GCVAE: Generalized-Controllable Variational AutoEncoder

Kenneth Ezukwoke, Anis Hoayek, Mireille Batton-Hubert, Xavier Boucher

HAL Id: emse-03712594
https://hal-emse.ccsd.cnrs.fr/emse-03712594v1

GCVAE: Generalized-Controllable Variational AutoEncoder

Kenneth Ezukwoke[*], Anis Hoayek[1], Mireille Batton-Hubert[1], and Xavier Boucher[2]

[1]Department of Mathematics and Industrial Engineering
[2]Department of Organisation and Environmental Engineering
[1,2]Mines Saint-Etienne, Univ Clermont Auvergne, INP Clermont Auvergne, CNRS, UMR 6158
LIMOS, F - 42023 Saint-Etienne, France

{ifeanyi.ezukwoke, anis.hoayek, batton, boucher}@emse.fr

Abstract

Variational autoencoders (VAEs) have recently been used for unsupervised disentanglement learning of complex density distributions. Numerous variants exist to encourage disentanglement in latent space while improving reconstruction. However, none have simultaneously managed the trade-off between attaining extremely low reconstruction error and a high disentanglement score. We present a generalized framework to handle this challenge under constrained optimization and demonstrate that it outperforms state-of-the-art existing models as regards disentanglement while balancing reconstruction. We introduce three controllable Lagrangian hyperparameters to control reconstruction loss, KL divergence loss and correlation measure. We prove that maximizing information in the reconstruction network is equivalent to information maximization during amortized inference under reasonable assumptions and constraint relaxation.

1 Introduction

Unsupervised learning describes an intelligent approach for extracting meaningful classes of decisions from unlabeled data using intrinsic properties such as similarity metrics. Numerous classical data-driven approaches based on multivariate statistical analysis including Principal Component Analysis [Karl, 1901, Hoffmann, 2007, Song et al., 2014] and Canonical Correlation Analysis [HOTELLING, 1936] have been used to obtain latent or subspace patterns of high correlation. Decision labels are then obtained by applying clustering algorithms [Martin et al., 1996, Breunig et al., 2000, Andrade et al., 2013, Liu et al., 2019, Ren et al., 2020]. Recently, deep learning-based approaches have gained traction for their improved performance at learning better latent representations with a high disentanglement metric [Hinton and Salakhutdinov, 2006] and have evolved into a new dimension of learning known as *Disentangled Representation Learning* [Bengio et al., 2014].

Disentanglement learning is a branch of unsupervised deep representation learning which involves uncovering the underlying independent factors responsible for making up an observation. For instance, an image with high dimensional features (eyes, nose, mouth, hair etc.), can be projected into a lower dimensional space of high disentanglement where different factored features of the images can be easily identified visually or clustered by an algorithm. Autoencoders are a specific type of a neural network, designed to encode the input into a compressed and meaningful representation, and then decode it back, so that the reconstructed input is as similar as possible to the original [Bank et al., 2021]. While learning the reconstructed space, autoencoders are not capable of inducing disentanglement in the latent space, hence, keeping the latent factors entangled.

Variational Autoencoders (VAE) [Kingma and Welling, 2014] are an improved generative version of autoencoder based on probabilistic graphical modeling. VAEs replace the encoder section of an autoencoder with an amortized inference process (from Bayesian machine learning) and sample the latent space using a reparameterization trick before reconstruction. VAEs have been employed for text disentanglement [John et al., 2018] and generation [Shen et al., 2019], as well as image generation [Kingma and Welling, 2014, Kulkarni et al., 2015, Gregor et al., 2015]. Despite the numerous benefits of VAEs, they experience certain drawbacks including the notable vanishing Kullback-Leibler (KL) divergence problem. However, β-VAE [Higgins et al., 2017] solves the challenges arising from using VAE by imposing a weight ($\beta \gg 1$) on KL-divergence, thus eliminating the problem. This is at the expense of poor reconstruction which leaves room for improvement. InfoGAN [Chen et al., 2016] belongs to a family of Generative Adversarial Networks (GAN) [Goodfellow et al., 2014], and involves an adversarial training of two models in a zero-sum game with the objective of inducing disentanglement in the latent space by maximizing variational mutual information. Despite the improved representation learned from using InfoGAN, a major drawback, as with all GANs, is the instability in training and strong assumption as regards the underlying distribution of the data.

It is imperative therefore to improve the amortized inference process of generative VAE networks which induces disentanglement with good reconstruction. InfoVAE [Zhao et al., 2018] takes a cue from InfoGAN by introducing a weighted mutual information term within the original objective of the VAE. While it proves to be an improvement in disentanglement and representation compared to VAE and β-VAE, reconstruction is still not optimal. It is similar in performance to FactorVAE [Kim and Mnih, 2019], whose objective also includes mutual information and a directional KL-divergence in the latent dimension. FactorVAE uses a discriminating network to compare the performance of the latent space at every epoch, achieving better disentanglement compared to its counterparts. A challenging question to answer when using FactorVAE is the choice of hyperparameter γ capable of inducing maximum disentanglement in the latent space while balancing reconstruction. ControlVAE [Shao et al., 2020] on the other hand tries to solve the problem of selecting the appropriate β automatically by introducing a non-linear Proportional-Integral-Derivative (**PID**) controller, to adaptively control the β hyperparameter.

In all the methods proposed in the literature, very little has been done in trying to adapt a model that encourages disentanglement while minimizing reconstruction error in the VAE model simultaneously. We propose a generalized framework similar that emphasizes information maximization in the reconstruction network rather than in the inference network, and tune the hyperparameters using the non-linear PID controllers proposed in controlVAE [Shao et al., 2020]. The proposed model is designed to maximize the mutual information between the reconstructed data and the latent spaces by adaptive weighting of the reconstruction term of the VAE. We show that factorizing latent space or increasing the weight of β on KL-divergence is not entirely sufficient for disentanglement. Our proposed model identifies the failures of existing models and introduces the squared Mahalanobis distance as a heuristic to improve independence in latent space.

The remaining sections of the paper are as follows: Section two discusses variational autoencoders and their variants; section three presents the proposed generalized-controllable variational autoencoder; section four contains experimental results; section five presents the final conclusion and finally the acknowledgements.

2 Variational Autoencoder (VAE)

VAEs [Kingma and Welling, 2014] are a class of generative model proposed to model complex distributions existing in images, natural language and functional data. VAEs serve a significant purpose in language modeling, especially in representation learning for text disentanglement, unsupervised abstractive sentence summarization [Schumann, 2018] and long and coherent text generation [Shen et al., 2019] among others. We formally define the VAE model by observing a d-dimensional input space $\{x_i\}_{i=1}^{N} \in \mathcal{X}$ consisting of N-independently and identically distributed (**i.i.d**) samples; k-dimensional latent space $\{z_i\}_{i=1}^{N} \in \mathcal{Z}$ (where $k \ll d$) over which a generative model is defined. We assume an empirical prior distribution $p_\theta(z) \sim \mathcal{N}(0, I)$ to infer an approximate posterior distribution $q_\phi(z|x) \sim \mathcal{N}(z|\mu_\phi(x), \sigma_\phi^2(x)I)$, with mean $\mu_\phi(x)$ and variance $\sigma_\phi^2(x)I$ used for reparameterization sampling of the latent space z [Kingma and Welling, 2014]. We model the data using conditional distribution $p_\theta(x|z) \sim \mathcal{N}(x|\mu_\theta(x), \sigma_\theta^2(x)I)$. Let us suppose that the underlying distribution of the input space $p_(x)$ follows a normal distribution, and its empirical distribution is denoted by $p_\mathcal{D}(x)$.

