Bayes-Factor-VAE: Hierarchical Bayesian Deep Auto-Encoder Models for Factor Disentanglement

Minyoung Kim[1], Yuting Wang[1,2], Pritish Sahu[2], and Vladimir Pavlovic[1,2]

[1]Samsung AI Center, Cambridge, UK
[2]Dept. of Computer Science, Rutgers University, NJ, USA
mikim21@gmail.com, {yw632,ps851,vladimir}@cs.rutgers.edu, v.pavlovic@samsung.com

Abstract

We propose a family of novel hierarchical Bayesian deep auto-encoder models capable of identifying disentangled factors of variability in data. While many recent attempts at factor disentanglement have focused on sophisticated learning objectives within the VAE framework, their choice of a standard normal as the latent factor prior is both suboptimal and detrimental to performance. Our key observation is that the disentangled latent variables responsible for major sources of variability, the relevant factors*, can be more appropriately modeled using long-tail distributions. The typical Gaussian priors are, on the other hand, better suited for modeling of nuisance factors. Motivated by this, we extend the VAE to a hierarchical Bayesian model by introducing hyper-priors on the variances of Gaussian latent priors, mimicking an infinite mixture, while maintaining tractable learning and inference of the traditional VAEs. This analysis signifies the importance of partitioning and treating in a different manner the latent dimensions corresponding to relevant factors and nuisances. Our proposed models, dubbed Bayes-Factor-VAEs, are shown to outperform existing methods both quantitatively and qualitatively in terms of latent disentanglement across several challenging benchmark tasks.*

1. Introduction

Data, such as images or videos, are inherently high-dimensional, a result of interactions of many complex factors such as lighting, illumination, geometry, etc. Identifying those factors and their intricate interplay is the key not only to explaining the source of variability in the data but also to efficiently representing the same data for subsequent analysis, classification, or even re-synthesis. To tackle this problem, deep factor models such as the VAE [17] have been proposed to principally, mathematically concisely, and computationally efficiently model the nonlinear generative relationship between the ambient data and the latent factors.

However, solely identifying some factors beyond the sources of variability is not sufficient; it is ultimately desirable that the identified factors also be *disentangled*. Although there are several different, sometimes opposing, views of disentanglement [3, 12], the most commonly accepted definition aligns with the notion of *apriori independence*, where each aspect of independent variability in data is exclusively sourced in one latent factor. Identifying these disentangled factors will then naturally lead to an effective, succinct representation of the data. In this paper we aim to solve this disentangled representation learning task in the most challenging, *unsupervised* setting, with no auxiliary information, such as labels, provided during the learning process.

While there have been considerable recent efforts to solve the latent disentanglement problem [7, 21, 13, 5, 18, 15, 6], most prior approaches have failed to produce satisfactory solutions. One fundamental reason for this is their inadequate treatment of the key factors supporting the disentanglement, which have in prior works been almost universally tied to i.i.d. Gaussian priors. In contrast, to accomplish high-quality disentanglement *one needs to distinguish, and treat separately, the relevant latent variables, responsible for principal variability in the data, from the nuisance sources of minor variation. Specifically, the relevant factors may exhibit non-Gaussian, long-tail behavior, which discerns them from statistically independent Gaussian nuisances.* We will detail and justify this requirement in Sec. 2.

Our goal in this paper is to develop principled factor disentanglement algorithms that meet this requirement. In particular, we propose three different hierarchical Bayesian models that place hyper-priors on the parameters of the latent prior. This effectively mimics employing infinite mixtures while maintaining tractable learning and inference of traditional VAEs. We begin with a brief background on VAEs, describe our motivation and requirement to achieve the disentanglement in a principled way (Sec. 2), followed by the definition of specific models (Sec. 3).

