets and the location and scale of each generative space mode could have more variation.

In representation learning the aim is to have latent spaces that capture the generative space well, with a high degree of overlap between these spaces. We hypothesise that **multimodality in the generative space** will impact the learnt **latents' modality**. The resultant latents would thus be **more sensitive to latent regularisation bias**.

When the **forward KL** is used for latent regularisation, this will exacerbate **mean-matching bias** in the latent space. When the **reverse KL** is used for latent regularisation, this will exacerbate **mode-seeking bias** in the latent space. In particular, disentangling attempts to learn isotropic latents that capture this (multimodal) generative space. The dataset is designed to assess what impact multimodal generative spaces have when learning representations.

The *mSprites* group of datasets consists of four datasets, each with varying number of spatial generative factor modes. Each dataset consists of 8×8 square generative space. Our images are 64×64 and a 8×8 square sprite object moves around a 64×64 black image following equations 4.4 and 4.5.

We define our $\mathbf{x} \sim \mathrm{Sim}(\mathbf{v}, \mathbf{w})$ by specifying the relationship between the data's sprite x, y spatial coordinates in the generated image, $\mathbf{x}_{x-location}$ and $\mathbf{x}_{y-location}$, and their corresponding generative factors $\mathbf{v}_{x-location}, \mathbf{v}_{y-location}$ as:

$$\mathbf{x}_{x-location} = 8\mathbf{v}_{x-location} \tag{4.4}$$

$$\mathbf{x}_{y-location} = 8\mathbf{v}_{y-location} \tag{4.5}$$

[2]"m" for multimodal generative space

Importantly, for each dataset we include the same number of spatial modes in **both** generative factor $\mathbf{v}_{x-location}$ and $\mathbf{v}_{y-location}$. For example the *2-mode* generative factor dataset has 2 modes for the x-locations and two 2 modes for the y-locations. The *4-mode* generative factor dataset has 4 modes for the x-locations and 4 modes for the y-locations. Finally, the *8-mode* generative factor dataset has 8 modes for the x-locations and 8 modes for the y-locations.

mSprite: uniform generative factor dataset

The *mSprite: uniform* generative factor dataset, shown in figure 4.2a does **not limit spatial movement** of the sprite. Both the x-location and y-location generative factors may be sampled from the green region in figure 4.2a. As such, both the x-location and y-location generative factors may be sampled from the all 8×8 coordinates available in the whole frame.

mSprite: 2-mode generative factor dataset

The *mSprite: 2-mode* generative factor dataset, shown in figure 4.2b, **limits the spatial movement** of the spite, by limiting what values of the generative factors may be sampled. Both the x-location and y-location generative factors may be sampled from the green regions in figure 4.2b. As such, the generative factors may only be sampled from **two** 4×4 **regions** in the generative factor space.

mSprite: 4-mode generative factor dataset

The *mSprite: 4-mode* generative factor dataset, shown in figure 4.2c, **limits spatial movement** of the sprite, by limiting what values of the generative factors may be sampled. Both the x-location and y-location generative factors may be sampled from the green regions in figure 4.2c. As such, the generative factors may only be sampled from **four** 2×2 **regions** in the generative factor space.

mSprite: 8-mode generative factor dataset

Finally, the *mSprite: 8-mode* generative factor dataset, shown in figure 4.2d, **limits spatial movement** of the sprite, by limiting what values of the generative factors may be sampled. Both the x-location and y-location generative factors may be sampled from the green regions in figure 4.2d. As such, the generative factors may only be sampled from **eight** 1×1 **regions** in the generative factor space.

We now explore related disentangling datasets, giving details on the other datasets that will also be used in our work.

4.2.1 Related work

Most disentangling research is typically focused on handful of synthetic datasets with known factors of variation that typically vary incrementally over the entire concept

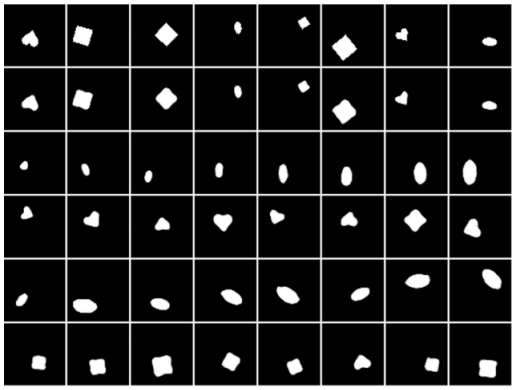


Figure 4.3: Matthey *et al.* [2017]'s *dSprites* disentangling dataset. 737,280 procedurally generated 2D shapes from 6 ground truth independent latent factors: x and y spatial factors and colour, shape, scale and rotation.

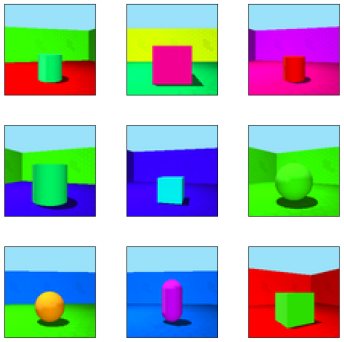
space in a uniform manner. A summary of the datasets considered is given in the table from Gondal *et al.* [2019] and is extended to include our synthetic data. All datasets explored in this work are in **bold**.

Dataset	Factors of Variation	Resolution	Number of Images
mSprites: uniform	**2**	**64 ×64**	**64**
mSprites: 2-mode	**2**	**64 ×64**	**64**
mSprites: 4-mode	**2**	**64 ×64**	**64**
mSprites: 8-mode	**2**	**64 ×64**	**64**
dSprites	**6**	**64 ×64**	**737,280**
Colored dSprites	6	64 ×64	737,280
Noisy dSprites	6	64 ×64	737,280
Scream dSprites	6	64 ×64	737,280
SmallNORB	5	128 ×128	48,600
Cars3D	3	64 ×64	17,568
3D shapes	**6**	**64 ×64**	**480,000**
MPI3D-toy	7	64 ×64	1,036,800
MPI3D-realistic	7	256 ×256	1,036,800
MPI3D-real	7	512 × 512	1,036,800

Table 4.1: A summary disentangling datasets from Gondal *et al.* [2019], extended to include our synthetic data. **All datasets considered in this work are in bold.**

Locatello *et al.* [2019] suggest that experiments should be on data sets of varying degrees of difficulty. While all datasets in table 4.1 are related to our *mSprites* dataset, since they supervised disentangling datasets, the other datasets chosen in this work are *dSprites* and *3D shapes*. The reason for this selection is that there is an increase in visual complexity moving from *mSprites*, to *dSprites* and finally to *3D shapes*. While both *dSprites* and *3D shapes* have the same number of factors of variation, *3D shapes* is a much more difficult dataset, displaying object complexity, colour and perspective. We now discuss these datasets.

Figure 4.4: *3D shapes* from Burgess and Kim [2018] procedurally generated from 6 ground truth independent latent factors: floor colour, wall colour, object colour, scale, shape and orientation.



dSprites dataset

The primary influence for the *mSprites* group of datasets is Matthey *et al.* [2017] *dSprite* dataset. The dataset consists of $737,280$ procedurally generated 2D shapes from 6 ground truth independent latent factors: colour, shape, scale, rotation, $\mathbf{x}$ and $\mathbf{y}$ of sprite positions. All possible combinations of these latents are present exactly once. Latents are varied one at a time. Latents values are chosen to have the smallest step changes while ensuring different pixel outputs. No noise was added. The dataset is similar to our multimodal dataset with generative factors $\mathbf{x}$ and $\mathbf{y}$ for sprite spatial positions. However, its generative factors are **not multimodal** and it includes more factors of variation. The dataset is illustrated in figure 4.3.

3D shapes dataset

Burgess and Kim [2018]'s *3D shapes* dataset, like the *mSprites* group of datasets and *dSprites*, is a supervised disentangling research dataset. The dataset consists of $64 \times 64 \times 3$ images of *3D shapes*, procedurally generated from 6 ground truth independent latent factors: floor colour, wall colour, object colour, scale, shape and orientation. Again, its generative factors are **not multimodal** and it includes more factors of variation. The dataset is illustrated in figure 4.4.

In this chapter we introduced our key contributions. The first of these is $\beta\gamma$ VAE. In addition, we outlined *mSprites*, a supervised toy disentangling dataset, designed to have multimodal generative spaces, along with more difficult datasets that will be explored in our study. The proposed methods outlined in this chapter are used to answer the research questions associated with the hypotheses from the previous chapter. We now discuss our experimental method in chapter 5, that follows.

Chapter 5

Experimental methods

We begin exploring our experimental methods by detailing our experimental datasets in section 5.1. We then focus on the models explored in our work in 5.2, which **share the same architecture and differ only in their** $\mathcal{ELBO}(q)$. The network architecture used throughout the study is then outlined in section 5.3, which is heavily influenced by that used by Burgess *et al.* [2018].

Locatello *et al.* [2019] note that experiments should be explicit about the role of inductive biases and (implicit) supervision, investigate concrete benefits of enforcing disentanglement of the learned representations, and consider a reproducible experimental setup covering several data sets. This motivates our experimental setup that follows in this chapter.

Implementations from Abdi *et al.* [2019] and Michlo [2021] were used as references for base models and metrics.

5.1 Data

Locatello *et al.* [2019] suggest that, in future research, experiments should be on data sets of varying degrees of difficulty. With this in mind, our study examines the following datasets from section 5.1:

- *mSprites: uniform* generative factor space dataset from section 4.2
- *mSprites: 2-mode* generative factor dataset from section 4.2
- *mSprites: 4-mode* generative factor dataset from section 4.2
- *mSprites: 8-mode* generative factor dataset from section 4.2
- *dSprites* from section 4.2.1
- *3D shapes* from section 4.2.1

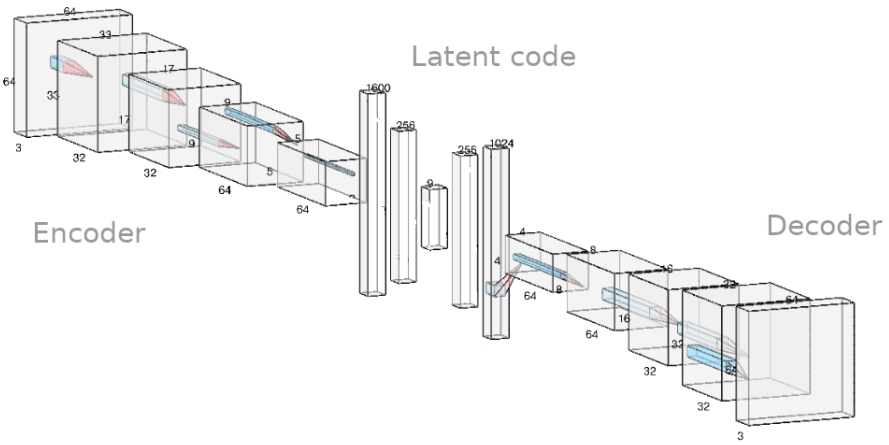


Figure 5.1: VAE architecture used, same to Higgins *et al.* [2017]. Original figure, constructed for this report.

The *mSprites* collection of datasets is initially used to explicitly test the performance of the considered models from section 5.2 on data sampled from multi-modal generative spaces. The factors of variation are low for these datasets, and they are specifically designed to test the impact of the multi-modal generative factors, **without** further nuisance factors. We then conduct further experiments with the popular *dSprites* and *3D shapes* datasets, which are more difficult than *mSprites*, being larger with more factors of variation.

5.2 Models

Our study confines itself to the two models described in chapters 2 and 4, namely:

- β VAE from section 2.5.1
- $\beta\gamma$ VAE from section 4.1

Importantly the β VAE may be recovered from the $\beta\gamma$ VAE if we include $\gamma = 0$ in our $\beta\gamma$ VAE hyperparameter sweep and effectively examine both models.