The objective function $\mathcal{L}(\theta, \phi)$ to be maximized is given by,

$$\mathbb{E}_{p_{\mathcal{D}}} \mathbb{E}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p_\theta(z)) \tag{1}$$

where ϕ and θ represent the parameters of the neural network encoder and decoder respectively. This objective function is known as the **E**vidence **L**ower **Bo**und (**ELBO**), upper bounded by the input data log-likelihood $\ln p_\theta(x)$. The first term of the objective is a *reconstruction error*. Ideally, maximizing negative reconstruction loss, simultaneously minimizes non-symmetric divergence between the distributions of the *Kullback-Leibler (KL)-divergence* terms.

Equivalently, maximizing ELBO maximizes negative KL-divergence between the inference prior $q_\phi(z|x)$ and generative prior $p_\theta(z|x)$, where $p_\theta(z|x) \propto p_\theta(x, z)$ and $q_\phi(z|x) \propto q_\phi(x, z)$.

$$\mathcal{L}(\theta, \phi) = -\mathbb{E}_{p_{\mathcal{D}}}[D_{KL}(q_\phi(z|x)||p_\theta(z|x))] \tag{2}$$

$$\ln p_\theta(x) \geq \mathbb{E}_{p_{\mathcal{D}}}\left[\ln p_\theta(x|z) - D_{KL}(q_\phi(z|x)||p_\theta(z))\right] \tag{3}$$

We prove equation (3) in Appendix A.1.

β-**VAE** [Higgins et al., 2017] introduces a fixed weight on the KL divergence (βD_{KL}) with the aim of (i) eliminating the vanishing KL problem (where $D_{KL} \to 0$) and (ii) inducing disentanglement by adjusting the value of β.

2.1 ControlVAE

Due to the difficulty in modeling complex entangled functions and the lack of statistical independence of the latent variables, [Higgins et al., 2017] has proposed an adjustable β-hyperparameter that (i) balances latent bottlenecks and independence with an improved reconstruction compared to VAE (ii) solves the vanishing KL-divergence problem where $D_{KL}(q_\phi(z|x)||p_\theta(z)) = 0$.

$$\mathbb{E}_{p_{\mathcal{D}}} \mathbb{E}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)) \tag{4}$$

β-VAE ensures that the hyperparameter on the KL increases the independence of the latent variables for very large values of β ($\beta > 1$) while ensuring posterior conditional inference probability $q_\phi(z|x)$ is relatively close to its prior $p_\theta(z)$. However, there is no a priori limit for β as the value of β depends on the data; which results in overfitting of the prior to the approximate posterior $p_\theta(z) \approx q_\phi(z|x)$. This implies that the information about x preserved in z decreases with $\beta \to +\infty$. **ControlVAE** [Shao et al., 2020] proposes the use of a PID controller for an adaptive β. Its objective function is as follows,

$$\mathbb{E}_{p_{\mathcal{D}}} \mathbb{E}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - \beta_t D_{KL}(q_\phi(z|x)||p_\theta(z)) \tag{5}$$

Where,

$$\beta_t = \frac{K_p}{1 + exp(e_t)} - K_i \sum_{j=0}^{t} e_j + \beta_{min} \tag{6}$$

where K_p and K_i are constants; e_t is the error between the actual KL-divergence value and the desired value at time t; β_{min} is an application-specific constant which shifts the range to within which β_t is allowed to vary.

2.2 InfoVAE

InfoVAE [Zhao et al., 2018] adds a weighted mutual information criteria to the VAE objective function and maximizes the mutual information between the data x and latent space z. The objective function of InfoVAE is given by,

$$\mathbb{E}_{p_{\mathcal{D}}} \mathbb{E}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - (1 - \alpha)D_{KL}(q_\phi(z|x)||p_\theta(z)) - (\alpha + \lambda - 1)D_{KL}(q_\phi(z)||p_\theta(z)) \tag{7}$$

which is reduced to a family of variational autoencoding objectives when $\alpha = 1$,

$$\mathbb{E}_{p_{\mathcal{D}}} \mathbb{E}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - \lambda D_{KL}(q_\phi(z)||p_\theta(z)) \tag{8}$$

3

Where $D_{KL}(q_\phi(z)||p_\theta(z))$ is a non-negative distance called Maximum Mean Discrepancy (**MMD**) and is defined as a kernel function,

$$D_{MMD}^2(q||p) = \mathop{\mathbb{E}}_{p(z)\cdot p(z')}[k(z,z')] - 2\mathop{\mathbb{E}}_{q(z)\cdot p(z')}[k(z,z')] + \mathop{\mathbb{E}}_{q(z)\cdot q(z')}[k(z,z')] \tag{9}$$

In Equation (9), $k(z,z')$ represents a positive semi-definite kernel matrix [Steinwart and Christmann, 2008]. $D_{MMD}^2(q||p)$ measures the distance between the sample distribution z and true distribution $z' \sim \mathcal{N}(0,I)$ and $\lambda > 0$ is a scaling coefficeint [Zhao et al., 2018].

3 Generalized-Controllable VAE

3.1 Alternate formulation of ELBO, β-VAE and InfoVAE objective

We consider an alternate formulation of the InfoVAE objective based on mutual information maximization. Let $I_p(x',z)$ be the mutual information between the reconstructed data x' and z under joint distribution $p_\theta(x'|z)p_\theta(z)$, where $p_\theta(z) \sim \mathcal{N}(0,I)$. $I_q(x,z)$ is the mutual information between x and z under joint distribution $q_\phi(z|x)p_\mathcal{D}(x)$. Since the objective of the VAE is to minimize a certain reconstruction loss, and considering that this loss is an l_2-norm, we assume that the expected log-likelihood $\mathbb{E}[\ln p_\theta(x'|z)]$ is approximately equal to $\mathbb{E}[\ln p_\theta(x|z)]$ and $p_\mathcal{D}(x') \approx p_\mathcal{D}(x)$ at the optimal point for n-number of experiments.

Proposition 1. Let $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ be random variables with $q_\phi(x)$ and $q_\phi(z)$ corresponding to their marginal probability density distribution. We denote the joint probability density function between x and z as $q_\phi(x,z)$ where $q_\phi(z) \sim \mathcal{N}(\mu_\phi, \sigma_\phi^2 I)$ and $p_\theta(x) \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2 I)$. For known parameters of x and z, the mutual information between x and z for the inference network is expressed as

$$I_q(x,z) = \int_x \int_z q_\phi(x,z) \ln \frac{q_\phi(x,z)}{q_\phi(z)q_\phi(x)} \, dx \, dz = D_{KL}(q_\phi(x,z)p_\theta(z)||q_\phi(z)q_\phi(x)p_\theta(z)) \tag{10}$$

$$= D_{KL}(q_\phi(z|x)||p_\theta(z)) - D_{KL}(q_\phi(z)||p_\theta(z)) \tag{11}$$

Furthermore, we can reformulate Equation (10) by introducing the joint generative distribution so that,

$$I_q(x,z) = D_{KL}(q_\phi(z|x)p_\theta(x,z)||q_\phi(z)p_\theta(x,z)) \tag{12}$$

$$= D_{KL}(q_\phi(z|x)||p_\theta(z)) - D_{KL}(q_\phi(z)||p_\theta(z)) - \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] + \ln p_\theta(x) \tag{13}$$

The complete derivation is found in the Appendix (section A.2). This maximization is also known to improve reconstruction of generative models, as seen for GANs [Belghazi et al., 2018] while the minimized term ensures an effective information bottleneck. The constraint optimization formulation used here will be applied in a later section.