Background. We denote by $\mathbf{x} \in \mathbb{R}^D$ the observation (e.g., image) and by $\mathbf{z} \in \mathbb{R}^d$ the underlying latent vector. The variational auto-encoder (VAE) [17] is a deep probabilistic model that represents the joint distribution as:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I}), \quad (1)$$
$$p_\theta(\mathbf{x}|\mathbf{z}) = p(\mathbf{x}; \theta(\mathbf{z})), \quad (2)$$

where $p(\mathbf{x}; \theta(\mathbf{z}))$ is a density model with the parameters $\theta(\mathbf{z})$ whose likelihood can be tractably computed (e.g., Gaussian or Bernoulli), and $\theta(\mathbf{z})$ is the output of a deep model with its own weight parameters. In the unsupervised setting, with ambient data $\{\mathbf{x}^n\}_{n=1}^N$, the model can be learned by the MLE, i.e., maximizing $\sum_{n=1}^N \log p(\mathbf{x}^n)$. This requires posterior inference $p(\mathbf{z}|\mathbf{x})$, but as the exact inference is intractable, the VAE adopts the variational technique: approximate $p(\mathbf{z}|\mathbf{x}) \approx q_\nu(\mathbf{z}|\mathbf{x})$, where $q_\nu(\mathbf{z}|\mathbf{x}) = q(\mathbf{z}; \nu(\mathbf{x}))$ is a freely chosen tractable density with parameters modeled by deep model $\nu(\mathbf{x})$. A typical choice, assumed throughout the paper, is independent Gaussian,

$$q_\nu(\mathbf{z}|\mathbf{x}) = \prod_{j=1}^d \mathcal{N}(z_j; m_j(\mathbf{x}), s_j(\mathbf{x})^2), \quad (3)$$

where $\nu(\mathbf{x}) = \{m_j(\mathbf{x}), s_j(\mathbf{x})\}_{j=1}^d$ for some deep networks $m_j(\mathbf{x})$ and $s_j(\mathbf{x})$. The negative data log-likelihood admits the following as its upper bound,

$$\text{Rec}(\theta, \nu) + \mathbb{E}_{p_d(\mathbf{x})}\left[\text{KL}(q_\nu(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right], \quad (4)$$

which we minimize wrt $\theta(\cdot)$ and $\nu(\cdot)$. Here, $p_d(\mathbf{x})$ is the empirical data distribution of $\{\mathbf{x}^n\}_{n=1}^N$, and

$$\text{Rec}(\theta, \nu) = -\mathbb{E}_{p_d(\mathbf{x})}\left[E_{q_\nu(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\right] \quad (5)$$

is the negative expected log-likelihood, identical to the reconstruction loss.

2. Our Motivation

Although minimizing (4) can yield a model that faithfully explains the observations, the learned model does not necessarily exhibit *disentanglement of latent factors*. In this section, we begin with a common notion of latent disentanglement[1], and consider a semi-parametric extension of VAE to derive a principled objective function to achieve latent disentanglement under this notion. Our analysis also suggests to discriminate relevant latent variables from nuisances, and separately treat the two.

Notion of Disentanglement. Consider a set of aspects that can be observed in $\mathbf{x}$, where the value of each aspect varies independently from the others in the data. In the facial image data, for instance, we typically observe images where the variability of each aspect, say (pose, gender, facial expression), is independent from the others (e.g., the distribution of pose variability in images is the same regardless of gender or expression). We then say the latent vector $\mathbf{z}$ is disentangled if each variable z_j is statistically correlated with only a single aspect, exclusive from other z_{-j}. That is, varying z_j while fixing z_{-j}, results in the exclusive variation of the j-th aspect in $\mathbf{x}$.