5.3 Architecture

The architecture is the same as Higgins *et al.* [2017], using the convolutional encoder and decoder outlined in their seminal disentangling work. A figure of our network architecture is illustrated in 5.1. The same architecture is used in all experiments, as we are focused on the impact of changing the $\mathcal{ELBO}(q)$ objective in our work.

The encoder for the VAEs consists of 4 convolutional layers, with 32 or 64 channels, 4×4 or 2×2 kernels and a stride of 2. This is followed by 2 fully connected layers of 256 units. The latent distribution consisted of one fully connected layer of 9 units parameterising the mean and log standard deviation of 9 Gaussian random variables.

The decoder architecture is broadly the transpose of the encoder. Rectified Linear Unit (ReLU) activations were used throughout.

5.4 Initialisation

Latent variables are assumed to follow a normal distribution with weights instantiated via *Xavier initialisation* proposed by Glorot and Bengio [2010].

5.5 Training length and random seeds

Each epoch consisted of 57600 steps. Runs were done *at least* 10 times and averaged out for each β-γ hyperparameter combination to account for stochasticity that is present, such as the mini-batching in the optimisation process and model reparameterisation trick.

From section 2.8.6, we are aware that random seeds can play a big role in the disentangling score variation. With this in mind, more random seeds per hyperparameter calibration are preferred. However, due to computational limits, our work has a minimum of 10 random seeds per hyperparameter combination.

5.6 Stochastic optimiser

Liu *et al.* [2019] note that typically learning rate warm-up heuristics achieve remarkable success in stabilising training, accelerating convergence and improving generalisation for adaptive stochastic optimisation algorithms like RMS Prop and Adam. They identify a problem of adaptive learning rates, namely problematically large variances in the early stages of training, and propose that warm-up works as a variance reduction technique. Their proposed variant of Adam, RAdam, introduces a term to rectify the variance of the adaptive learning rate.

We used RAdam in our work, finding via empirical experimentation that a learning rate of $5e^{-4}$ converges well.

5.7 Evaluation

To assess the relative performance of the various models and datasets outlined, we propose using all metrics outlined in section 2.6. These include:

- Factor score from section 2.7.2
- Mutual information (MI) from section 2.8.1
- Mutual information gap (MIG) from section 2.8.2
- Separated Attribute Predictability (SAP) from section 2.8.3
- Disentangling score (DCI-D) from section 2.8.4
- Completeness score (DCI-C) from section 2.8.4
- Informativeness score (DCI-I) from section 2.8.4

5.8 Finding an appropriate baseline calibration

Assessing appropriate hyperparameter scale

Here we perform exploratory hyperparameter sweeps to tune for the appropriate hyperparameter scale. All datasets described in section 5.1, above, are tested for *both* the β VAE and the $\beta\gamma$ VAE.

An initial sweep is performed to assess reconstruction performance at different β-γ hyperparameter scales. The β-γ scales for both β and γ tested are $10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 0, 10$ resulting in 36 pairs **per dataset, per model**. The intention here is to broadly assess which hyperparameter scale is the most suitable combination to produce high-fidelity reconstructions. Setting up the sweep in this manner allows both the β VAE and β-γ VAE to be examined in one sweep.

Latent traversals are used to qualitatively assess the reconstructions formed via inspection. A large number of hyperparameter combinations produced traversals with blank reconstructions. It is precisely these hyperparameter combinations we wish to avoid with this exercise. Values of 10^{-3} produced suitable results for all datasets, with 10^{-2} also producing good results for the *3D shapes* dataset. It should be noted that other scale combinations resulted in blank reconstructions.

Given the robustness of the 10^{-3} scale, we refine the sweep space to $[0, 25, 50, 75] * 10^{-3}$. All 16 combinations of β-γ are investigated in the findings that follow in chapter 6.

Importantly we consider the reconstruction without disentangling scores when choosing the hyperparameter sweep scale. Disentangling metrics are *not* considered when choosing the hyperparameters, as this would introduce bias into our experimental setup.

Assessing appropriate latent size

The latent space size $\mathbf{z}$ should be large enough to capture *all* independent, interpretable ground-truth factors $\mathbf{z}$. Referring to table 4.1 we note that the 3D shapes dataset has 6 factors of variation. With this in mind, we chose a latent size of 9 as it was deemed to have sufficient excess capacity to capture 6 factors of variation, and the same latent size is used throughout.

Chapter 6

Findings and discussion

Having outlined our experimental methods in chapter 5, here we discuss our study's key findings. A total of 179 days of compute were required to produce the following findings. 951 experiments were run on single NVIDIA Tesla V100s. All computations were performed using the high-performance computing infrastructure provided by the Mathematical Sciences Support unit at the University of the Witwatersrand.

In this chapter, we assess the research questions outlined in section 3.2, namely:

- 3.2.1 Do multimodal generative spaces distort **global latent fit** of β VAE?
- 3.2.2 Do multimodal generative spaces distort **disentangling** of β VAE?
- 3.2.3 Does our proposed $\beta\gamma$ VAE provide robust **global latent fit** when compared to the β VAE. Is this impacted by multimodal generative spaces?
- 3.2.4 Does our proposed $\beta\gamma$ VAE provide **robust disentangling** when compared to the β VAE. Is this impacted by multimodal generative spaces?

We find that when using the **reverse KL** for latent regularisation, multimodal generative spaces largely **distort the overall information content** captured by the learnt representation, distorting the global fit of the learnt representation. The reduced global fit from increased generative space modality in *mSprites* is related to the mode-seeking behaviour of the reverse KL.

Our work shows that when adding an additional constrained **forward KL** for latent regularisation, as done in our $\beta\gamma$ VAE, this distortion in the global overlap between latents and generative factors, from increased generative space modality, is remedied by the forward KL's mean-matching behaviour. The impact of multimodal generative spaces on disentangling is less clear. We provide evidence that **multimodal generative spaces *can* negatively distort disentangling; however, this is not robust**. In particular, **axis alignment between latent and generative dimensions seems clearly impacted by increased generative space modality**.

Finally, we observe that **while our $\beta\gamma$ VAE *can* improve certain disentangling metrics and eliminate the impact of multimodal generative spaces, this is not consistent across metrics**. Our **$\beta\gamma$ VAE is thus suitable for representation learning with inconsistent evidence that it is perhaps useful in disentangling.**

These findings highlight that the **goals of disentangling and representations learning are not perfectly aligned**. Improving the global overlap between generative and latent factors boosts the informational content of the learnt representations, perhaps improving axis alignment. However, this does not necessarily improve the one-to-one relationship between generative and latent dimensions synonymous with disentangling.

Our findings are summarised as follows (the corresponding research question they address are given in brackets):

- **6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β (3.2.1)**
- **6.3 Multimodal generative factors do *not consistently* distort disentangling, but improve axis alignment (3.2.2)**
- **6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and eliminates the negative impact of multimodal generative space (3.2.3)**
- **6.5 $\beta\gamma$ VAE do not consistently disentangle better than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces (3.2.4)**

We begin with section 6.1, outlining how the empirical distributions of disentangling metrics should be interpreted.

6.1 Interpreting results

Recalling sections 2.8.5 and 2.8.6, for a *particular run*, disentangling metrics are heavily correlated, and Locatello *et al.* [2019] show that scores of the **same** disentangling model, run on the **same** dataset, are **heavily overlapping** and that the **level of correlation depends on the dataset.** Locatello *et al.* [2019] hypothesises that variation in scores *between runs* primarily driven by random seeds and hyperparameter tuning.

The disentangling metrics' empirical distributions observed by Locatello *et al.* [2019] typically have long-tailed distributions that are highly correlated between models. They note that disentangling scores typically exhibit a high degree of variation and are heavily overlapping. Furthermore, they note that: **a good run with a bad hyperparameter can beat a bad run with a good hyperparameter**.

A preferred disentangling score distribution is one that leads to both a **higher maximum score** and an empirical distribution that have **empirical distribution mass that is more positive.** With this in mind, we now discuss our findings.

6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β

In this section, we seek to answer the research question 3.2.1, repeated below:

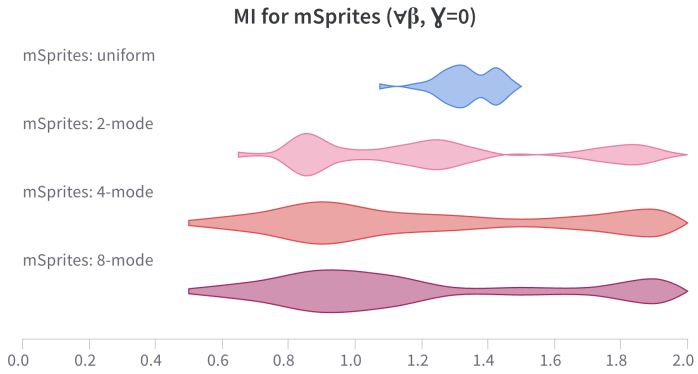
- 3.2.2 Do multimodal generative spaces distort **global latent fit** in the β VAE?

In order to assess if multimodal generative spaces distort global latent fit of β VAE, we examine the mutual information (MI) from section 2.8.1 over our *mSprites* collection of datasets, described in section 4.2. MI may be seen as a global measurement for the overall overlap between the distributions of the generative ground truth and the latent representation learnt. Our *mSprites* collection datasets has an increased number of modes in the generative space as we move from *mSprites: uniform* to *mSprites: 8-mode*. We hypothesise that the mode-seeking behaviour of the reverse KL in the β VAE biases the latents towards local correlations in the latent space at the expense of the desired global overlap with ground-truth generative factors. As such, we expect increased generative space modality to reduce the overall global overlap between the generative space and the learnt latents, resulting in reduced MI.

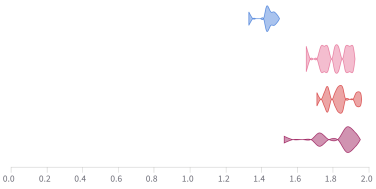
In figure 6.1, we examine the distribution of our empirical mutual information score described in section 2.8.1. The violin plots show the combined score distribution over all our β hyperparameter values and each dataset in the *mSprites* dataset group.

In , somewhat surprisingly, we see an *increase* in maximum mutual information when moving from *mSprites: uniform* to *mSprites: 2-mode*. When moving from *mSprites: 2-mode* to *mSprites: 8-mode* we see a *decrease* in this maximum mutual information. Since the relative maximum position is difficult to see graphically in this plot, we provide these maximum values. Moving from *mSprites: uniform* to *mSprites: 8-mode* are $1.4917, 1.9179, 1.9481$ and 1.9035 respectively.

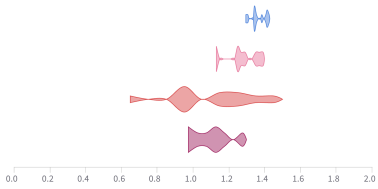
Further examining the impact for specific values of β we see that the *increase* in mutual information when moving from *mSprites: uniform* to *mSprites: 2-mode* is driven by the auto-encoder at value $\beta = 0$ in figure 6.1b. For $\beta \neq 0$ in figures 6.1c to 6.1e we see that the presence of generative space modality seemingly does decrease the mutual information, as expected.



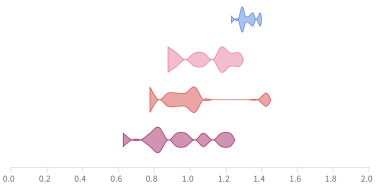
(a) All β values



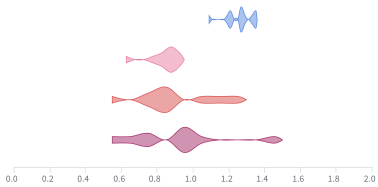
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE



(e) $\beta = 75 * 10^{-3}$
β VAE

Figure 6.1: Mutual information (MI) profile of β VAE on *mSprites* dataset: multimodal generative spaces reduce MI, indicating reduced global overlap between the latents and ground-truth generative factors. This reduction from increased modality in the generative space is diminished with increased β constraint, and we see the slope of the change reduce when moving from 6.1c, and even inverts at the highest β in 6.1d.