Mutual information from the generative network is given as,

$$I_p(x',z) = D_{KL}(p_\theta(x',z)||p_\theta(z)p_\theta(x')) = \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)}[\ln p_\theta(x'|z)] - \ln p_\theta(x') \tag{14}$$

The marginal log-likelihood $\ln p_\theta(x')$, given as $\int_z p_\theta(x'|z)p_\theta(z)dz$ with an integral over z is computationally intractable for a large number of z-variables. We can drop the marginal since it does not significantly impact the likelihood term. We reformulate the objective function of a family of InfoVAE (Equation 8) as a function of the mutual information of both generative and reconstruction processes as follows,

$$\max_{\phi,\theta} \quad I_p(x',z) - I_q(x,z) \tag{15}$$

We call this formulation (Equation 15), the total mutual information of a VAE, which is further explained in Figure 1. This implies therefore, that accurately reconstructing the original distribution $p_\mathcal{D}(x)$ requires us to maximize the mutual information $I_p(x',z)$ in the reconstructed space while reducing information loss $I_q(x,z)$ during inference. The constraint optimization formulation of Equation (15) can be written as,

$$\begin{aligned} \max_{\phi,\theta} \quad & I_p(x',z) \\ s.t \quad & D_{KL}(q_\phi(z|x)||p_\theta(z)) \leq \xi_1 \\ s.t \quad & D_{KL}(q_\phi(z)||p_\theta(z)) \leq \xi_2 \\ & \xi_1, \xi_2 \geq 0 \end{aligned} \tag{16}$$

ξ_1 and ξ_2 in (16) are permissible error terms. We apply the Karush–Kuhn–Tucker (KKT) conditions [Karush, 1939, Kuhn and Tucker, 1951] and solve for optimality, by applying the Lagrangian multipliers to obtain the following,

$$\max_{\phi,\theta} \quad \mathcal{L}(\phi,\theta,\lambda,\beta) \tag{17}$$

$$\max_{\phi,\theta} \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} [\ln p_\theta(x|z)] - \beta D_{KL}(q_\phi(z|x)||p_\theta(z)) - \lambda D_{KL}(q_\phi(z)||p_\theta(z)) \tag{18}$$

$$\mathcal{L}(\phi,\theta,\lambda,\beta) = \begin{cases} ELBO & \text{if } \beta = 1, \lambda = 0 \\ \beta - VAE & \text{if } \beta > 1, \lambda = 0 \\ InfoVAE & \text{if } \beta = 0, \lambda \geq 1 \\ FactorVAE & \text{if } \beta = 1, \lambda = -1 \end{cases} \tag{19}$$

We derive from Equation (15), that maximizing mutual information under reasonable constraints in the generative network of VAE is equivalent to mutual information minimization in the inference network. Hence, maximizing $I_p(x', z)$ subject to inference constraints is equivalent to maximizing the total mutual information. The constraints to which mutual information is subjected in the InfoVAE optimization framework does not simultaneously balance the high disentanglement-low reconstruction error trade-off. In the next section, we propose an optimization framework based on information maximization in a generative network, that simultaneously manages the trade-off between high disentanglement and loss reconstruction error.
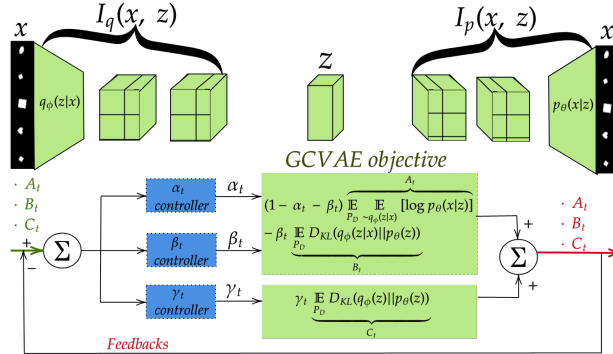
3.2 GCVAE: Generalized-Controllable VAE

Figure 1: GCVAE framework. α_t, β_t and γ_t respectively provide automatic balancing of the log-likelihood and KL divergences for optimal reconstruction and disentanglement. The feed-ins A_t, B_t and C_t are expectations of variational loss.

We propose a generalized framework for variational inference modeling, taking into account the idea of information maximization in the reconstruction network. We prioritize the disentanglement of the latent space and balancing the trade-off between disentanglement metric and reconstruction loss by maximizing the mutual information between that reconstructed data x' and latent space z. Furthermore, we show that weighting the reconstruction term of this new lower bound by the controllable weight of its KL-divergence significantly improves the disentangling factor. Under the constraint optimization framework, we follow a similar reformulation, as shown in Equation (16), to propose the following formal problem.

Proposition 2. Let $x \in \mathcal{X}$ and $z \in \mathcal{Z}$ be continuous spaces representing the input and latent space respectively. $I_p(x', z)$ is the joint mutual information space between x' and z generated from the posterior $p_\theta(x|z)$ after obtaining an inference posterior $q_\phi(z|x)$. Let $0 \leq \beta_t \leq \alpha_t \leq 1$ be controllable optimizable hyperparameters (Lagrangian multipliers). We reach an optimum when $q_\phi(z|x) \approx p_\theta(z|x)$. We maximize the mutual information in the reconstruction space subject to

5

inference constraints as follows,

$$\max_{\theta,\phi,\xi^+,\xi^-,\xi_p \in \mathbb{R}} \quad I_p(x',z)$$

$$s.t \quad \mathbb{E}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z|x) \parallel p_\theta(z)) + I_p(x',z) \leq \xi^-$$

$$s.t \quad - \mathbb{E}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z)||p_\theta(z)) \leq \xi^+$$

$$s.t \quad I_p(x',z) \leq \xi_p$$

$$s.t \quad \xi_i^+, \xi_i^-, \xi_{ip} \geq 0, \quad \forall i = 1, \ldots, n \tag{20}$$

We apply the KKT conditions and solve for optimality as previously seen in section 3.1 (proof provided in the Appendix A.3). Hence,

$$\mathcal{L}(\theta,\phi,\xi^+,\xi^-,\xi_p,\alpha,\beta,\gamma) = (1 - \alpha_t - \beta_t) \mathbb{E}_{p_{\mathcal{D}}} E_{z \sim q_\phi(z|x)}[\ln p_\theta(x|z)] - \beta_t \mathbb{E}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z|x)||p_\theta(z))$$

$$+ \gamma_t \mathbb{E}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z)||p_\theta(z)) \tag{21}$$

α_t, β_t and γ_t and Lagrangian multipliers and are used here as controllable hyperparameters. The controllable weight on the reconstruction loss $1 - \alpha_t - \beta_t$, implies the effect of the KL-divergence loss on the reconstruction loss, which simultaneously control the disentanglement in the latent space with respect to the reconstruction quality. Empirical results show that this new lower bound is free from the vanishing KL-divergence problem even for fixed β ($\beta \gg 1$) and does not overfit during training. Exponentially small values of $\alpha, \beta \to 0$ penalize reconstruction error and KL divergence respectively. The last term in (21), $D_{KL}(q_\phi(z)||p_\theta(z))$ is the expected Mahalanobis distance (MD) denoted as $\mathbb{E}D_{MAH}^2(q||p)$ between the density function of two continuous variables p and q as opposed to MMD used in InfoVAE. MD is a positive semi-definite distance and obeys the triangle inequality. We show that MD distance is a better disentangling metric than MMD with better reconstruction quality.