Relevant vs. Nuisance Variables. It is natural to assume the exact number of meaningful aspects is a priori unknown, but a sufficiently large upper bound d may be known. Only some variables in $\mathbf{z}$ will have correspondence to aspects, with the rest attributed to nuisance effects (e.g., acting as a conduit to the data generation process). We thus partition the latent dimensions into two disjoint subsets, $\mathbf{R}$ (relevant) and $\mathbf{N}$ (nuisance), $\mathbf{R} \cup \mathbf{N} = \{1, \ldots, d\}$ and $\mathbf{R} \cap \mathbf{N} = \emptyset$. Formally, index j is said to be relevant ($j \in \mathbf{R}$) if z_j and $\mathbf{x}$ are statistically *dependent* and j is called nuisance ($j \in \mathbf{N}$) if z_j and $\mathbf{x}$ are statistically *independent*. Analysis in this section assumes *known* $\mathbf{R}$ and $\mathbf{N}$.

The above notion implies the latent variables z_j's be apriori independent of each other, in agreement with the goals and framework of the *independent component analysis* (ICA) [14], the task of blind separation of statistically independent sources. In particular, our derivation is based on the *semi-parametric* view [4, 2], in which the only assumption made is that of a fully factorized $p(\mathbf{z})$, with no restrictions on the choice of the density $p(\mathbf{z})$.

For ease of exposition, we consider a *deterministic* decoder/encoder pair, $\mathbf{x} = \text{dec}_\theta(\mathbf{z})$ and $\mathbf{z} = \text{enc}_\nu(\mathbf{x})$ with parameters θ and ν, constrained to be the inverses of each other, $\text{enc}_\nu(\cdot) = \text{dec}_\theta^{-1}(\cdot)$. In the semi-parametric ICA, we seek to solve the MLE problem:

$$\min_{p(\mathbf{z}), \theta} \text{KL}\left(p_d(\mathbf{x})||p_\theta(\mathbf{x})\right) \quad \text{s.t.} \;\; p(\mathbf{z}) = \prod_{j=1}^d p(z_j), \quad (6)$$

where $p_\theta(\mathbf{x})$ is the density derived from $\mathbf{x} = \text{dec}_\theta(\mathbf{z})$, with $\mathbf{z} \sim p(\mathbf{z})$. The latent prior $p(\mathbf{z})$ is now a part of our model to be learned, instead of being fixed as in VAE. We let $p(\mathbf{z})$ be of free form (semi-parametric) but fully factorized, the key to the latent disentanglement.

Directly optimizing (6) is intractable, and we solve it in the $\mathbf{z}$ space. Using the fact that KL divergence is invariant to invertible transformations[2], we have:

$$\text{KL}(p_d(\mathbf{x})||p_\theta(\mathbf{x})) = \text{KL}(q_\nu(\mathbf{z})||p(\mathbf{z})), \quad (7)$$

where $q_\nu(\mathbf{z})$ is the density of $\mathbf{z} = \text{enc}_\nu(\mathbf{x})$ with $\mathbf{x} \sim p_d(\mathbf{x})$. Our original problem (6) then becomes:

$$\min_{p(\mathbf{z}), \nu} \text{KL}_{\mathbf{z}} := \text{KL}\left(q_\nu(\mathbf{z}) \Bigg\| \prod_{j=1}^d p(z_j)\right) \quad (8)$$

[1] While there is no universal definition, the one we use shares the main concepts with other definitions, including the recent symmetric transformation view [12].

[2] See Supplement for the proof.

In case when the encoder/decoder pair becomes *stochastic* (2) and (3), three modifications are needed: i) stochastic inverse[3], ii) the invariance of KL (7) turns into an approximation, and iii) $q_\nu(\mathbf{z})$ is defined as:

$$q_\nu(\mathbf{z}) = \mathbb{E}_{p_d(\mathbf{x})}\Big[q_\nu(\mathbf{z}|\mathbf{x})\Big] = \frac{1}{N}\sum_{n=1}^{N} q_\nu(\mathbf{z}|\mathbf{x}^n), \quad (9)$$

a well known quantity in the recent disentanglement literature, dubbed *aggregate posterior*.