When examining the impact of increased generative space modality, when moving from *mSprites: 2-mode* to *mSprites: 8-mode*, we see in figures 6.1c to 6.1e that the exact value of β greatly impact this relationship. In figure 6.1c, we see a **decrease in the maximum mutual information achieved when increasing generative space modality, as expected**.

However this relationship inverts for a high β in figure 6.1e. This is somewhat surprising, and the increased constraint of $\beta = 75 * 10^{-3}$ applied to the reverse KL in the β VAE was expected to increase latent mode-seeking behaviour. This is expected to result in lower overlap between the generative and latent space as the generative space

modality increases. However, we see the opposite impact as we moving from *mSprites: 2-mode* to *mSprites: 8-mode* in figure 6.1e.

We now investigate the impact of multimodal generative spaces on disentangling in the section that follows.

6.3 Multimodal generative factors do *not consistently* distort disentangling, but improves axis alignment.

In this section, we seek to answer the research question 3.2.2, repeated below:

- 3.2.2 Do multimodal generative spaces distort **disentangling** of β VAE?

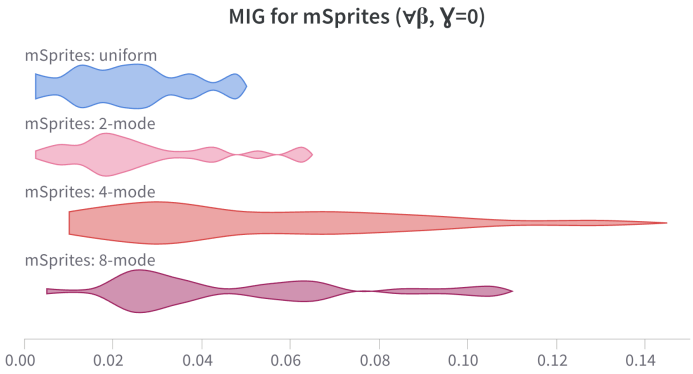
To answer research questions 3.2.2, assessing if multimodal generative spaces distort disentangling of β VAE, we examine the performance of our disentangling metrics from section 2.6 over our mSprites collection of datasets from section 4.2.

Mutual information (MI) measures the global overlap between the generative factors and the learnt latent factors. A high MI indicates that the latents overlap the generative space more and are thus reflective of this generative space. From section 6.2 we know that multimodal generative spaces distort the global latent fit, reducing MI. This distortion is heavily impacted by the β value of the β VAE. However, it is not immediately obvious how such multimodal generative spaces impact the relationships between the **individual latent factors learnt**. In particular, we are interested in how the one-to-one relationship between latents that characterises disentangling is impacted.

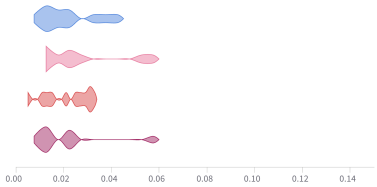
Latent traversals may provide some insight into this, however, this is a subjective assessment. A more objective assessment may be provided by the plethora of disentangling metrics described in section 2.6. Our synthetic mSprites dataset group provide the necessary ground-truth supervision needed for the disentangling metrics we use in assessment.

6.3.1 Mutual information gap (MIG)

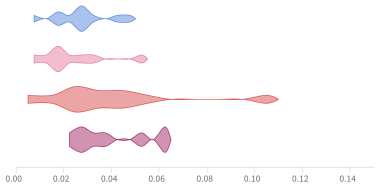
We first examine the empirical Mutual Information Gap (MIG), as described in section 2.8.2. From section 6.2, we have evidence that the mutual information score is degraded through generative space modality. Does less overall mutual information mean that the difference between the mutual information of each latent, quantified by the MIG, also decreases? Intuitively one may expect that if total mutual information in the representation reduces, the MIG may scale in line with this. The MIG represents the inter-dimension difference in mutual information and, if the MIG scales in line with the MI, this indicates that correlations between latent factors remain stable.



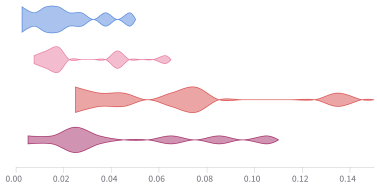
(a) All β values



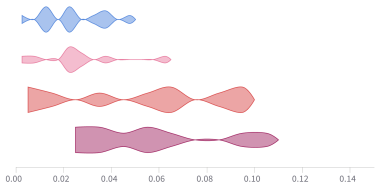
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE



(e) $\beta = 75 * 10^{-3}$
β VAE

Figure 6.2: Mutual information gap (MIG) profile of β VAE on mSprites datasets: increased modality in the generative space increases MIG, indicating increased disentangling of learnt latent space. This increase is exacerbated with increased constraint, and we see the slope of the change increase when moving from 6.2c to 6.2e.

Figure 6.2 examines how this gap changes in the presence of increased modality in the generative space. Figure 6.2a examines how increased modality in the generative space changes the MIG score over *all* β hyperparameter calibrations for the β VAE. There appears to be an improvement in the mutual information gap when moving from the *mSprites: uniform* to *mSprites: 2-mode*. This impact is preserved at all calibrations in figures 6.2b to 6.2e. This is the *opposite impact to what is observed in the mutual information* in figure 6.1a. This implies that **despite the degradation of the overall fit, disentangling improves** when moving from the *mSprites: uniform* to *mSprites: 2-mode*. Still looking at figure 6.2a, we see that increasing generative space modality when moving from the 2-mode dataset to the 8-mode dataset improves disentangling performance, increasing the MIG.

Figures 6.2b to 6.2e reveal how the different hyperparameter β tunings impact the relationship. Notably the profile of the mutual information and the mutual information gap align when moving from *mSprites: 2-mode* to *mSprites: 8-mode*, when $\beta \neq 0$. In general, datasets with multimodal generative spaces seem more reactive to hyperparameter tuning than datasets with a uniform generative space. This indicates that data with multimodal generative spaces is impacted more when varying the constraint on the β reverse KL divergence in the β VAE.

6.3.2 Factor score

Next, we examine the factor score outlined in section 2.6. Like the Higgins' metric, it is a supervised disentangling metric. Notably, it extends the Higgins' metric to include a **measure of axis alignment**. In order to assess the impact of generative space modality on disentangling and axis alignment, we examine the distribution of this factor score on the various datasets in *mSprites* in figure 6.3.

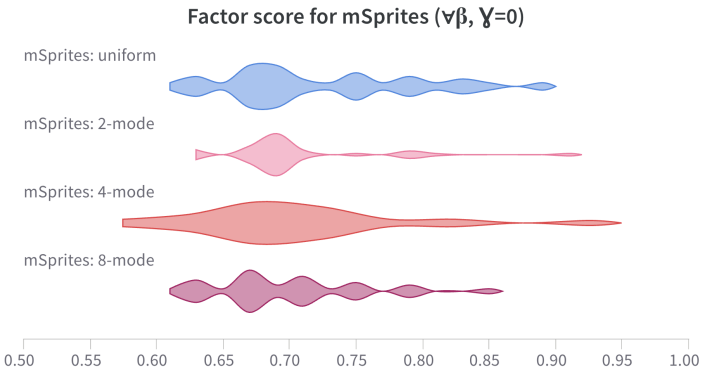
Examining 6.3, the factor score seems to show that generative space modality, seen when moving from the *mSprites: uniform* to *mSprites: 8-mode*, negatively impacts the score for non-zero βs. This is indicative of improved axis alignment, as the factor score is an error. Notably, this impact aligns with MIG in figure 6.1. It is also worth noting that this negative impact is enhanced when the hyperparameter value of β increases, and we see more pronounced distribution changes when moving from figure 6.3b to figure 6.3e. **Importantly this is indicative of increased generative space modality having a positive impact on axis alignment in the learnt generative space.**

6.3.3 SAP score

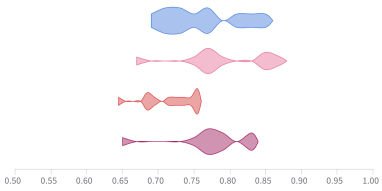
We examine the SAP score described in section 2.6. Like the factor score, the SAP score supposedly improves on the Higgins' metric. A low SAP score is indicative of poor disentangling. Importantly the SAP differs from the factor score as it is **not classifier based**. The introduction of the score thus tests disentanglement in multimodal generative spaces using a score that is not classifier based, a key criticism of many disentangling metrics.

Figure 6.4 examines the impact of multimodal generative spaces on the SAP score. In figure 6.4 we examine the distribution of the empirical SAP score over all β hyperparameters. We see that the presence of modality in the generative space negatively impacts the SAP score, seen when moving from the *mSprites: uniform* to *mSprites: 2-mode*. However, the presence of increased modality in the generative space seems to slightly increase the SAP score, seen when from the *mSprites: 2-mode* to *mSprites: 8-mode*.

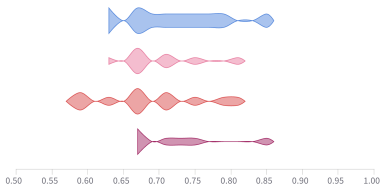
The reduction in SAP score is impacted by the β hyperparameterisation as seen in figures 6.4c to 6.4e. Higher values of β improve the performance of high generative space modality. In figure 6.4c at $\beta = 25 * 10^{-3}$ the 4-mode generative space performs



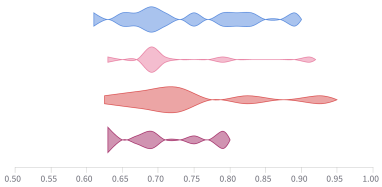
(a) All β values



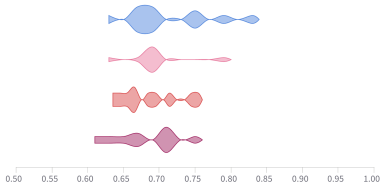
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE

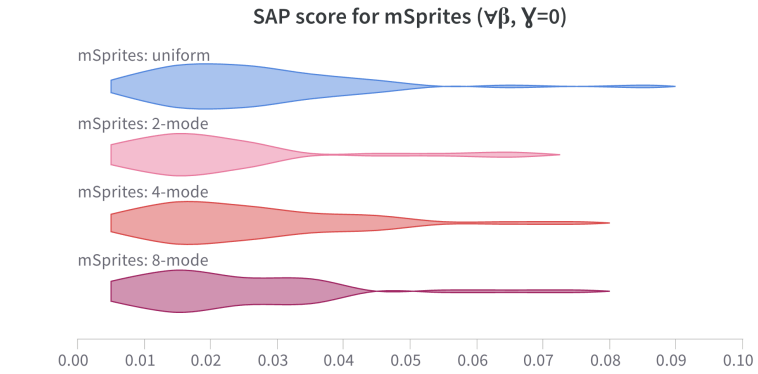


(e) $\beta = 75 * 10^{-3}$
β VAE

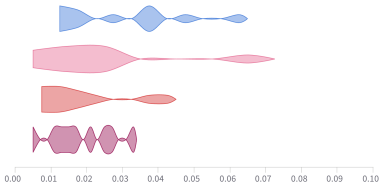
Figure 6.3: Factor score profile of β VAE on *mSprites* dataset: generative space modality reduces factor score indicating reduce disentangling of learnt latent space. Figure 6.3b to 6.3e show the factor score is less sensitive to the β parameterisation than other metrics, like the MIG in figure 6.2, but β does impact the relationship between datasets in *mSprites*.

best. In figure 6.4d at $\beta = 50 * 10^{-3}$ the 2-mode and 4-mode generative space performs best. Finally in figure 6.4e at $\beta = 75 * 10^{-3}$ the 8-mode generative space performs best.