Proposition 3. Let $x \in \mathcal{X}$ be independent random variables with distribution $p(x)$ and similarly $y \in \mathcal{Y} \sim q(y)$. x' and y' are transposes of x and y respectively. We suppose that x and y are mapped to a Reproducing Kernel Hilbert Space (RKHS) denoted as $\mathcal{H}$ [Aronszajn, 1950] so that $x \sim \varphi(x)$ and $y \sim \varphi(y)$ with a common covariance matrix $\Sigma_{d \times d}$. Let $z := \{x_i, y_i\}_{i=1}^n \in \mathcal{Z}$ be n i.i.d random samples drawn from a distribution $\mathcal{D}_z$. The expected squared Mahalanobis distance between distributions p and q in a reproducing kernel Hilbert space is given by $\Sigma^{-1} D_{MMD}^2(q||p)$, where Σ^{-1} is the diagonal covariance.

$$\mathbb{E}D_{MAH}^2(q||p) = \mathbb{E}[(\varphi(x) - \varphi(y))' \Sigma^{-1}(\varphi(x) - \varphi(y))]$$

$$\geq \mathbb{E}[tr(\Sigma^{-1}(\varphi(x) - \varphi(y))'(\varphi(x) - \varphi(y)))]$$

$$\geq tr\left(\parallel \mathbb{E}_{x \sim p} \varphi(x) - \mathbb{E}_{y \sim q} \varphi(y) \parallel_{\Sigma^{-1}, \mathcal{H}}^2 \right)$$

$$\geq tr(\Sigma^{-1} D_{MMD}^2(q||p)) \tag{22}$$

The expectation of the right is $\geq tr(\mathbb{E}\Sigma^{-1} D_{MMD}^2(q||p))$. The proof of Equation (22) is given in Appendix A.4. $\mathbb{E}D_{MAH}^2(q||p)$ is a measure of the average dissimilarity between the distributions p and q in the Hilbert space, as it normalizes the similarity with feature variances, therefore encouraging class discrimination. $\mathbb{E}D_{MAH}^2(q||p)$ reduces to $D_{MMD}^2(q||p)$ when Σ^{-1} is identity. We note that x is zero-centered before it is projected into the Hilbert space.

We deduce other variants of VAE from GCVAE loss as follows,

$$\mathcal{L}(\theta,\phi,\xi^+,\xi^-,\xi,\alpha,\beta,\gamma) = \begin{cases} ELBO & \text{if } \alpha_t = \alpha = -1, \beta_t = \beta = 1, \gamma = 0 \\ ControlVAE & \text{if } \alpha_t = \alpha = 0, \beta_t > 0, \gamma = 0 \\ InfoVAE & \text{if } \alpha_t = \alpha = 0, \beta_t = \beta = 0, \gamma_t > 1 \\ FactorVAE & \text{if } \alpha_t = \alpha = -1, \beta_t = \beta = 1, \gamma = -1 \end{cases} \tag{23}$$

While the above models focus on penalizing KL-divergence to induce disentanglement, GCVAE takes a parameter into account to control reconstruction loss (α_t) while gradually improving disentanglement. GCVAE achieves optimality for $\alpha_t \in [0, 1 - \beta_t], \beta_t \in [0, 1 - \alpha_t], \gamma_t \in [0, 1]$. This new objective function allows us to significantly control the amount of meaningful information encoded in the latent space.

4 Experiment

4.1 Experimental setup and parameter optimization

We design and select an experimental setup that places emphasis on the relevance of disentanglement without overlooking the issue of reconstruction trade-off encountered by some state-of-the art models. We validate the performance of the proposed model both qualitatively (reconstruction) and quantitatively (disentanglement), using a simplified version of [Kim and Mnih, 2019] neural architecture. We train for $250K$ epochs with batch of 64 and a learning rate of $1e^{-3}$. The number of dimensions of the latent z is 10 for the DSprites dataset and 6D for the 3D Shapes dataset. We prefer dimension 2 to analyze the reconstruction and generative quality of the model. In all cases, the Adam optimizer is used for training the models.

Model architecture. The preferred model is a 4-layer, 2D convolutional neural network with (2, 2) filters. A full description of the network architecture is available in Appendix B.1 of this paper.

Parameter optimization. The GCVAE lower bound can be reduced to the ELBO of other variational models, while the choice of values for the hyperparameters can be selected according to the original papers for the training phase. InfoVAE only enables the λ weight on the MMD to be optimized ($\lambda = 1000$ is used in our experiment). The value of $\beta = 10$ is used for training β-VAE. ControlVAE optimizes the β and we set an expected $\beta = 10$ for training. However, the GCVAE model requires the optimization of three PID-controllable parameters α_t, β_t and γ_t. The expected α, β and γ hyperparameters for GCVAE are set to $10, 30, 0.1$ before training. These set points are selected after extensive trial and errors. To evaluate the strength of disentanglement and the quality of reconstruction, we propose three families of GCVAE according to the metric selected for the $D_{KL}(q_\phi(z)||p_\theta(z))$:

(1) **GCVAE-I**: $D_{KL}(q_\phi(z)||p_\theta(z)) = D^2_{MMD}(q||p)$;
(2) **GCVAE-II**: $D_{KL}(q_\phi(z)||p_\theta(z)) = D^2_{MAH}(q||p)$
(3) **GCVAE-III**: $D_{KL}(q_\phi(z)||p_\theta(z)) = \mathbb{E}\Sigma^{-1}D^2_{MMD}(q||p)$.

4.2 Dataset

MNIST [Deng, 2012] is a 28×28-dimensional image data set with $70,000$ observation. The data set is split into training observations ($60,000$) and testing observations ($10,000$) and contains images of numbers between $0 - 9$ values together with their labels.

DSprites 2D Shapes DSprites [Matthey et al., 2017] is a data set of 2D 64×64-dimensional image data set with 737,280 binary observations. Five ground truth factors are available with properties (shape: 3; scale: 6; orientation: 40; x-position: 32; y-position: 32).

4.3 Evaluation metric

Mutual Information Gap (MIG) score [Chen et al., 2019]. Normalized mutual information gap between the top two latent factors $(I(y_k, z_I) - I(y_k, z_{II}))/(\sum_{j=1}^{d} I(y_i, z_j))$. The average MIG score is taken by normalizing the total mutual information. The score reports the compactness of the latent code by ensuring that the information contained in a fixed ground truth y_k is expressed by only one latent factor z_j at a time. High values imply a high level of disentanglement in the latent space. **Joint Entropy Minus Mutual Information Gap (JEMMIG) score** [Do and Tran, 2021]. This indirectly measures the modularity in latent space, since a single latent factor may explain more than one ground truth factor. It is expressed as $H(y_k, z_I) - I(y_k, z_I) + I(y_k, z_{II})$. A lower JEMMIG score is preferred or a high $(1 - \text{JEMMMIG})$ score. The average JEMMIG score $(\frac{1}{K}\sum_{k=0}^{K-1} JEMMIG(y_k))$ quantifies the interpretability of the latent variables, measuring both its compactness and its explicitness. **Modularity score** [Ridgeway and Mozer, 2018] expresses the number of latent factors z_j with high mutual information and explains the ground truth factors.

4.4 Quantitative and Qualitative Evaluation

Improving disentanglement in latent space without compromising reconstruction, interpretatbility and informativeness is challenging yet achievable. We present a generalized controllable model that automatically adjusts to learning parameters seeking to increase disentanglement, while reducing reconstruction error. We also demonstrate the strength of this model for generative tasks, using well-known data sets and comparing it with benchmark models. In the rest of this paper, we focus on information-based metrics like MIG (compactness), JEMMIG (explicitness) and Modularity

	MIG ↑	Modularity ↑	JEMMIG ↑	Reconstruction loss ↓	KL loss ↗
VAE	0.1268	0.798	0.233	3.339	3.0025
β-VAE	0.0778	**0.881**	0.238	**0.012**	35.0295
ControVAE	0.1213	0.782	0.312	0.016	24.3809
InfoVAE	0.1501	0.757	0.188	0.079	10.0621
GCVAE-I	0.1507	0.844	0.236	**0.012**	24.3739
GCVAE-II	**0.2793**	0.858	**0.312**	**0.012**	**24.4316**
GCVAE-III	0.1337	0.825	0.294	0.015	24.2937

Table 1: Performance comparison of different models on **DSprites** after training on 737 samples. Comparison metrics MIG [Chen et al., 2019], Modularity [Ridgeway and Mozer, 2018] and JEMMIG [Do and Tran, 2021] for 10-D Latent representation. The direction of the arrow indicates the best performing model. Higher is better for MIG, Modularity and JEMMIG($1 - JEMMIG$). **GCVAE-II** performs best on MIG disentanglement metric, robustness and interpretability; plus having the lowest reconstruction error. **GCVAE-I, III** and **ControlVAE** also measure up in disentanglement compared to other benchmark models. **GCVAEs** have the least reconstruction error partly due to the normalization introduced by the inverse precision matrix in $D_{KL}(q_\phi(z)||p_\theta(z))$ and the weight on the first term of the GCVAE loss 21

(robustness) scores, as they are more robust for unsupervised scenarios. The information based performance metric from Table 1 measures the quality of disentanglement per model. The MIG metric indicates the best performing model is GCVAE-II and GCVAE-I respectively.