Further imposing the independence constraint for the nuisance variables, our optimization problem becomes:

$$\min_{p(\mathbf{z}),\nu} \mathrm{KL}_\mathbf{z} \;\; \text{s.t.} \;\; q_\nu(z_j|\mathbf{x}) = q_\nu(z_j) \;\; \forall \mathbf{x}, \;\; j \in \mathbf{N}, \quad (10)$$

where $q_\nu(z_j|\mathbf{x})$ and $q_\nu(z_j)$ are marginals from $q_\nu(\mathbf{z}|\mathbf{x})$ and $q_\nu(\mathbf{z})$, respectively. We will often omit the subscript ν in notation. It is not difficult to see that the objective $\mathrm{KL}_\mathbf{z}$ in (8) and (10) can be decomposed as follows (see Supplement):

$$\mathrm{KL}_\mathbf{z} = \mathrm{TC} + \sum_{j\in\mathbf{R}} \mathrm{KL}(q(z_j)||p(z_j)) + \sum_{j\in\mathbf{N}} \mathrm{KL}(q(z_j)||p(z_j)) \quad (11)$$

where TC is the *total correlation*, a measure of the degree of factorization of $q(\mathbf{z})$:

$$\mathrm{TC} := \mathrm{KL}\Bigg(q(\mathbf{z})\Bigg\|\prod_{j=1}^{d} q(z_j)\Bigg). \quad (12)$$

With the freedom to choose $p(\mathbf{z})$ and ν (of $q_\nu(\mathbf{z}|\mathbf{x})$) to minimize $\mathrm{KL}_\mathbf{z}$ within the constraint (10), we tackle the last two terms in (11) individually.

3rd Term. For nuisance z_j, to satisfy the constraint (10), we have $q(z_j|\mathbf{x}) = \mathcal{N}(z_j; m_j, s_j^2)$ for some fixed m_j and s_j. Then $q(z_j) := \int q(z_j|\mathbf{x})p_d(\mathbf{x})d\mathbf{x} = \mathcal{N}(z_j; m_j, s_j^2)$, allowing one to choose a Gaussian prior $p(z_j) = \mathcal{N}(z_j; 0, 1)$, leading to $m_j = 0$, $s_j = 1$, vanishing the KL.

2nd Term. For z_j a relevant factor variable, z_j and $\mathbf{x}$ should not be independent, thus $q(z_j)$ is a Gaussian mixture with heterogeneous components $q(z_j|\mathbf{x})$. The VAE's Gaussian prior $p(z_j) = \mathcal{N}(z_j; 0, 1)$ implies that the divergence can never be made to vanish in general. To remedy this, one either i) chooses $p(z_j)$ different from $\mathcal{N}(0, 1)$ (potentially, non-Gaussian), or ii) retains a Gaussian prior but lets the mean and variance of $p(z_j)$ be flexibly chosen, perhaps differently over $j \in \mathbf{R}$, to maximally diminish this KL divergence. The former approach may raise a nontrivial question of which prior to choose[4]. Instead, we propose a solution that builds a hierarchical Bayesian prior of $p(z_j)$ and infers the posterior (Sec. 3.2 and 3.3). In this strategy, we regard the variances of Gaussian $p(z_j)$ as parameters to be learned, and minimize $\mathrm{KL}(q(z_j)||p(z_j))$ wrt the VAE parameters as well as the prior variances (Sec. 3.1).

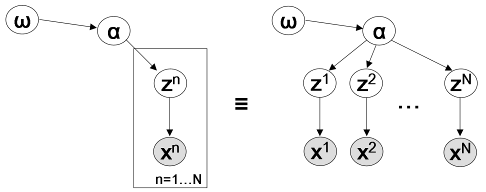

Figure 1. Graphical model representation for BF-VAE-1 and BF-VAE-2: (Left) plate, (Right) unrolled version. The hyperparameter $\boldsymbol{\omega}$ is either $\{a_j\}$ (BF-VAE-1) or $\{r_j\}$ (BF-VAE-2).