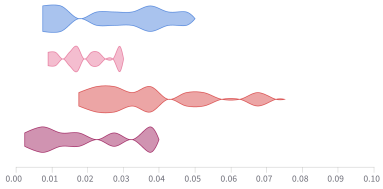
Importantly the overall profile in figure 6.4 is broadly aligned with figure 6.3, confirming the relationship seen when using the factor score, but using the unsupervised SAP score. **This indicates that when using an unsupervised metric, generative space modality harms disentangling.**



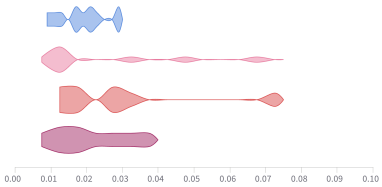
(a) All β values



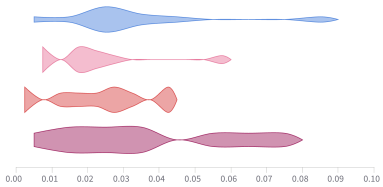
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE



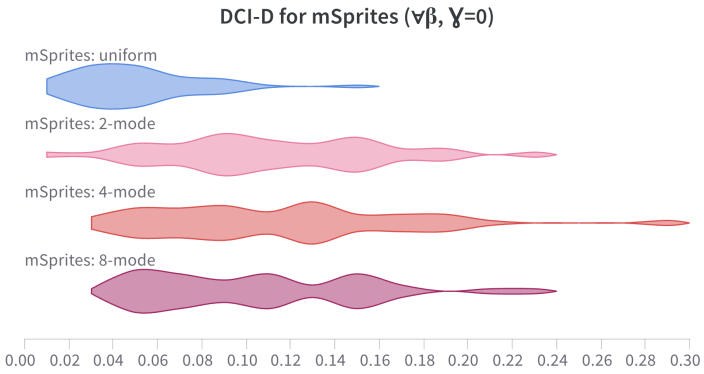
(e) $\beta = 75 * 10^{-3}$
β VAE

Figure 6.4: SAP score profile of β VAE on *mSprites* dataset: modality in the generative space reduces the SAP score. 6.4b to 6.4e show how the β value can drastically alter the impact of increased modality in the generative space.

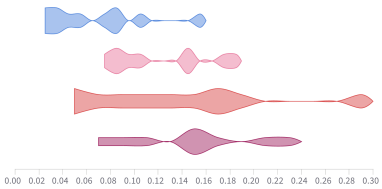
6.3.4 DCI-D

The DCI metrics were introduced to provide a framework for the qualitative evaluation of disentangled representations when the **ground-truth is available**. In addition, the three criteria were introduced to compare models on an equal basis in an attempt to remove the need for visual evaluation.

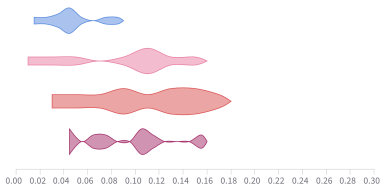
The Disentanglement score (DCI-D) captures the extent to which the learnt latent code $\mathbf{z}$ factorises (disentangles) the underlying factors of variation, with each variable capturing at most one ground-truth factor. The scores quantify deviation from ideal one-to-one mapping synonymous with operating in the direction **from one factor to many latents** as seen in figure 2.8.



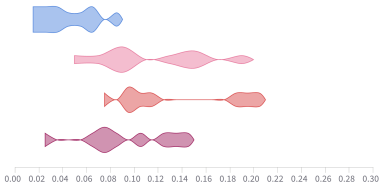
(a) All β values



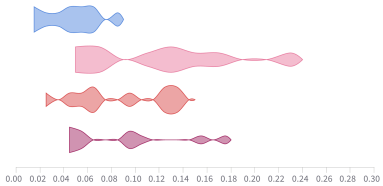
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE



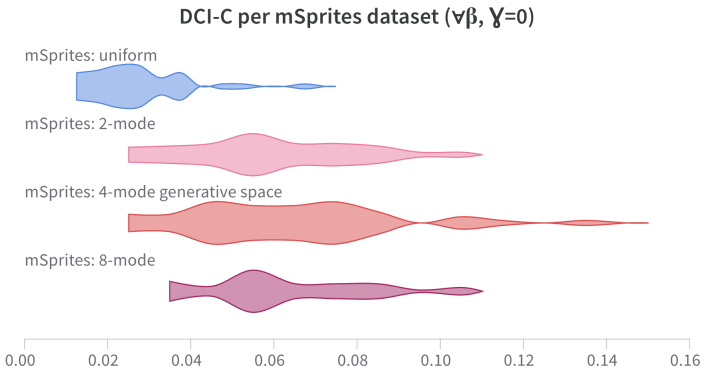
(e) $\beta = 75 * 10^{-3}$
β VAE

Figure 6.5: DCI-Disentangling (DCI-D) profile of β VAE on *mSprites* dataset: generative space modality increases DCI-D, indicating increased disentangling of learnt latent space. This relationship varies widely based on the parameterisation of β as seen in figures 6.5b to 6.5e.

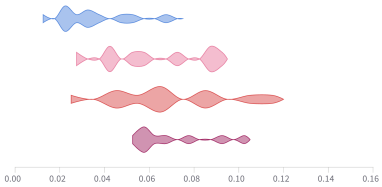
Examining figure 6.5a, we see a DCI-D profile over all hyperparameters is broadly similar to the MIG in figure 6.2a. Differences are seen at individual hyperparameter calibrations. However, they both seem to capture the broad trend that **increased generative space modality seems to improve disentangling.** The alignment between DCI-D and MIG seems appropriate as they measure similar things, and importantly, from section 2.8.5 are known to be highly correlated.

6.3.5 DCI-C

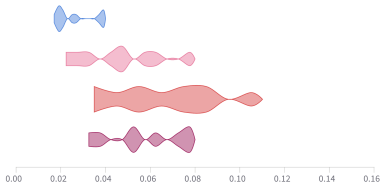
The Completeness score (DCI-C) captures the extent to which each underlying factor is captured by a single latent code. The score quantifies deviation from the ideal one-to-one mapping synonymous with disentangling. It checks if one generative factor is



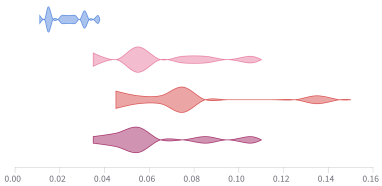
(a) All β values



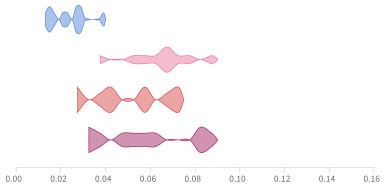
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE

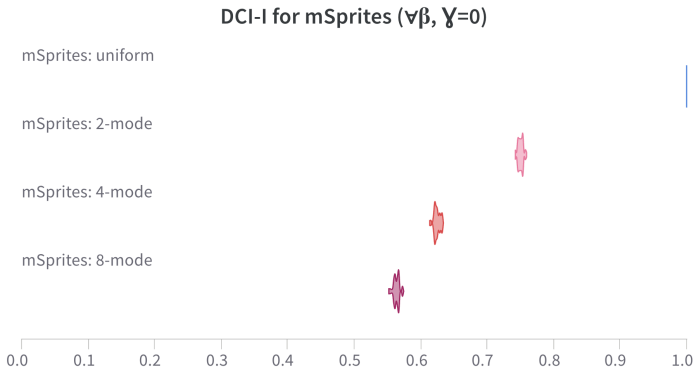


(e) $\beta = 75 * 10^{-3}$
β VAE

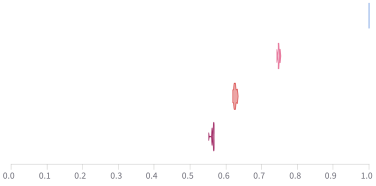
Figure 6.6: Completeness score (DCI-C) profile of β VAE on *mSprites* dataset: generative space modality increases DCI-C indicating increases disentangling of learnt latent space. The impact of increased generative space modality varies widely based on the parameterisation of β as seen in figures 6.6b to 6.6e.

captured by many latents. It operates in the direction from **one latent to many factors**. This is opposite to the direction DCI-D operates in.

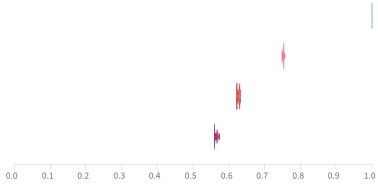
Comparing figure 6.6 and 6.5a, we see a profile similar at all hyperparameters, albeit at different scales. **This implies that generative space modality seems to *improve* disentangling similarly in either direction.** The one-to-one relationship **from latents to factors** is impacted similarly to the one-to-one relationship **from factors to latents**.



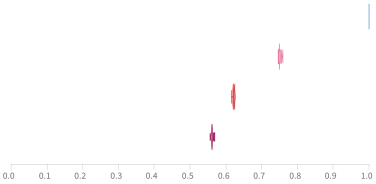
(a) All β values



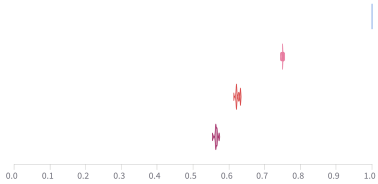
(b) $\beta = 0$
Auto-encoder



(c) $\beta = 25 * 10^{-3}$
β VAE



(d) $\beta = 50 * 10^{-3}$
β VAE



(e) $\beta = 75 * 10^{-3}$
β VAE

Figure 6.7: Informativeness score (DCI-I) profile of β VAE on *mSprites* dataset: generative space modality increases DCI-I. This relationship seems stable regardless of hyperparameterisation of β as seen in figures 6.6b to 6.6e.

6.3.6 DCI-I

The Informativeness score (DCI-I) is the amount of information about the underlying factors of variation captured by the representation. Examining figure 6.3.6 we see a similar profile over all β values and for individual β values in figures 6.7b to 6.7e. It seems that multimodal generative spaces distinctly cause less information about the underlying factors to be captured. This is confirmed by examining figure 6.7b to 6.7e. This seems similar to the MI scores from figure 6.1a, but DCI-I lacks the sensitivity to β.

6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and *consistently* eliminates negative impacts of multimodal generative spaces

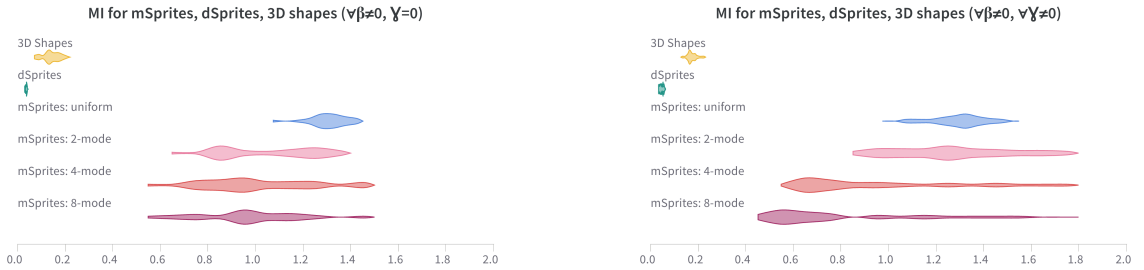
Here we attempt to answer research question 3.2.3, repeated below:

- 3.2.3 Does our proposed $\beta\gamma$ VAE provide robust **global latent fit** when compared to the β VAE? Is this impacted by multimodal generative spaces?

In order to do this and assess whether our $\beta\gamma$ VAE has better global latent alignment than the β VAE, we examine the mutual information (MI) over all datasets outlined in section 5.1. As in section 6.2, we use MI as it is an indicator of the degree of overlap between the ground-truth generative factors and the learnt latents. In representation learning, we want our learnt representations to capture as much information about ground-truth generative factors as possible, as indicated by a higher MI.

From figure 6.8 we see that the $\beta\gamma$ VAE has better capacity than the β VAE, as indicated by the higher MI. We examine what impact $\beta\gamma$ VAE has on the mutual information gap (MIG), in figure 6.9a in section 6.5 that follows.