The GCVAE-II and III both require more computational time per epoch as a result of the approximate cubic time $O(n^3)$ of inverting the precision matrix Σ. The conditional independence of variables i, j in z noted $\Sigma_{z_{ij}} > 0$ given a third variable increases the precision when estimating the partial correlation of i, j given the rest of the variables. The higher the diagonal covariance, the more stable the approximation of $D_{KL}(q_\phi(z)||p_\theta(z))$ and robust estimate of $I(y_k, z)$, hence, an 85% increase in estimation of MIG for GCVAE-II compared to I. This represents an advantage over other models, as we now know that a robust estimation of the conditional independence of variables in latent z induces disentanglement.

The reconstruction quality for GCVAE-II is shown to the the least indicating high level of clarity in the reconstruction of the observation.
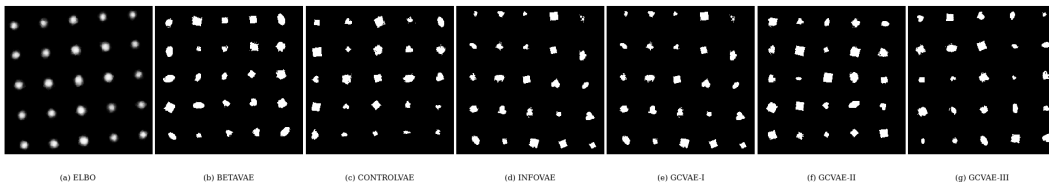
4.5 Generation

Figure 2: Generative process comparison for the different models on DSprites after training on less than 800 samples. Model GCVAE-I, II and ControlVAE clearly outperformed other models. The reconstruction error of GCVAE-II is the lowest from Table 1

We evaluate the quality of generation by considering the explicitness and coherency of the encoded latent variables. The generation by GCVAE-I, II and III far outperformed those of the benchmark models (especially on the MNIST dataset). It is worthy of mention that, increasing KL-divergence does not directly correlate to increasing the MIG disentanglement metric (Table 1). While the Mahalanobis correlation metric is not entirely responsible for disentanglement, it serves as a good heuristic for improving generative quality as well as disentanglement observed in GCVAE-II.

4.6 Comparing GCVAE-1, II & III

We compare performance of GCVAEs both on reconstruction and disentanglement using a stopping criteria algorithm. Using a fixed ϵ for $\alpha_t = 10e - 5$ and $\beta_t = 10e - 4$, we obtain reasonably low

(a) ELBO (b) BETAVAE (c) CONTROLVAE (d) INFOVAE (e) GCVAE-I (f) GCVAE-II (g) GCVAE-III

Figure 3: Generative process comparison for the different models after training the MNIST dataset for 500 epochs. GCVAE-II and ELBO (VAE) have a similar reconstruction quality with better interpretation. GCVAE-II clearly outperformed the benchmark models in generating clear and meaningful representations of the original data. β-VAE is the least performing in terms of generating an interpretable image of the original data. ControlVEA and InfoVAE generate meaningful representations.
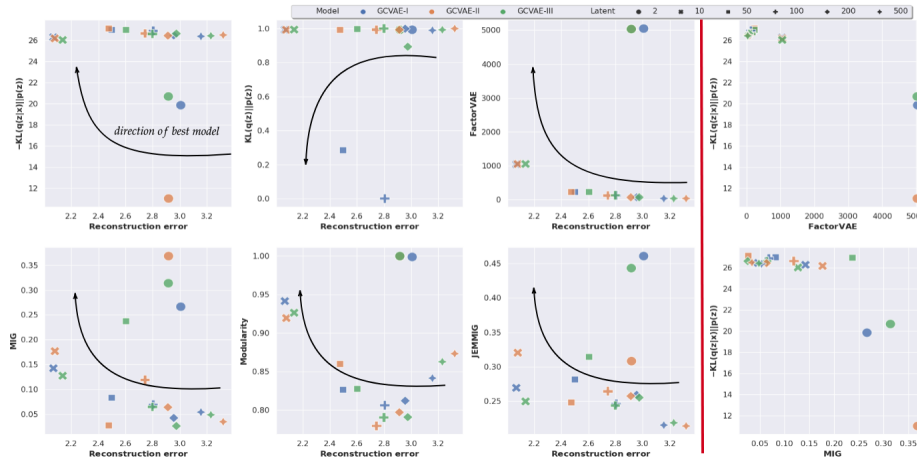

Figure 4: Model performance comparison on 737 samples of DSprites data. Arrows indicate direction of best performing model. **Top**: Comparison of reconstruction error against KL divergence , $D_{KL}(q_\phi(z|x)||p_\theta(z))$ and correlation, $D_{KL}(q_\phi(z)||p_\theta(z))$. High KL-Low reconstruction error observed for Latent-10. Low reconstruction error implies high disentanglement, while high KL does not. High disentanglement using FactorVAE is observed on Latent-2 followed by Latent 10. **Bottom**: Comparing disentanglement metrics with reconstruction loss. The highest disentanglement on the MIG metric is observed for GCVAE-**II** on Latent-2, however, the best scores are observed for GCVAE-II on Latent-10. JEMMIG is similar in behavior to MIG. **RHS**: Validating the statement *Latent disentanglement is not correlated to KL maximization.*

reconstruction loss with high disentanglement without having to train for a lengthy period. The average time required to train GCVAE using the stopping criterion is 6 hours while it takes more than 3days to train for 250K iterations. The performance of the model compared for increasing latent dimension is illustrated in Figure 4.

5 Conclusion

The challenge of latent space disentanglement for variational models often comes with finding the right reconstruction trade-off. As a result, we propose a new lower bound we refer to as Generalized-Controllable Variational Autoencoder (GCVAE: a model built from a constraint optimization perspective to maximize mutual information in the generative phase, subject to inference constraints to encourage disentanglement in latent space. We use the Mahalanobis distance metric as a heuristic to encourage independence between latent space variables and show that the representation obtained GCVAE is both meaningful and interpretable, with low reconstruction loss. GCVAE-II demonstrates notable strength in disentangling latent space and reconstructing with minimal mutual information loss compared to other variants.

References

Guilherme Andrade, Gabriel Ramos, Daniel Madeira, Rafael Sachetto, Renato Ferreira, and Leonardo Rocha. G-dbscan: A gpu accelerated algorithm for density-based clustering. *Procedia Computer Science*, 18:369–378, 2013. ISSN 18770509. doi: 10.1016/j.procs.2013.05.200.

N. Aronszajn. Deep learning-based clustering approaches for bioinformatics. *Transactions on American Mathematical Society*, 68:337–404, 1950. doi: 10.1090/S0002-9947-1950-0051437-7. URL https://doi.org/10.1090/S0002-9947-1950-0051437-7.

Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/belghazi18a.html.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives, 2014.

Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. Lof: Identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, page 93–104, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335388. URL https://doi.org/10.1145/342009.335388.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets, 2016.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations, 2021.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation, 2015.

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

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

Heiko Hoffmann. Kernel pca for novelty detection. *Pattern Recognition*, 40(3):863–874, 2007. ISSN 00313203. doi: 10.1016/j.patcog.2006.07.009.