Learning Objective. Based on the above analysis, the overall learning goal can be defined as:

$$\begin{aligned} &\min_{\theta,\nu,p(\mathbf{z})} \;\; \mathrm{Rec}(\theta,\nu) + \mathbb{E}_{p_d(\mathbf{x})}[\mathrm{KL}(q(z_j|\mathbf{x})||p(z_j))] + \gamma\mathrm{TC} \\ &\quad \text{s.t.} \quad p(z_j) = \mathcal{N}(z_j; 0, 1) \;\; \text{for} \;\; j \in \mathbf{N}, \end{aligned} \quad (13)$$

where we include $\mathrm{Rec}(\theta,\nu)$ of (5) to impose the stochastic inverse, and replace the difficult-to-evaluate $\mathrm{KL}(q(z_j)||p(z_j))$ by the expected KL, an upper bound[5] admitting a closed form. The TC term will be estimated through its density ratio proxy, using an adversarial discriminator similarly as [15], where its impact is controlled by γ.

Our learning objective in (13) is similar to those of recent disentanglement algorithms (see Sec. 4) in that the VAE loss is augmented with the additional loss of independence of latent variables, such as the TC term. However, a key distinction is our separate treatment of relevant and nuisance variables, with the additional aim to learn a non-Gaussian relevant variable prior $p(z_j)$. The optimization (13) assumes a known relevance partition $\mathbf{R}$ and $\mathbf{N}$. In the next section we will deal with how to learn this partition automatically from data, either implicitly (Sec. 3.1 and 3.2) or explicitly (Sec. 3.3) via hierarchical Bayesian treatment.

3. Bayes-Factor-VAE (BF-VAE)

The key insight from Sec. 2 is that, for relevant factors, it is necessary to have $p(z_j)$ different from $\mathcal{N}(0, 1)$. In this section we propose three different prior models to accomplish this goal in a principled Bayesian manners.

3.1. Adjustable Gaussian Prior (BF-VAE-0)

We first define a base model, also needed for subsequent more complex variations, which relaxes the fixed, identical variance assumption for priors $p(z_j)$:

$$p(\mathbf{z}|\boldsymbol{\alpha}) = \prod_{j=1}^{d} p(z_j|\alpha_j) \; = \; \prod_{j=1}^{d} \mathcal{N}(z_j; 0, \alpha_j^{-1}), \quad (14)$$

[3] Such that θ and ν minimize the reconstruction loss $\mathrm{Rec}(\theta,\nu)$ (5).

[4] One may employ a flexible model for $p(z_j)$, e.g., a finite mixture approximation or the VampPrior [26]. However, this may lead to overfitting; see our empirical study in Sec. 5.1.

[5] See Supplement for the proof.

where $\boldsymbol{\alpha} > 0$ are the precision parameters to be learned from data[6].

We expect the learned α_j to be close to (apart from) 1 for nuisance (relevant, resp.) j. To explicitly express our preference of encouraging many dims j to be nuisance, and avoid redundancy in the learned relevant variables, we add the regularizer, $(\alpha_j^{-1} - 1)^2$, which leads to:

$$\min_{\theta,\nu,\boldsymbol{\alpha}} \sum_{j=1}^{d} \mathbb{E}_{p_d(\mathbf{x})}\Big[\mathrm{KL}(q(z_j|\mathbf{x})||\mathcal{N}(z_j;0,\alpha_j^{-1}))\Big] + \mathrm{Rec}(\theta,\nu) + \gamma\mathrm{TC} + \eta\sum_{j=1}^{d}(\alpha_j^{-1}-1)^2. \quad (15)$$

We denote this model by **BF-VAE-0**. The expected KL in (15) admits a closed form, resulting in added flexibility without extra computation, compared to e.g., [15]. Another benefit is the trade-off parameter η acts as a proxy to control the cardinality of relevant factors; small η encourages more relevant factors than large η.