Recall from section 2.4.1 increasing mutual information $\mathbb{I}(\mathbf{x};\mathbf{z})$ can be encouraged by increasing both bounds of $\mathbb{H} - \mathbb{D} \leq \mathbb{I}(\mathbf{x};\mathbf{z}) \leq \mathbb{R}$. Increasing the upper bound $(\mathbb{R} := \mathbb{D}_{\mathbb{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\right))$ may be seen as a means to control the maximum capacity of the encoder channel. A similar effect may be achieved by reducing the distortion $(\mathbb{D} := -\log p_\theta(\mathbf{x} \mid \mathbf{z}))$, with stronger decoder.



(a) $\beta = 25, 50, 75 * 10^{-3}, \gamma = 0$ β VAE

(b) $\beta = 25, 50, 75 * 10^{-3}, \gamma = 25, 50, 75 * 10^{-3}$ $\beta\gamma$ VAE

Figure 6.8: Empirical mutual information (MI), on average, is higher for $\beta\gamma$ VAE, in figure 6.8a, when compared to the γ VAE, indicating better capacity, in figure 6.8b.

Since our decoder remains unchanged, it seems the boost in our empirical mutual information score (MI) for the $\beta\gamma$ VAE over the β VAE stems from the additional γ constrained forward KL increasing the model's latent capacity. This is seen in figure 6.8, and the score in **every dataset** in figure 6.8a has a longer tail than that in 6.8b. Scores in figure 6.8a reach higher values, with the empirical distributions moving upward.

We continue with findings that may be derived from figure 6.8. Comparing *mSprites* datasets in figures 6.8a and 6.8b we see that the impact of moving from *mSprites: uniform* to *mSprites: 2-mode* changes. In figure 6.8a we see that MI decreases when modality is introduced in the generative space. By contrast this MI increases in figure 6.8b. In addition, in figure 6.8a the maximum MI increases when moving from *mSprites: 2-mode* to *mSprites: 8-mode*. This increase is largely eliminated in figure 6.8b.

Importantly, the $\beta\gamma$ VAE's increased MI is evidence that the model produces latents that have better global overlap with the generative ground-truth factors. Furthermore, the impact of generative space modality is diminished as the $\beta\gamma$ counteracts the mode-seeking latent behaviour of the β VAE. We now attempt to assess if this better global overlap improves disentangling in the section that follows.

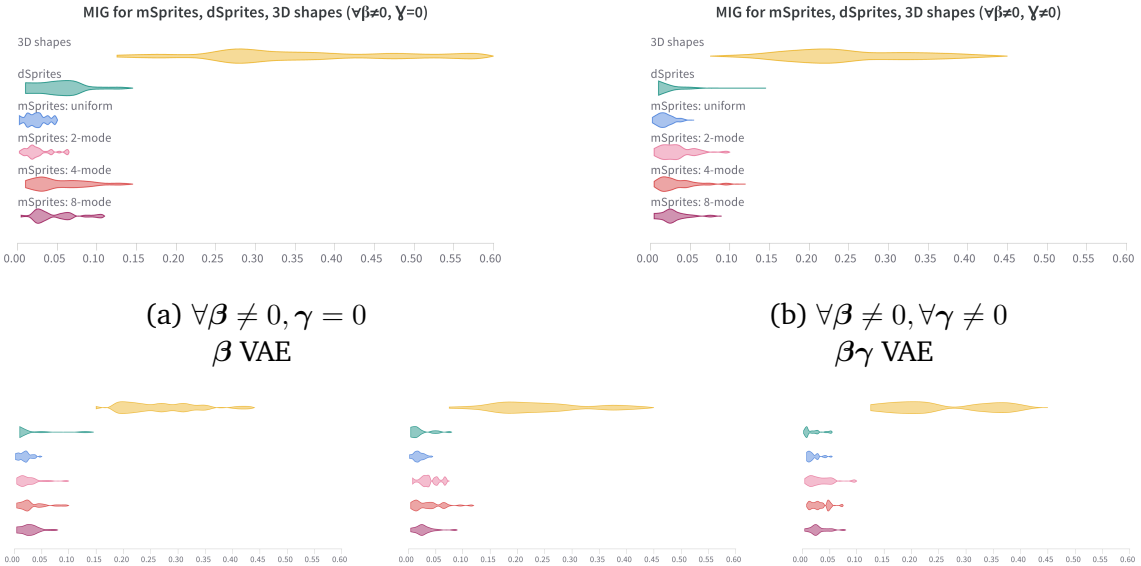
6.5 $\beta\gamma$ VAE does *not consistently* disentangle better than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces

From section 6.4 we see that the $\beta\gamma$ VAE has greater latent capacity than the β VAE. This is because both models have the same architecture, yet the $\beta\gamma$ results in greater empirical mutual information. However, it is not immediately obvious that this additional latent capacity will improve disentangling. Here we look at our range of disentangling metrics, from section 2.6, over our datasets, from section 5.1, in order to answer research question 3.2.4, repeated below:

- 3.2.4 Does our proposed $\beta\gamma$ VAE provide **robust disentangling** when compared to the β VAE? Is this impacted by multimodal generative spaces?

From figure 6.8 we see that the $\beta\gamma$ VAE has better latent capacity than the β VAE; with this in mind, we first examine what impact $\beta\gamma$ VAE has on the mutual information gap (MIG) in figure 6.9. Intuitively one may expect that if total mutual information in the representation improves, the inter-dimension difference in mutual information, as shown by the MIG, may scale in line with this if the correlations between latent factors remain stable.

6.5.1 Mutual information gap (MIG)



(a) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 0$ $\boldsymbol{\beta}$ VAE

(b) $\forall \boldsymbol{\beta} \neq 0, \forall \boldsymbol{\gamma} \neq 0$ $\boldsymbol{\beta\gamma}$ VAE

(c) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 25 * 10^{-3}$ $\boldsymbol{\beta\gamma}$ VAE

(d) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 50 * 10^{-3}$ $\boldsymbol{\beta\gamma}$ VAE

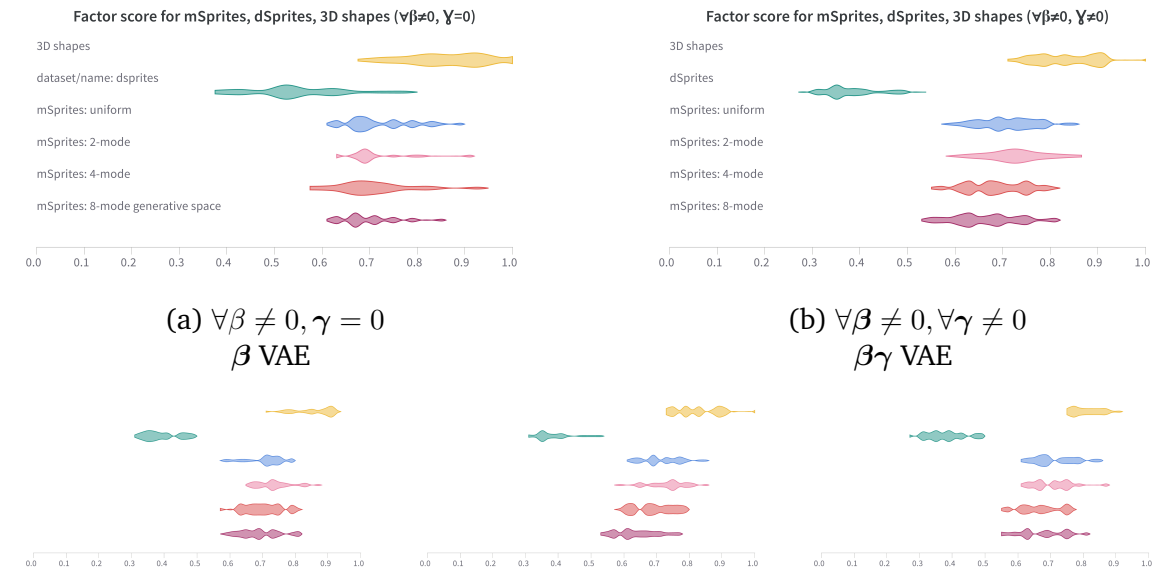
(e) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 75 * 10^{-3}$ $\boldsymbol{\beta\gamma}$ VAE

Figure 6.9: Mutual information gap (MIG) profile of $\boldsymbol{\beta}$ VAE, compared to $\boldsymbol{\beta\gamma}$ VAE for all datasets: Comparing figures 6.9a and 6.9b we see that the introduction of the $\boldsymbol{\gamma}$ term, with the exception of *mSprites: 2-mode*, diminishes MIG, **indicating worse disentangling**. Figures 6.2c to 6.2e indicate that the reduction in MIG in the $\boldsymbol{\beta\gamma}$ VAE changes with the $\boldsymbol{\gamma}$ hyperparameterisation. A higher $\boldsymbol{\gamma}$ tends to lead to a lower MIG, though this relationship is not universal across datasets.

Figure 6.9 examines the MIG over all datasets explored in the investigation. Comparing figures 6.9a and 6.9b, we see that the $\boldsymbol{\beta\gamma}$ **VAE, when aggregated across all $\boldsymbol{\gamma}$ values, reduces the maximum empirical MIG, leading to lower scores for every dataset.**

Figures 6.9c to 6.9e explore the impact of the $\boldsymbol{\gamma}$ parameter. In general, $\boldsymbol{\gamma} \neq 0$ leads to a lower maximum MIG, with higher $\boldsymbol{\gamma}$ tending to lead to a lower MIG, though this relationship is not universal across datasets. Of the datasets in *mSprites*, the presence of generative space modality in *mSprites: 2-mode* seems to **improve the MIG** over to *mSprites: 2-mode*, across both models and all datasets. The relationship seen when moving from *mSprites: 2-mode* to *mSprites: 8-mode* is heavily dependant on the $\boldsymbol{\gamma}$ hyperparameterisation but typically **increased modality in the generative space reduces the MIG, indicative of worse disentangling in the $\boldsymbol{\beta\gamma}$ VAE**. Importantly, this relationship is an inverse of that seen in the $\boldsymbol{\beta}$ VAE.

6.5.2 Factor score



(a) $\forall\beta \neq 0, \boldsymbol{\gamma = 0}$
$\boldsymbol{\beta}$ VAE

(b) $\forall\boldsymbol{\beta} \neq 0, \forall\boldsymbol{\gamma} \neq 0$
$\boldsymbol{\beta\gamma}$ VAE

(c) $\forall\beta \neq 0, \gamma = 25 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(d) $\forall\beta \neq 0, \boldsymbol{\gamma = 50 * 10^{-3}}$
$\boldsymbol{\beta\gamma}$ VAE

(e) $\forall\beta \neq 0, \boldsymbol{\gamma = 75 * 10^{-3}}$
$\boldsymbol{\beta\gamma}$ VAE

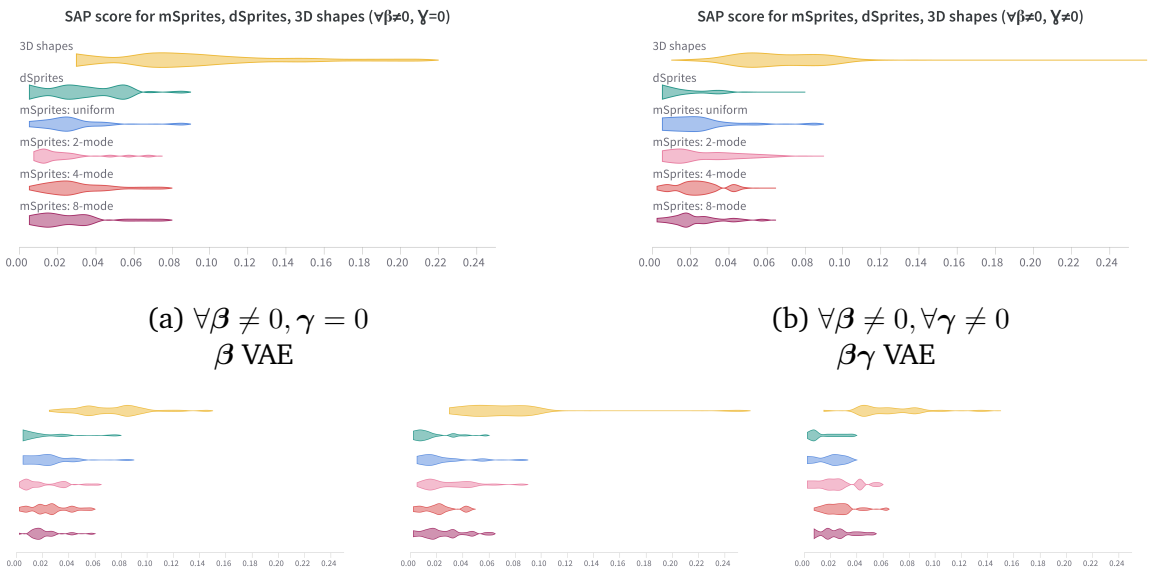
Figure 6.10: Factor score profile: comparing figures 6.10a and 6.10b, the factor scores leads to lower factor scores on $\boldsymbol{\beta\gamma}$ VAE, when compared to the $\boldsymbol{\beta}$ VAE.