HAROLD HOTELLING. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28(3-4):321–377, 12 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL https://doi.org/10.1093/biomet/28.3-4.321.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. Disentangled representation learning for non-parallel text style transfer, 2018.

Pearson Karl. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559–572, 1901.

William Karush. Minima of functions of several variables with inequalities as side conditions. Master's thesis, Department of Mathematics, University of Chicago, Chicago, IL, USA, 1939.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2019.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2014.

H. W. Kuhn and A. W. Tucker. Nonlinear programming. In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, pages 481–492, Berkeley and Los Angeles, 1951. University of California Press.

Tejas D. Kulkarni, Will Whitney, Pushmeet Kohli, and Joshua B. Tenenbaum. Deep convolutional inverse graphics network, 2015.

Ruhui Liu, Weiping Huang, Zhengshun Fei, Kai Wang, and Jun Liang. Constraint-based clustering by fast search and find of density peaks. *Neurocomputing*, 330:223–237, 2019. ISSN 09252312. doi: 10.1016/j.neucom.2018.06.058.

Ester Martin, Kriegel Hans-Peter, Sander Jorg, and Xiawei Xu. A density-based algorithm for discovering clusters in large spartial database with noise. *International Conference on Knowledge Discovery and Data Mining*, 2(6):226–231, 1996.

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Yazhou Ren, Ni Wang, Mingxia Li, and Zenglin Xu. Deep density-based image clustering. *Knowledge-Based Systems*, 197:105841, 2020. ISSN 09507051. doi: 10.1016/j.knosys.2020.105841.

Karl Ridgeway and Michael C. Mozer. Learning deep disentangled embeddings with the f-statistic loss, 2018.

Raphael Schumann. Unsupervised abstractive sentence summarization using length controlled variational autoencoder, 2018.

Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder, 2020.

Dinghan Shen, Asli Celikyilmaz, Yizhe Zhang, Liqun Chen, Xin Wang, Jianfeng Gao, and Lawrence Carin. Towards generating long and coherent text with multi-level latent variable models, 2019.

Bing Song, Hongbo Shi, Yuxin Ma, and Jianping Wang. Multisubspace principal component analysis with local outlier factor for multimode process monitoring. *Industrial & Engineering Chemistry Research*, 53(42):16453–16464, 2014. ISSN 0888-5885. doi: 10.1021/ie502344q.

Ingo Steinwart and Andreas Christmann. *Support Vector Machines*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387772413.

Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders, 2018.

A Proof

A.1 Maximizing Negative KL-divergence

Proof that maximizing the negative KL-divergence is equivalent to maximizing the ELBO.

$$\mathcal{L}(\theta, \phi) = - \mathop{\mathbb{E}}_{p_{\mathcal{D}}} \left[D_{KL}(q_\phi(z|x) || p_\theta(z|x)) \right] \tag{24}$$

$$= - \mathop{\mathbb{E}}_{p_{\mathcal{D}}} \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left[\ln \frac{q_\phi(z|x)}{p_\theta(z|x)} \right] \tag{25}$$

$$= - \mathop{\mathbb{E}}_{p_{\mathcal{D}}} \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left[\ln \frac{q_\phi(z|x)}{p_\theta(z)} \cdot \frac{1}{p_\theta(x|z)} \cdot p_\theta(x) \right] \tag{26}$$

$$\ln p_\theta(x) \geq \mathop{\mathbb{E}}_{p_{\mathcal{D}}} \left[\ln p_\theta(x|z) - D_{KL}(q_\phi(z|x) || p_\theta(z)) \right] \tag{27}$$

A.2 Minimizing $I_q(x, z)$

We demonstrate that minimizing the mutual information between x and latent z equivalently maximizes components of the ELBO and consequently, $I_p(x', z)$ subject to inference constraints.

$$\min_{\theta, \phi \in \mathbb{R}} \quad I_q(x, z) \tag{28}$$

$$I_q(x, z) = D_{KL}(q_\phi(z|x) || q_\phi(z)) \tag{29}$$

$$= \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left[\ln \frac{q_\phi(z|x)}{q_\phi(z)} \cdot \frac{p_\theta(x, z)}{p_\theta(x, z)} \right] \tag{30}$$

$$= \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left[\ln \frac{q_\phi(z|x)}{p_\theta(x|z) p_\theta(z)} \cdot \frac{p_\theta(z|x) p_\theta(x)}{q_\phi(z)} \right] \tag{31}$$

$$= \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} \left[\ln \frac{q_\phi(z|x)}{p_\theta(z)} \cdot \frac{p_\theta(z|x)}{q_\phi(z)} \cdot \frac{p_\theta(x)}{p_\theta(x|z)} \right] \tag{32}$$

$$\begin{aligned} = D_{KL}(q_\phi(z|x) || p_\theta(z)) &- D_{KL}(q_\phi(z) || p_\theta(z)) \\ &- \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} [\ln p_\theta(x|z)] + \ln p_\theta(x) \end{aligned} \tag{33}$$

Note that $p_\theta(z|x) \sim p_\theta(z)$ so that,

$$\begin{aligned} \ln p_\theta(x) \geq \mathop{\mathbb{E}}_{z \sim q_\phi(z|x)} [\ln p_\theta(x|z)] &- D_{KL}(q_\phi(z|x) || p_\theta(z)) \\ &+ D_{KL}(q_\phi(z) || p_\theta(z)) \end{aligned} \tag{34}$$

Therefore, minimizing $I_q(x, z)$ is equivalent to maximizing terms on the Right Hand Side (R.H.S) of 34, which is a lower bound.

A.3 GCVAE: Proof of proposition

We recall the constraint optimization loss presented in section (3) and prove it accordingly. Given the maximization problem,

$$\begin{aligned} \max_{\theta, \phi, \xi^+, \xi^-, \xi_p \in \mathbb{R}} \quad & I_p(x', z) \\ s.t \quad & \mathop{\mathbb{E}}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z|x) \, || \, p_\theta(z)) + I_p(x', z) \leq \xi^- \\ s.t \quad & - \mathop{\mathbb{E}}_{p_{\mathcal{D}}} D_{KL}(q_\phi(z) || p_\theta(z)) \leq \xi^+ \\ s.t \quad & I_p(x', z) \leq \xi_p \\ s.t \quad & \xi_i^+, \xi_i^-, \xi_{ip} \geq 0, \quad \forall i = 1, \ldots, n \end{aligned} \tag{35}$$

The expand of the above equations using sets of Lagrangian multipliers is as follows,

$$\mathcal{L}(x, z; \theta, \phi, \xi^+, \xi^-, \xi, \alpha, \beta, \gamma, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\nu})$$

$$= I_p(x', z) - \beta \left(\underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z|x)||p_\theta(z)) + I_p(x', z) - \sum_{i=1}^n \xi_i^- \right)$$

$$+ \gamma \left(\underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z)||p_\theta(z)) + \sum_{i=1}^n \xi_i^+ \right) - \alpha \left(I(x'; z) - \sum_{i=1}^n \xi_{ip} \right)$$

$$+ \sum_{i=1}^n \boldsymbol{\eta}_i \xi_i^+ + \sum_{i=1}^n \boldsymbol{\tau}_i \xi_i^- + \sum_{i=1}^n \boldsymbol{\nu}_i \xi_{ip} \tag{36}$$

$$\mathcal{L}(x, z; \theta, \phi, \xi^+, \xi^-, \xi, \alpha, \beta, \gamma, \boldsymbol{\eta}, \boldsymbol{\tau}, \boldsymbol{\nu})$$

$$= (1 - \alpha - \beta) I_p(x', z) - \beta \underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z|x)||p_\theta(z))$$

$$+ \gamma D_{KL}(q_\phi(z)||p_\theta(z)) + (\beta - \boldsymbol{\tau}) \sum_{i=1}^n \xi_i^- + (\gamma - \boldsymbol{\eta}) \sum_{i=1}^n \xi_i^+$$