3.2. Hierarchical Bayesian Prior (BF-VAE-1)

To extend BF-VAE-0 to a Bayesian hierarchical setting, in conjunction with (14), we adopt a conjugate prior on $\boldsymbol{\alpha}$,

$$p(\boldsymbol{\alpha}) = \prod_{j=1}^{d} p(\alpha_j) = \prod_{j=1}^{d} \mathcal{G}(\alpha_j; a_j, b_j), \quad (16)$$

where $\mathcal{G}(y; a, b) \propto y^{a-1}e^{-by}$ is the Gamma distribution with parameters a (shape) and b (inverse scale) with $a, b > 0$. We further set $b_j = a_j - 1$, $a_j > 1$, to express our preference for $\mathrm{Mode}[p(\alpha_j)] = 1$[7]. We let $\{a_j\}_{j=1}^d$ be the model parameters that can be learned from data. This model, named **BF-VAE-1**, has a graphical model representation shown in Fig. 1.

A key aspect of this model is that by marginalizing out $\boldsymbol{\alpha}$, the prior $p(\mathbf{z})$ becomes an infinite Gaussian mixture, $p(\mathbf{z}) = \int p(\boldsymbol{\alpha})\mathcal{N}(\mathbf{z}; \mathbf{0}, \boldsymbol{\alpha}^{-1})d\boldsymbol{\alpha}$, a desideratum for relevant factors. Because $\mathrm{Var}[p(\alpha_j)] \approx (a_j - 1)^{-1}$, large a_j will lead to $\lim_{a_j\to\infty} p(z_j|a_j) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{1})$, a nuisance factor.

We describe the variational inference for the model where we introduce variational densities $q(\boldsymbol{\alpha})$ and $q(\mathbf{z}|\mathbf{x})$ to approximate the true posteriors as follows:

$$p(\boldsymbol{\alpha}, \{\mathbf{z}^n\}_{n=1}^N | \{\mathbf{x}^n\}_{n=1}^N) \approx \overbrace{\prod_{j=1}^{d} \mathcal{G}(\alpha_j; \hat{a}_j, \hat{b}_j)}^{q(\boldsymbol{\alpha})} \prod_{n=1}^{N} q(\mathbf{z}^n|\mathbf{x}^n). \quad (17)$$

This allows the average negative marginal data log-likelihood, $-\frac{1}{N}\log p(\{\mathbf{x}^n\})$, to be upper-bounded by[8]:

$$\mathcal{U}_1 := \mathrm{Rec}(\theta,\nu) + \frac{1}{N}\mathrm{KL}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha})) + \mathbb{E}_{q(\boldsymbol{\alpha})}\mathbb{E}_{p_d(\mathbf{x})}\big[\mathrm{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\boldsymbol{\alpha}))\big]. \quad (18)$$

$\mathrm{Rec}(\theta,\nu)$ in (18) is identical to that of VAE, while the other two admit closed forms; see Supplement for the details. The TC term becomes an average over $q(\boldsymbol{\alpha})$:

$$\mathrm{TC}_1 := \mathbb{E}_{q(\boldsymbol{\alpha})}\left[\mathrm{KL}(q(\mathbf{z}|\boldsymbol{\alpha})||\prod_{j=1}^{d} q(z_j|\boldsymbol{\alpha}))\right], \quad (19)$$

which turns out to be equal to TC in (12), since $q(\mathbf{z}|\boldsymbol{\alpha}) := \int q(\mathbf{z}|\boldsymbol{\alpha},\mathbf{x})p_d(\mathbf{x})d\mathbf{x} = \int q(\mathbf{z}|\mathbf{x})p_d(\mathbf{x})d\mathbf{x} = q(\mathbf{z})$. The final optimization is then minimizing $(\mathcal{U}_1 + \gamma\mathrm{TC}_1)$ wrt (θ,ν) and $\{a_j, b_j, \hat{a}_j, \hat{b}_j\}_{j=1}^d$ with the constraint $b_j = a_j - 1$.