Recall that the factor score extends on the Higgins' metric to include a **measure of axis alignment**. The **maximum factor score typically reduces for the $\boldsymbol{\beta\gamma}$ VAE** in figure 6.10b, when compared to the $\boldsymbol{\beta\gamma}$ VAE in figure 6.10a. Since the factor score is an error metric, this is **indicative of improved disentangling and axis alignment**. Despite this the mass of the empirical scores move upward, resulting in higher mean scores for the *cSprites* datasets.

Examining the impact of the $\boldsymbol{\beta\gamma}$ VAE across the *cSprites* datasets, we see there seems to be a downward trend in maximum factor score when moving from *mSprites: uniform* to *mSprites: 2-mode*, indicating **improved disentangling and axis alignment when there is modality in the generative space**. This **relationship in *cSprites* is largely not impacted by the change in model**, nor change in $\boldsymbol{\gamma}$ value.

6.5.3 SAP score

Recall that, like the factor score, the SAP score improves on the Higgins' metric. A low SAP score is indicative of poor disentangling. Importantly the SAP score differs from the factors score as it **not classifier based**. Figure 6.11 shows that when comparing figures 6.11a and 6.11b, we see that the $\boldsymbol{\beta\gamma}$ VAE **elongates the empirical distribution of the SAP scores only for 3D shapes**. The profile seems **largely similar to that of the factor scores** in section 6.5.2.



(a) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 0$
$\boldsymbol{\beta}$ VAE

(b) $\forall \boldsymbol{\beta} \neq 0, \forall \boldsymbol{\gamma} \neq 0$
$\boldsymbol{\beta\gamma}$ VAE

(c) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 25 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(d) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 50 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

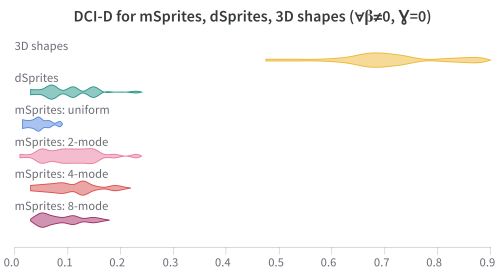
(e) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 75 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

Figure 6.11: SAP score profile shows, on average, the $\boldsymbol{\beta\gamma}$ VAE in figure 6.11b performs better than $\boldsymbol{\beta}$ VAE in figure 6.11b. Figures 6.11c to 6.11c show the best performing $\boldsymbol{\gamma}$ value varies according to dataset.

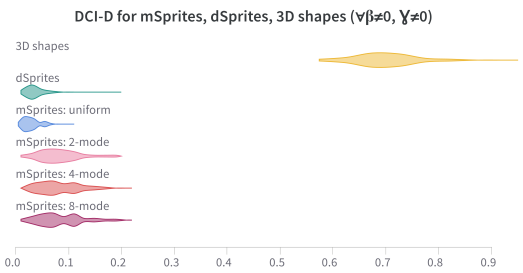
Figure 6.11c to 6.11e explore the impact of the $\boldsymbol{\gamma}$ parameter. In general $\boldsymbol{\gamma} \neq 0$ leads to better maximum empirical SAP scores. It is worth noting that for higher value of $\boldsymbol{\gamma} = 75 * 20^{-3}$, in figure 6.11e, the negative impact of moving from *mSprites: uniform* to *mSprites: 4-mode* is inverted, and the presence of multimodal generative space increases the SAP score. **Moving to *mSprites: 8-mode* still decreases the SAP score**, as is the typically seen across all hyperparameter values.

6.5.4 DCI-D

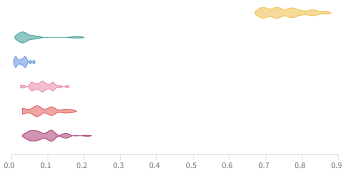
Recall, the Disentanglement score (DCI-D) quantifies the deviation from the ideal one-to-one mapping, operating in the direction **from one factor to many latents** as seen in figure 2.8.



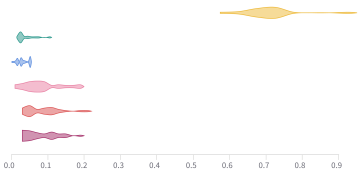
(a) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 0$
$\boldsymbol{\beta}$ VAE



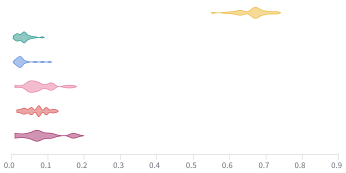
(b) $\forall \boldsymbol{\beta} \neq 0, \forall \boldsymbol{\gamma} \neq 0$
$\boldsymbol{\beta\gamma}$ VAE



(c) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 25 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE



(d) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 50 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE



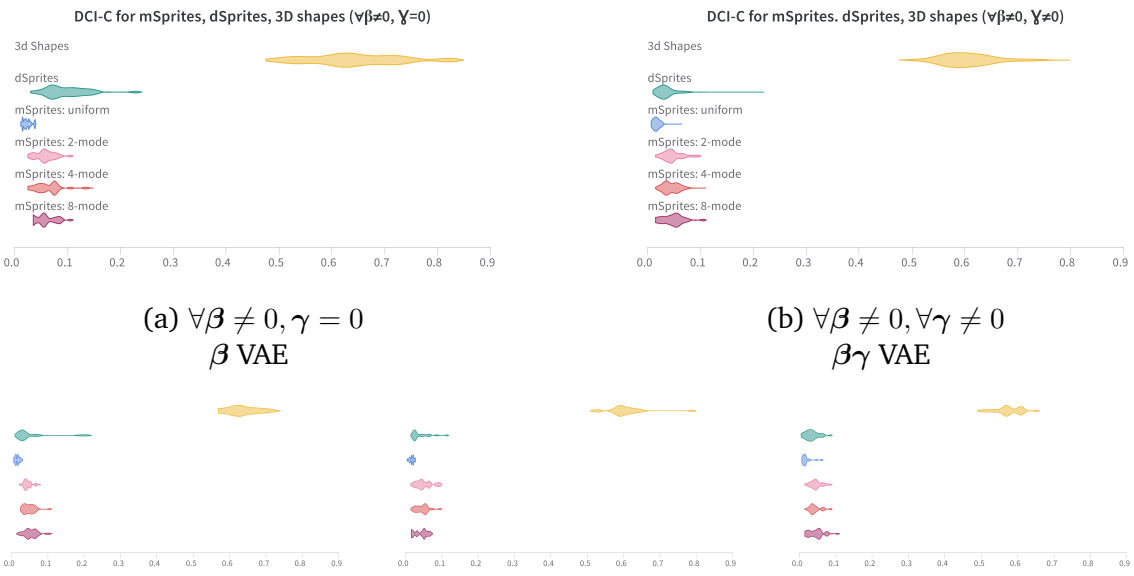
(e) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 75 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

Figure 6.12: Comparing 6.12a and 6.12b, in general $\boldsymbol{\beta\gamma}$ VAE improves the empirical distribution of the DCI-D score for most datasets. Notable exceptions are *mSprites: 2-mode* and *mSprites: uniform*. Figures 6.12c to 6.12e show that the impact of multimodal generative spaces is impacted by the $\boldsymbol{\gamma}$ value.

Examining figure 6.12, we compare figures 6.12a and 6.12b to see the impact of introducing the $\boldsymbol{\gamma}$ term. Notably, on average, the $\boldsymbol{\gamma}$ term does seem to improve the DCI-D score. This is most prominent in *3D shapes*, *mSprites: uniform* and *mSprites: 8-mode*. Importantly, in figure 6.12b the $\boldsymbol{\beta\gamma}$ **seems to eliminate the drop in the score when moving from the *mSprites: 2-mode* to *mSprites: 8-mode*** seen in figure 6.12a. Figures 6.12c to 6.12e indicate that the score is sensitive to the $\boldsymbol{\gamma}$ value.

6.5.5 DCI-C

Recall from section 2.8.4 that the Completeness score (DCI-C) quantifies deviation from the ideal one-to-one mapping synonymous with disentangling. It operates in the direction from **one latent to many factors**, opposite to the direction DCI-D operates in. By assessing both the DCI-D and DCI-C, we get an indication of the "symmetry" of the disentangling.



(a) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 0$
$\boldsymbol{\beta}$ VAE

(b) $\forall \boldsymbol{\beta} \neq 0, \forall \boldsymbol{\gamma} \neq 0$
$\boldsymbol{\beta\gamma}$ VAE

(c) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 25 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(d) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 50 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(e) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 75 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

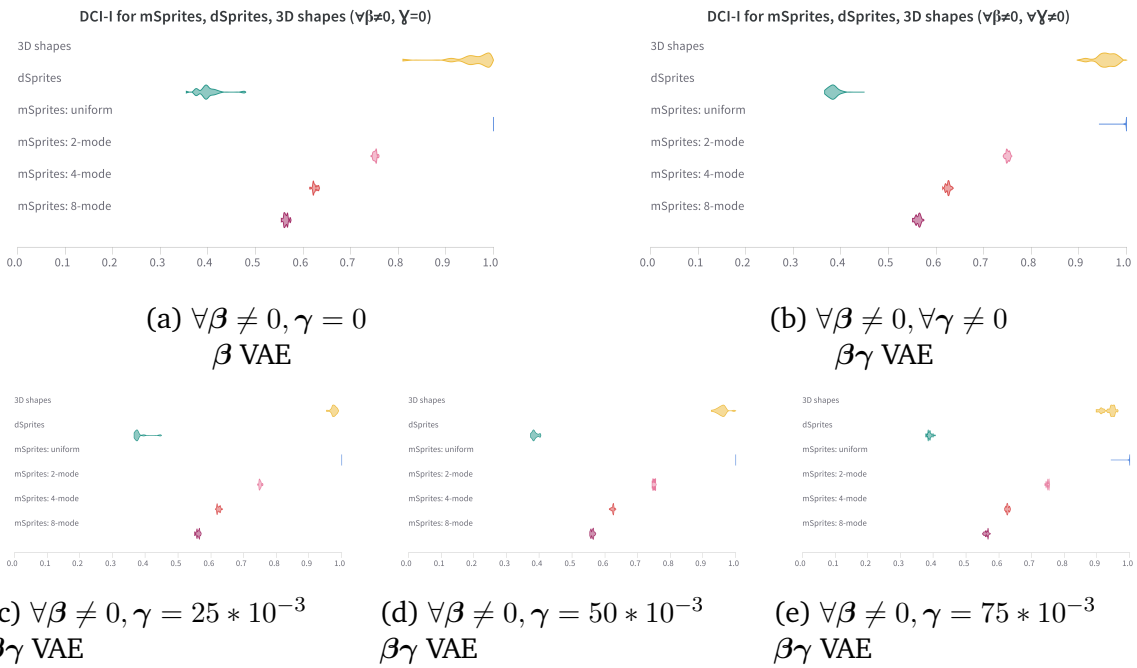
Figure 6.13: Comparing 6.13a and 6.13b, in general $\boldsymbol{\beta\gamma}$ VAE do not typically improve DCI-C score. Figures 6.13c to 6.13e show that the value of $\boldsymbol{\gamma}$, where this lack of improvement is largely unaffected by $\boldsymbol{\gamma}$ value.