$$+ (\alpha - \boldsymbol{\nu}) \sum_{i=1}^n \xi_{ip} \tag{37}$$

We take the gradient over the loss, $\nabla \mathcal{L}$ for ξ^-, ξ^+, ξ_p and apply KKT optimality conditions to obtain,

$$\mathcal{L}(x, z; \theta, \phi, \boldsymbol{\xi}^+, \boldsymbol{\xi}^-, \boldsymbol{\xi_p}, \alpha, \beta, \gamma)$$

$$= (1 - \alpha - \beta) I_p(x', z) - \beta \underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z|x)||p_\theta(z))$$

$$+ \gamma D_{KL}(q_\phi(z)||p_\theta(z)) \tag{38}$$

$$= (1 - \alpha - \beta) \underset{z \sim q_\phi(z|x)}{\mathbb{E}} [\ln p_\theta(x|z)]$$

$$- \beta \underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z|x)||p_\theta(z))$$

$$+ \gamma D_{KL}(q_\phi(z)||p_\theta(z)) \tag{39}$$

We set the Lagrangian adaptive hyperparameters following [Shao et al., 2020] as follows,

$$\mathcal{L}(x, z; \theta, \phi, \alpha, \beta, \gamma)$$

$$= (1 - \alpha_t - \beta_t) \underset{z \sim q_\phi(z|x)}{\mathbb{E}} [\ln p_\theta(x|z)]$$

$$- \beta_t \underset{p_\mathcal{D}}{\mathbb{E}} D_{KL}(q_\phi(z|x)||p_\theta(z))$$

$$+ \gamma_t D_{KL}(q_\phi(z)||p_\theta(z)) \tag{40}$$

The adaptive weight α_t controls the reconstruction error while β_t ensures the posterior latent factor $q_\phi(z|x)$ does not deviate significantly from its prior $p_\theta(z)$. Varying both terms gives us better control of the degree of disentanglement and helps us to understand the parameters affecting density disentanglement.

A.4 Expected Squared Mahalanobis distance: $\mathbb{E} D^2_{MAH}(q||p)$

Suppose that a data $x \in \mathcal{X}$ with probability $p(x)$ is projected into a reproducing kernel Hilbert space $< \varphi(x) >_\mathcal{H}$, the expectation of the squared Mahalanobis distance is as follows,
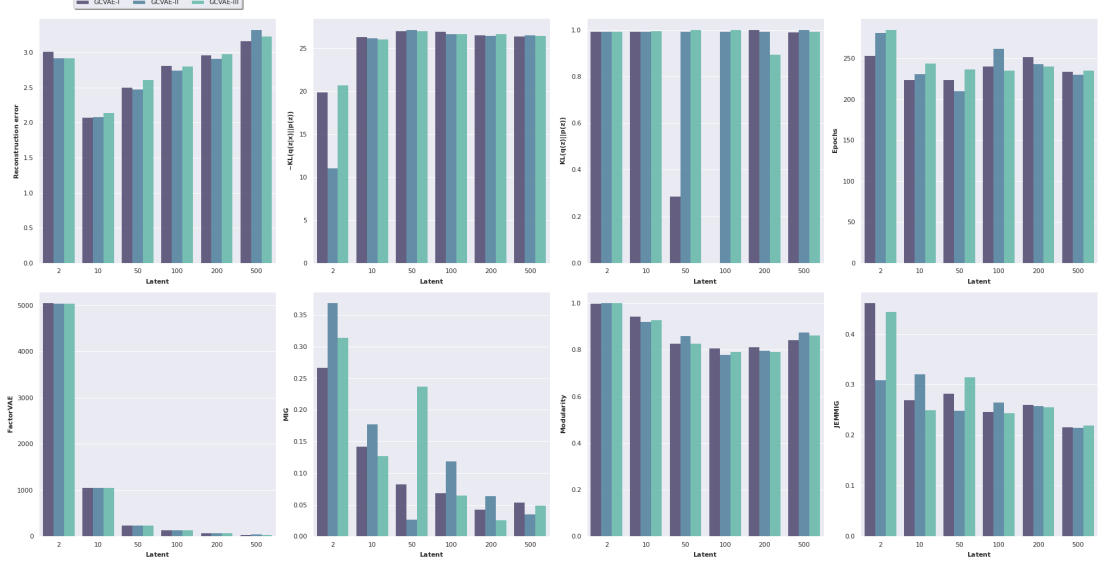
Figure 5: Comparison of metrics for DSprites 2D shapes dataset. **Top**: Comparison of GCVAE-I, II & III losses over increasing latent dimensions. The lowest reconstruction error is observed for GCVAE-I on Latent-10 and is monotone increasing thereafter. $-D_{KL}(q_\phi(z|x))||p_\theta(z))$ is stable, increasing above $z = 10$ dimensions and highest for GCVAE-II and III on Latent-50. Note that the original values of $D_{KL}(q_\phi(z))||p_\theta(z))$ are scaled to $[0, 1]$.
Bottom: Disentanglement metric over different dimensions. FactorVAE metric is monotone decreasing for latent space greater than 2. MIG is similar in behavior to FactorVAE metric and best for GCVAE-II on Latent-2.

$$\mathbb{E}D^2_{MAH}(q(z)||p(z))$$

$$\geq \| \mathop{\mathbb{E}}_{x\sim p} \varphi(x) - \mathop{\mathbb{E}}_{y\sim q} \varphi(y) \|^2_{\Sigma^{-1},\mathcal{H}} \tag{41}$$

$$\geq (\mathop{\mathbb{E}}_{x\sim p} \varphi(x) - \mathop{\mathbb{E}}_{y\sim q} \varphi(y))' \Sigma^{-1} (\mathop{\mathbb{E}}_{x\sim p} \varphi(x) - \mathop{\mathbb{E}}_{y\sim q} \varphi(y)) \tag{42}$$

$$\geq \Sigma^{-1} (\mathop{\mathbb{E}}_{x\sim p} \varphi(x) - \mathop{\mathbb{E}}_{y\sim q} \varphi(y))' (\mathop{\mathbb{E}}_{x\sim p} \varphi(x) - \mathop{\mathbb{E}}_{y\sim q} \varphi(y)) \tag{43}$$

$$\geq \Sigma^{-1} [\mathop{\mathbb{E}}_{x'\sim p, x\sim p} \langle \varphi(x'), \varphi(x)\rangle_\mathcal{H} - \mathop{2\mathbb{E}}_{x'\sim p, y\sim q} \langle \varphi(x'), \varphi(y)\rangle_\mathcal{H}$$

$$+ \mathop{\mathbb{E}}_{y'\sim q, y\sim q} \langle \varphi(y'), \varphi(y)\rangle_\mathcal{H}] \tag{44}$$

We suppose that the feature map $\varphi(x)$ takes the canonical form $k(x, \cdot)$, so that $\langle \varphi(x'), \varphi(x)\rangle_\mathcal{H} = k(x', x)$ [Steinwart and Christmann, 2008], where $k(x', x)$ represents a positive semi-definite kernel (for instance the Gaussian kernel, $exp(-\frac{\|x'-x\|^2}{2\sigma^2})$). Hence,

$$\mathbb{E}D^2_{MAH}(q(z)||p(z))$$

$$\geq \Sigma^{-1} [\mathop{\mathbb{E}}_{x'\sim p, x\sim p} k(x', x) - 2 \mathop{\mathbb{E}}_{x'\sim p, y\sim q} k(x', y)$$

$$+ \mathop{\mathbb{E}}_{y'\sim q, y\sim q} k(y', y)]$$

$$\geq \Sigma^{-1} [\mathop{\mathbb{E}}_{z'\sim p, z\sim p} k(z', z) - 2 \mathop{\mathbb{E}}_{z'\sim p, z\sim q} k(z', z)$$

$$+ \mathop{\mathbb{E}}_{z'\sim q, z\sim q} k(z', z)]$$

$$= \Sigma^{-1} D^2_{MMD}(q(z)||p(z))$$

$$\tag{45}$$

14

B Performance comparison of GCVAE

We compare the performance of GCVAE-I, II & III using a stopping criterion. We design a stopping criterion using the adaptive convergence of α and β below a certain threshold ϵ. ϵ_a for stopping reconstruction loss and ϵ_b stops KL divergence, $-D_{KL}(q_\phi(z|x)||p_\theta(z))$ at the optimum point.