BF-VAE-1 can capture the uncertainty in the precision parameters $\boldsymbol{\alpha}$ with no computational overhead as all of the objective terms admit closed forms. Having learned the model from data $\mathcal{D} = \{\mathbf{x}^n\}_{n=1}^N$, the *data corrected prior*, $\overline{p}(z_j) := \int p(z_j|\alpha_j)p(\alpha_j|\mathcal{D})d\alpha_j$, is approximated as:

$$\overline{p}(z_j) \approx \int p(z_j|\alpha_j)q(\alpha_j)d\alpha_j = t_{2\hat{a}_j}\left(z_j; 0, \frac{\hat{b}_j}{\hat{a}_j}\right), \quad (20)$$

where $t_f(0, v)$ is the generalized Student's t distribution with dof f and shape v. $\overline{p}(z_j)$ informs us about the relevance of z_j: Large dof implies nuisance (as the t becomes close to Gaussian), while small suggests a relevant variable.

3.3. Prior with Relevance Indicators (BF-VAE-2)

BF-VAE-1 allows only implicit control over the cardinality of relevant dims, assuming no explicit differentiation between relevant factors and nuisances. In this section we propose another model that can address these issues.

The key idea[9] is to introduce relevance indicator variables $\mathbf{r} \in [0, 1]^d$ (high r_j indicating relevance of z_j). We let $\mathbf{r}$ determine the shape of the hyper prior $p(\boldsymbol{\alpha})$: If $r_j \approx 1$ (relevant), we make $p(\alpha_j)$ uninformative, thus z_j far from $\mathcal{N}(0, 1)$. In contrast, if $r_j \approx 0$ (nuisance), $p(\alpha_j)$ should strongly peak at $\alpha_j = 1$, with $p(z_j)$ close to $\mathcal{N}(0, 1)$. The following reparametrization of (16) enables this control:

$$p(\boldsymbol{\alpha}|\mathbf{r}) = \prod_{j=1}^{d} \mathcal{G}\left(\alpha_j; \frac{1+2\epsilon}{r_j+\epsilon}, \frac{1+2\epsilon}{r_j+\epsilon} - 1\right), \quad (21)$$

[6] Note that we fix the mean as 0, and only learn the (inverse) variances α_j. Although we can easily parametrize the mean as well, the form of (14) is equally flexible in terms of minimizing the KL, as shown in Supplement.

[7] This preference also improved empirical performance.

[8] See Supplement for the derivations.

[9] It is related to the well-known (Bayesian) variable selection problem [24], but clearly different in that the latter is typically framed within the standard regression setup where the variables (covariates) of interest are *observed* in the data. In our case, we aim to select the most relevant *latent* variables z_j's that explain the major variation in the observed data.

where ϵ is a small positive number (e.g., 0.001).

The indicator $\mathbf{r}$ naturally defines the relevant index set $\mathbf{R} = \{j : r_j \approx 1\})$, allowing us to decompose $q(\mathbf{z})$ over $\mathbf{R}$ and $\mathbf{N}$ as $q(\mathbf{z_R}) \cdot \prod_{j \in \mathbf{N}} q(z_j)$[10], making TC into:

$$\mathrm{KL}\left(q(\mathbf{z_R}) || \prod_{j \in \mathbf{R}} q(z_j) \right) \approx \mathbb{E}_{q(\mathbf{z_R})} \left[\log \frac{D(\mathbf{z_R})}{1 - D(\mathbf{z_R})} \right], \tag{22}$$

focused only on relevant variables. Note that we suggest using the discriminator density ratio proxy, rhs of (22), to evaluate TC, with $D(\cdot)$ optimized to discern samples from $q(\mathbf{z_R})$ from those of $\prod_{j \in \mathbf{R}} q(z_j)$.