Comparing figure 6.13a and 6.13b we see a profile similar at all hyperparameters. Unlike 6.12, **DCI-C do not typically improve with the introduction of the $\boldsymbol{\gamma}$ term**. However, like DCI-D, in figure 6.13b, the $\boldsymbol{\beta\gamma}$ **seems to eliminate the drop in the score when moving from the *mSprites: 2-mode* to *mSprites: 8-mode***, as seen in figure 6.13a. Figure 6.12e indicates that higher values in $\boldsymbol{\gamma}$ eliminate the impact of generative factor modality, and maximum DCI-C is largely the same across the *cSprites* dataset.

This implies that multimodal generative spaces impact disentangling similarly in **either direction**. The one-to-one relationship **from latents to factors** is impacted similarly to the one-to-one relationship **from factors to latents**, and the impact of introducing the $\boldsymbol{\gamma}$ term may be seen as "symmetrical".

6.5.6 DCI-I

The informativeness score (DCI-I) quantifies the amount of information about the underlying factors of variation captured by the representation.



(a) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 0$
$\boldsymbol{\beta}$ VAE

(b) $\forall \boldsymbol{\beta} \neq 0, \forall \boldsymbol{\gamma} \neq 0$
$\boldsymbol{\beta\gamma}$ VAE

(c) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 25 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(d) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 50 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

(e) $\forall \boldsymbol{\beta} \neq 0, \boldsymbol{\gamma} = 75 * 10^{-3}$
$\boldsymbol{\beta\gamma}$ VAE

Figure 6.14: Similar profile over all hyperparameters in figures 6.14a and 6.14b. Score variation is increased when γ term is introduced.

In 6.14 we see a similar profile over all hyperparameters in figures 6.14a and 6.14b. Importantly $\gamma \neq 0$ increases the variation of the DCI-C score. Individual γ values in figures 6.14c to 6.14e indicate the γ parameterisation do not impact the relative position of the datasets' DCI-C but may have an impact of the score's empirical distribution range. It seems that **regardless of model, increased generative factor modality distinctly causes less information about the underlying factors to be captured.**

6.6 Summary of findings

In this chapter, we have provided empirical evidence that data sampled from **multimodal generative spaces distort global overlap between generative and learnt latents spaces when the reverse KL is used for latent regularisation**. In our experiments, we observe a drop in overall global overlap, as indicated by a drop in mutual information (MI) in section 6.2 and DCI-I in section 6.3.6. The global information content captured by the latents is thus reduced by the presence of generative space modality. This indicates that **for representation learning, latent regularisation using the mode-seeking reverse KL may not be optimal for data sampled from multimodal generative spaces.**

Despite this drop in the overall global overlap between the generative and latent spaces, the evidence indicates that for **latent regularisation using the mode-seeking reverse KL, multimodal generative spaces do not *consistently* negatively impact**

disentangling. In fact, some disentangling metrics indicate that multimodal generative spaces may **actually improve disentangling** in this context. Evidence for this is seen in the improvements of the MIG in section 6.3.1, DCI-D in 6.3.4 and DCI-C in section 6.3.5. However, for latent regularisation using the mode-seeking reverse KL, multimodal generative spaces do seem to **negatively impact *alignment* between the generative space and learnt latents**. This is indicated by negative impacts on the factor score in section 6.3.2 and the SAP score in section 6.3.3.

Investigating the potential merit of an additional constrained forward KL for latent regularisation, we find that our $\beta\gamma$ **VAE remedies the distortion in the global overlap between multimodal generative spaces and latents learnt**. This is evidenced by the elimination of the reduction in MI observed with the β VAE. This improved MI is also observed more broadly on other disentangling datasets that do not explicitly have multimodal generative spaces. This indicates that our $\beta\gamma$ **VAE generally has better global overlap between the generative space and the learnt latent space when compared to the β VAE**. This suggests that there is **merit for latent regulation using an additional constrained forward KL in representation learning.**

However, the impact that the $\beta\gamma$ VAE has on disentangling is less clear. We find that the $\beta\gamma$ **VAE do not consistently disentangle better than the β VAE, nor does it consistently eliminate the impact of multimodal generative spaces. Negative impacts on alignment from multimodal generative spaces are still largely observed**, as indicated by reduced factor and SAP scores in sections 6.5.2 and 6.5.3 respectively. There is some evidence that **disentangling *can* improve in a "symmetrical" manner**, indicated by improved DCI-D in section 6.5.4 and DCI-C in section 6.5.5 across datasets. Furthermore, these scores indicate that the $\beta\gamma$ **VAE *can* eliminate impact of multimodal generative spaces have on disentangling.** Importantly, though, the findings on disentangling are less robust, varying across scores and datasets.

Our study thus observes that when using the reverse KL for latent regularisation, multimodal generative spaces largely distort the overall information content captured by the representation. This is related to the mode-seeking behaviour of the reverse KL as evidenced by reduced fit from increased generative space modality. This distortion may be remedied by introducing an additional constrained forward KL for latent regularisation, as done in our $\beta\gamma$ VAE. Despite this, the impact of disentangling is less clear. While our study shows that multimodal generative spaces negatively distort alignment, it is not clear that this robustly hinders disentangling. Furthermore, we observe that while our $\beta\gamma$ VAE can improve specific disentangling metrics and eliminate the impact of multimodal generative spaces, this is not consistent across metrics, and results are less robust and transparent.

These findings emphasise that improving the global overlap between generative and latent factors may boost the informational content of the learnt representations but does not necessarily improve the individual inter-dimensional relationship that is the concern of disentangling.

Our findings are summarised as follows (the corresponding research question they address are given in brackets):

- **6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β (3.2.1)**
- **6.3 Multimodal generative factors do *not consistently* distort disentangling but improves axis alignment. (3.2.2)**
- **6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and *consistently* eliminates negative impacts of multimodal generative spaces (3.2.3)**
- **6.5 $\beta\gamma$ VAE do not *consistently* disentangle better than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces (3.2.4)**

Having discussed our key findings, we explore future research that may be conducted in this space before concluding in the chapter that follows.

Chapter 7

Future research and conclusion

7.1 Future research

Section 2.8.6 stresses the importance of random seeds in disentangling. With this in mind, our study is admittedly on a much smaller scale than what is needed to truly account for random variation. Our experiments have only 10 runs per β-γ hyperparameter combination described in section 5.5. The study in Locatello *et al.* [2019], by contrast, have 50 runs per hyperparameter combination. Expanding the scale of our study would provide for more credible results.

A deeper study into the inconsistent impacts shown by different disentangling metrics in our study would also be beneficial. This would involve an examination into exactly what detection biases are present in each disentangling metric and how this impacts results. Furthermore, we could perform time-consuming latent traversals to see if the metrics align with what is observed.

We could expand our model to be a mixture of Gaussians in the latent space. It is likely that this model would handle multimodal generative factors well. However, this model would not be challenged by multimodal generative spaces, and so the study of the failures would not be as rich.

Our study also focuses on the multimodal generative spaces, inferring that this impacts the modality of the latent space via informational metrics. These datasets could be improved, and the location and scale of each generative space mode could be random and the impact examined. A more thorough investigation into how more general generative space structures impacts the latent space would be fruitful. The multimodal structure of our generative space is admittedly narrow, and generative spaces that vary over a wider variety of generative space structures should be explored.

Studies into what drives generative space shape would also be beneficial. Our current hypothesis is that a driver of structure in the generative space is when objects in the

dataset do not fully examine all potential options that could theoretically be available. For example in *mSprites* we limit the sprite locations in order to produce generative space structure in *cSprites: 2-mode* to *cSprites: 8-mode*. However, the sprite may have broader locations as seen in *cSprites: uniform*. It is by limiting the full variability of the object that we get structure in the generative space.

Exploring the impact of non-Gaussian variational and prior distributions would be insightful. The unimodal nature of our variational latent posterior and latent prior played a key role in our study. It would be interesting to see the posterior bias induced by alternative, multimodal latent prior and posterior distributions.

Lastly, exploring latent regularisation using alternative divergences other than the (forward and reverse) KL may prove beneficial when investigating the implications of generative space structure. Other divergences will certainly induce different latent solution biases, which may be beneficial for certain generative space structures.

7.2 Conclusion

Our work focuses on examining the bias from latent regularisation. In particular, we examine the mode-seeking behaviour of the reverse KL, the typical divergence used in variational inference. Since the goal of representation learning is to capture the generative space well in the learnt latent space, we investigate this bias with a made-for-purpose collection of datasets *mSprites* with multimodal generative spaces. Furthermore, we introduce an additional constrained forward KL latent regulariser in our proposed model, the $\beta\gamma$ VAE.

When using the reverse KL for latent regularisation, we find multimodal generative spaces distort the overall information content captured by the learnt representation. We provide evidence that this is related to the mode-seeking behaviour of the reverse KL, and the reduced global fit is seen when increasing generative space modality in *mSprites*. We remedy this impact by introducing an additional constrained forward KL for latent regularisation in our $\beta\gamma$ VAE.

We find that the impact of multimodal generative spaces on disentangling is less clear. Our work provides evidence that while multimodal generative spaces can negatively impact disentangling, this impact is not robust. Multimodal generative spaces do robustly improve the axis alignment of the representation learnt relative to the generative space. However, we observe inconsistent results, with some disentangling metrics negatively impacted and others positively impacted.

Our $\beta\gamma$ VAE can improve certain disentangling metrics and eliminate the impact of multimodal generative spaces. However, once again, this is not consistent across

metrics. While our $\beta\gamma$ VAE seems suitable for representation learning, results are less clear that the model is beneficial for disentangling benefits when compared to the β VAE.

Our findings highlight that the goals of disentangling and representation learning are not perfectly aligned. Improving the global overlap between generative and latent factors boosts the informational content of the learnt representations, improving axis alignment globally. However, this does not necessarily improve the one-to-one relationship between generative and latent dimensions synonymous with disentangling. As such improved global correlation in the latent space, when compared to the generative space, does not necessarily improve disentangling.

Our findings are summarised as follows (the corresponding research question they address are given in brackets):

- **6.2 Multimodal generative spaces distort global latent fit of β VAE, particularly at lower values of β (3.2.1)**
- **6.3 Multimodal generative factors do not *not consistently* distort disentangling but improves axis alignment. (3.2.2)**
- **6.4 $\beta\gamma$ VAE has *consistently* better global latent fit than β VAE and eliminates negative impacts of multi- modal generative space (3.2.3)**
- **6.5 $\beta\gamma$ VAE does *not consistently* disentangle better than β VAE, *nor* does it *consistently* eliminate impact of multimodal generative spaces (3.2.4)**

References

[Abdi *et al.* 2019] Amir H. Abdi, Purang Abolmaesumi, and Sidney Fels. Variational learning with disentanglement-pytorch. *arXiv preprint arXiv:1912.05184*, 2019.

[Achille and Soatto] A Achille and S Soatto. On the emergence of invariance and disentangling in deep representations. arxiv 2017. *arXiv preprint arXiv:1706.01350*.

[Akinola and Fobi 2019] Iretiayo Akinola and Simone Fobi. Progressive disentanglement using relevant factor vae. *Proceedings of Machine Learning Research*, 1:1–4, 2019.

[Alanov *et al.* 2019] Aibek Alanov, Max Kochurov, Artem Sobolev, Daniil Yashkov, and D. Vetrov. Implicit λ-jeffreys autoencoders: Taking the best of both worlds. 2019.

[Alemi *et al.* 2018] Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International Conference on Machine Learning*, pages 159–168. PMLR, 2018.

[Arora *et al.* 2017] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 224–232. JMLR. org, 2017.

[Asperti and Trentin 2020] Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders. *IEEE Access*, 8:199440–199448, 2020.

[Barber *et al.* 2018] David Barber, Mingtian Zhang, Raza Habib, and Thomas Bird. Spread divergences. *arXiv preprint arXiv:1811.08968*, 2018.