Algorithm 1: GCVAE stopping criteria

Input : $\mathcal{L}(\phi, \theta, \dots), \alpha_t, \beta_t, \gamma_t, \epsilon_a, \epsilon_b$
Output : Optimal parameters of loss $\mathcal{L}$

1 **begin**
2   **Compute gradient:** $\nabla\mathcal{L}(\phi, \theta, \alpha_t, \beta_t, \gamma_t, \dots)$;
3   **Perform back propagation**;
4   **if** α_t - $\alpha_{t-1} < \epsilon_a$ && β_t - $\beta_{t-1} < \epsilon_b$ **then**
5 | return optimal $\mathcal{L}, \alpha_t, \beta_t, \gamma_t$;
6   **end if**
7 **end**

The stopping criteria reduces the constraint of stopping the algorithm at a specific learning epoch, serving to assist the model to conclude the final epoch, when the optimal reconstruction and KL divergences are obtained. The optimal parameters yield the best disentanglement with low reconstruction error. We observe that reconstruction error is not usually as low as when run for N-iterations, but the representations results are reasonably comparable.

Figure 5 illustrates the result of using the stopping criterion on the DSprites 2D data set. We observe in the very top right-hand corner of this figure that the number of iterations required to reach optimum point is below 300 for all latent spaces. The singular downside to using stopping criteria is with regards to reconstruction error. Minimum reconstruction error and maximum disentanglement are observed for GCVAE-III on Latent-10 and GCVAE-II on Latent-2 respectively.

We present an elaborate representation of the results of GCVAE for N-number of iterations in Figure 6. The running time is better represented in Figure 5 while 6 details reasons for why a GCVAE model obtains a lower MIG disentanglement metric over the other. This is evident in $D_{KL}(q_\phi(z)||p_\theta(z))$. We observe a collapse of the $D_{KL}(q_\phi(z)||p_\theta(z))$ when $D_{KL}(q_\phi(z)||p_\theta(z)) = 0$, which reduces the GCVAE model to ControlVAE. This is demonstrable, since α_t and β_t are usually very low. This effect corresponds to the lowest MIG disentanglement metric observed in Latent-50 and 200 for GCVAE-II. A GCVAE model whose posterior $q_\phi(z) \approx p_\theta(z)$ returns better disentanglement, since the second $D_{KL}(q_\phi(z|x)||p_\theta(z)) > 0$.

Secondly, we identify another reason for the collapse of the correlation term, $D_{KL}(q_\phi(z)||p_\theta(z))$. Given an inverse covariance matrix (Σ^{-1}) whose diagonals are infinitesimally small, this could in fact collapse the correlation and subsequently return low disentanglement score. This is indeed evident, since the generated factors of the latent z have a different posterior, $p_\theta(x|z)$ from the original factors of the given distribution.

Despite the emphasis on disentanglement as opposed to reconstruction, we observe that GCVAE with a stopping criteria running for very low number of iterations is capable of achieving reasonable reconstruction of the original data as well as generating meaningful data from an unseen sample.

B.1 Architecture

We use similar architecture to train GCVAE-I, II, III and all benchmark models. The proposed convolutional 2D convolutional architecture is given in the table below,
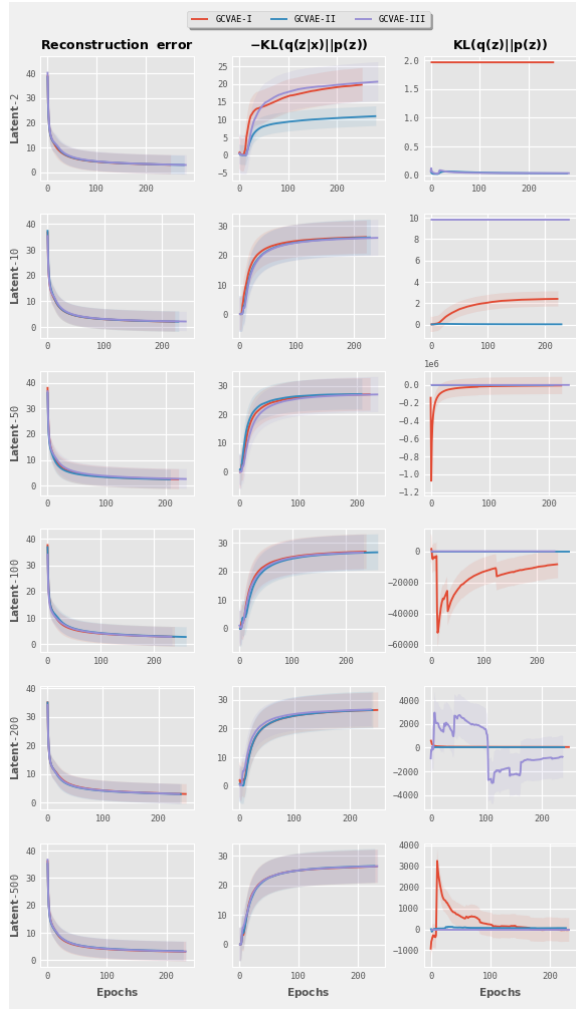
Figure 6: A visual comparison of the reconstruction, $-D_{KL}(q_\phi(z|x))||p_\theta(z))$ and $D_{KL}(q_\phi(z))||p_\theta(z))$ losses for GCVAE-I, II & III over different latent space. Behavior of reconstruction error per latent is relatively close and indistinguishable. In all cases of latents experimented with except for Latent-2, $-D_{KL}(q_\phi(z|x))||p_\theta(z))$ is comparable. GCVAE-I is unstable in $D_{KL}(q_\phi(z))||p_\theta(z))$ during training across all latents. While a lower value of $D_{KL}(q_\phi(z))||p_\theta(z))$ is preferred, we observe correlation with MIG disentanglement metric in Figure 5.
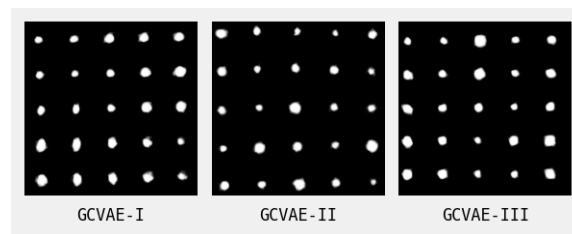

Figure 7: Generated samples from training with 737 observations and stopping after less than 300 iterations. The generated samples especially for GCVAE-II.

16

Encoder	Decoder
Input 64 x 64 binary image	Input $\mathbb{R}^{10}$
4×4 conv. 64 ReLU. stride 2	FC 25; FC 200 ReLU
4×4 conv. 64 ReLU. stride 2	4×4 upconv. 32 ReLU. stride 2
4×4 conv. 32 ReLU. stride 2	4×4 upconv. 32 ReLU. stride 2
4×4 conv. 32 ReLU. stride 2	4×4 upconv. 64 ReLU. stride 2
FC 200; FC 25; FC 2 X 10	4×4 upconv. 64 ReLU. stride 2
	4×4 upconv. 1. stride 2

Table 2: Encoder-Decoder architecture for 2D DSprites and MNIST data. We use a 2D MaxPooling and Batch normalization for the encoder while size 2 Upsampling with Batch normalization is used for the decoder.