[Bengio *et al.* 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[Bengio *et al.* 2013] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[Berthelot *et al.* 2017] David Berthelot, Thomas Schumm, and Luke Metz. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[Blei *et al.* 2017] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

[Burgess and Kim 2018] Chris Burgess and Hyunjik Kim. *3D Shapes Dataset*. https://github.com/deepmind/3dshapes-dataset/, 2018.

[Burgess *et al.* 2018] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in $beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

[Burgess *et al.* 2019] Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

[Chen *et al.* 2016] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[Chen *et al.* 2018] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[Chen *et al.* 2020] Nutan Chen, Alexej Klushyn, Francesco Ferroni, Justin Bayer, and Patrick Van Der Smagt. Learning flat latent manifolds with vaes. *arXiv preprint arXiv:2002.04881*, 2020.

[Choi and Min 2022] Dooseop Choi and KyoungWook Min. Hierarchical latent structure for multi-modal vehicle trajectory forecasting. *arXiv preprint arXiv:2207.04624*, 2022.

[Connor *et al.* 2021] Marissa Connor, Gregory Canal, and Christopher Rozell. Variational autoencoder with learned latent structure. In *International Conference on Artificial Intelligence and Statistics*, pages 2359–2367. PMLR, 2021.

[Cover and Thomas 1991] Thomas M Cover and Joy A Thomas. Elements of information theory. 1991.

[Creswell *et al.* 2017] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. Generative adversarial networks: An overview. *CoRR*, abs/1710.07035, 2017.

[Dai and Wipf 2019] Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

[Dai *et al.* 2020] Bin Dai, Ziyu Wang, and David Wipf. The usual suspects? reassessing blame for vae posterior collapse. In *International Conference on Machine Learning*, pages 2313–2322. PMLR, 2020.

[Davidson *et al.* 2018] Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

[Dembo *et al.* 1991] Amir Dembo, Thomas M Cover, Joy A Thomas, et al. Information theoretic inequalities. *IEEE Transactions on Information theory*, 37(6):1501–1518, 1991.

[Dieng *et al.* 2016] Adji B Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David M Blei. Variational inference via χ-upper bound minimization. *arXiv preprint arXiv:1611.00328*, 2016.

[Doan *et al.* 2019] Thang Doan, Joao Monteiro, Isabela Albuquerque, Bogdan Mazoure, Audrey Durand, Joelle Pineau, and R Devon Hjelm. On-line adaptative curriculum learning for gans. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3470–3477, 2019.

[Durugkar *et al.* 2016] Ishan Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. *arXiv preprint arXiv:1611.01673*, 2016.

[Eastwood and Williams 2018] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

[Fidler *et al.* 2012] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. *Advances in Neural Information Processing Systems*, 25:611–619, 2012.

[Fu *et al.* 2019] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.

[Ganguly and Earp 2021] Ankush Ganguly and Samuel WF Earp. An introduction to variational inference. *arXiv preprint arXiv:2108.13083*, 2021.

[Geman and Geman 1984] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984.

[Ghosh *et al.* 2018] Arnab Ghosh, Viveka Kulharia, Vinay P Namboodiri, Philip HS Torr, and Puneet K Dokania. Multi-agent diverse generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8513–8521, 2018.

[Glorot and Bengio 2010] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth International Conference on Artificial Intelligence and Statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[Gondal *et al.* 2019] Muhammad Waleed Gondal, Manuel Wüthrich, DJor dje Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

[Goodfellow *et al.* 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Goodfellow 2016] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[Goodfellow 2017] Ian J. Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. *CoRR*, abs/1701.00160, 2017.

[Graves *et al.* 2017] Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1311–1320. JMLR. org, 2017.

[Higgins *et al.* 2017] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework. *International Conference on Learning Representations*, 2(5):6, 2017.

[Hinton and Salakhutdinov 2006] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[Hinton and Van Camp 1993] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

[Hinton and Zemel 1994] Geoffrey E Hinton and Richard S Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in Neural Information Processing Systems*, pages 3–10, 1994.

[Hoang *et al.* 2017] Quan Hoang, Tu Dinh Nguyen, Trung Le, and Dinh Phung. Multi-generator generative adversarial nets. *arXiv preprint arXiv:1708.02556*, 2017.

[Hoffman and Johnson 2016] Matthew D Hoffman and Matthew J Johnson. Elbo surgery: yet another way to carve up the variational evidence lower bound. In *Workshop in Advances in Approximate Bayesian Inference, NIPS*, volume 1, page 2, 2016.

[Jordan *et al.* 1998] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer, 1998.

[Juefei-Xu *et al.* 2017] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Gang of gans: Generative adversarial networks with maximum margin ranking. *arXiv preprint arXiv:1704.04865*, 2017.

[Karras *et al.* 2017] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *CoRR*, abs/1710.10196, 2017.

[Kenfack *et al.* 2021] Patrik Joslin Kenfack, Daniil Dmitrievich Arapovy, Rasheed Hussain, SM Kazmi, and Adil Mehmood Khan. On the fairness of generative adversarial networks (gans). *arXiv preprint arXiv:2103.00950*, 2021.

[Kim and Mnih 2018] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

[Kim *et al.* 2019] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance factor vae: Learning and identifying disentangled factors. *arXiv preprint arXiv:1902.01568*, 2019.

[Kingma and Welling 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kingma and Welling 2019] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[Kumar *et al.* 2017] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

[Kynkäänniemi *et al.* 2019] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, pages 3929–3938, 2019.

[Lake *et al.* 2017] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.

[Lazarsfeld 1950] Paul F Lazarsfeld. The logical and mathematical foundation of latent structure analysis. *Studies in Social Psychology in World War II Vol. IV: Measurement and Prediction*, pages 362–412, 1950.

[Li and Turner 2016] Yingzhen Li and Richard E Turner. R'enyi divergence variational inference. *arXiv preprint arXiv:1602.02311*, 2016.

[Liu *et al.* 2018] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018):11, 2018.

[Liu *et al.* 2019] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[Locatello *et al.* 2019] 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 *International Conference on Machine Learning*, pages 4114–4124, 2019.

[Makhzani *et al.* 2015] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[Matiisen *et al.* 2019] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 2019.

[Matthey *et al.* 2017] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. *dSprites: Disentanglement testing Sprites dataset*. https://github.com/deepmind/dsprites-dataset/, 2017.

[Metropolis *et al.* 1953] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[Michlo 2021] Nathan Juraj Michlo. *Disent - A modular disentangled representation learning framework for pytorch*. Github, 2021.

[Minka and others 2005] Tom Minka et al. *Divergence measures and message passing*. Technical report, Citeseer, 2005.

[Munch 1893] Edvard Munch. *The Scream*, 1893.

[Murphy 2012] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[Murtagh and Farid 2001] F Murtagh and MM Farid. Pattern classification, by ro duda, pe hart, and dg stork. *JOURNAL OF CLASSIFICATION*, 18(2):273–275, 2001.

[Neal and Hinton 1998] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[Neyshabur *et al.* 2017] Behnam Neyshabur, Srinadh Bhojanapalli, and Ayan Chakrabarti. Stabilizing gan training with multiple random projections. *arXiv preprint arXiv:1705.07831*, 2017.

[Nguyen *et al.* 2010] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

[Nowozin *et al.* 2016] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in Neural Information Processing Systems*, 29, 2016.

[Pan *et al.* 2019] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE Access*, 7:36322–36333, 2019.

[Parisi 1988] G Parisi. *Statistical Field Theory,(Addison*, 1988.

[Peterson 1987] Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1:995–1019, 1987.

[Prokhorov *et al.* 2019] Victor Prokhorov, Ehsan Shareghi, Yingzhen Li, Mohammad Taher Pilehvar, and Nigel Collier. On the importance of the kullback-leibler divergence term in variational autoencoders for text generation. *arXiv preprint arXiv:1909.13668*, 2019.

[Pu *et al.* 2017] Yunchen Pu, Weiyao Wang, Ricardo Henao, Liqun Chen, Zhe Gan, Chunyuan Li, and Lawrence Carin. Adversarial symmetric variational autoencoder. *arXiv preprint arXiv:1711.04915*, 2017.

[Reed *et al.* 2015] Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems*, 28:1252–1260, 2015.

[Rezende and Viola 2018] Danilo Jimenez Rezende and Fabio Viola. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

[Rezende *et al.* 2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014.

[Rolinek *et al.* 2019] 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*, pages 12406–12415, 2019.

[Rubenstein *et al.* 2018] Paul K Rubenstein, Bernhard Schölkopf, and Ilya Tolstikhin. Learning disentangled representations with wasserstein auto-encoders. 2018.

[Salimans *et al.* 2015] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *International conference on machine learning*, pages 1218–1226. PMLR, 2015.

[Sason and Verdu 2016] Igal Sason and Sergio Verdu. F-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, nov 2016. Funding Information: This work has been supported by the Israeli Science Foundation (ISF) under Grant 12/12, by NSF Grant CCF-1016625, by the Center for Science of

Information, an NSF Science and Technology Center under Grant CCF-0939370, and by ARO under MURI Grant W911NF-15-1-0479.

[Sason 2018] Igal Sason. On f-divergences: Integral representations, local behavior, and inequalities. *Entropy*, 20(5):383, 2018.

[Shannon 1948] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[Shlens 2014] Jonathon Shlens. Notes on kullback-leibler divergence and likelihood. *arXiv preprint arXiv:1404.2000*, 2014.

[Skrondal and Rabe-Hesketh 2007] Anders Skrondal and Sophia Rabe-Hesketh. Latent variable modelling: A survey. *Scandinavian Journal of Statistics*, 34(4):712–745, 2007.

[Soviany *et al.* 2020] Petru Soviany, Claudiu Ardei, Radu Tudor Ionescu, and Marius Leordeanu. Image difficulty curriculum for generative adversarial networks (cugan). In *The IEEE Winter Conference on Applications of Computer Vision*, pages 3463–3472, 2020.

[Suter *et al.* 2019] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065. PMLR, 2019.

[Tishby *et al.* 2000] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[Tolstikhin *et al.* 2017a] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[Tolstikhin *et al.* 2017b] Ilya O Tolstikhin, Sylvain Gelly, Olivier Bousquet, Carl-Johann Simon-Gabriel, and Bernhard Schölkopf. Adagan: Boosting generative models. In *Advances in Neural Information Processing Systems*, pages 5424–5433, 2017.

[Tschannen *et al.* 2019] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019.

[Wan *et al.* 2020] Neng Wan, Dapeng Li, and Naira Hovakimyan. f-divergence variational inference. *Advances in Neural Information Processing Systems*, 33, 2020.

[Wang *et al.* 2017] Kunfeng Wang, Chao Gou, Yanjie Duan, Yilun Lin, Xinhu Zheng, and Fei-Yue Wang. Generative adversarial networks: introduction and outlook. *IEEE/CAA Journal of Automatica Sinica*, 4(4):588–598, 2017.

[Wang *et al.* 2019] Zhengwei Wang, Qi She, and Tomas E. Ward. Generative adversarial networks: A survey and taxonomy. *CoRR*, abs/1906.01529, 2019.

[Wolfowitz 1965] Jacob Wolfowitz. Asymptotic efficiency of the maximum likelihood estimator. *Theory of Probability & Its Applications*, 10(2):247–260, 1965.

[Yang 2017] Xitong Yang. Understanding the variational lower bound. *variational lower bound, ELBO, hard attention*, 13:1–4, 2017.

[Zhang *et al.* 2018] Cheng Zhang, Judith Bütepage, Hedvig Kjellström, and Stephan Mandt. Advances in variational inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018.

[Zhang *et al.* 2019] Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, and David Barber. Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*, 2019.

[Zhao *et al.* 2017] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[Zhu *et al.*] Kongtao Zhu, Xiwei Liu, and Hongxue Yang. A survey of generative adversarial networks. In *2018 Chinese Automation Congress (CAC)*, pages 2768–2773. IEEE.