To turn (22) into a continuous space optimization problem, we rewrite $D(\mathbf{z_R})$ as $D(\mathbf{r} \circ \mathbf{z})$, where $\circ$ is the element-wise (Hadamard) product, and introduce two additional regularizers to control the cardinality of $\mathbf{R}$ through $||\mathbf{r}||_1$ and the preference toward discrete values using the entropic prior $H(\mathbf{r}) = -\sum_{j=1}^{d} \big(r_j \log r_j + (1 - r_j) \log(1 - r_j) \big)$. This leads to the final objective:

$$\mathcal{U}_1 + \gamma \mathbb{E}_{q(\mathbf{z})} \left[\log \frac{D(\mathbf{r} \circ \mathbf{z})}{1 - D(\mathbf{r} \circ \mathbf{z})} \right] + \eta_S ||\mathbf{r}||_1 + \eta_H H(\mathbf{r}), \tag{23}$$

which is minimized over (θ, ν), $\mathbf{r}$, and $\{\hat{a}_j, \hat{b}_j\}_{j=1}^{d}$, together with alternating gradient updates for $D(\cdot)$. In this model, named **BF-VAE-2**, the trade-off parameters η_S and η_H control the cardinality of relevant factors[11] large η encourages few strong factors; for η small, many weak factors could be learned. The learned relevance vector $\mathbf{r}$ can serve as an indicator discerning relevant factors from nuisances.

4. Related Work

Most recent approaches to unsupervised disentanglement consider the learning objectives combining the VAE's loss in (4) with regularization terms that encourage prior latent factor independence. In **β-VAE** [13], the expected KL term of the VAE's objective is overemphasized, which can be seen as a proxy for the prior matching, i.e., minimizing $\mathrm{KL}(q(\mathbf{z}) || p(\mathbf{z}))$. In **AAE** [21], they aim to directly minimize the latter term via adversarial learning. As illustrated in our analysis in Sec. 2, the full independence of $q(\mathbf{z})$ imposed in the TC, is important in the factor disentanglement, where the TC was estimated by the discriminator density ratio in **Factor-VAE** [15], whereas **TC-VAE** [6] employed a weighted sampling strategy. Another alternative is the adversarial learning to minimize the Jensen-Shannon divergence in [5], instead of KL in the TC. Quite closely related to the TC are: **DIP-VAE** [18] that penalized the deviation of the variance of $q(\mathbf{z})$ from identity, and **InfoGAN** [7] that aimed to minimize the reconstruction error in the $\mathbf{z}$-space in addition to the reconstruction error in the $\mathbf{x}$-space.

Recent deep representational learning attempts to extend the VAE by either adopting non-Gaussian prior models or partitioning latent variables into groups that are treated differently, both seemingly similar to our approach. In [9], a hybrid model that jointly represents discrete and continuous latents was introduced. In [22], under the partially labeled data setup, they separately treated the factors associated with the labels from those that are not, leading to a conditional factor model. The Gaussian prior assumption in VAE has been relaxed to allow more flexibility and/or better fit in specific scenarios. In **VampPrior** [26], they came up with a reasonable encoder-based finite mixture model that approximates the infinite mixture model. In [8] the von Mises-Fisher density was adopted to account for a hyper-spherical latent structure. The recent **CHyVAE** [1] employed the inverse-Wishart prior (generalization of Gamma), however, it mainly dealt with situations where latents can be correlated with one another apriori, via full prior covariance. The **Hierarchical Factor VAE** [11] instead focused on independence of groups of latent variables (group disentanglement). Although these recent works are closely related to ours, they either focused on different disentanglement goals, or extended the priors for inreased model capacity.

5. Evaluation

We evaluate our approaches[12] on several benchmark datasets, where we assess the goodness of disentanglement both quantitatively and qualitatively. The former applies only to fully factor-labeled datasets, and we consider a comprehensive suite of disentanglement metrics in Sec. 5.1. Qualitative assessment is accomplished through visualizations of data synthesis via latent space traversal. We also verify in Sec. 5.2 that the visually relevant/important aspects accurately correspond to those determined by the indicators we hypothesized in each of our three models.

1) Datasets. We test all methods on the following datasets: `3D-Face` [25], `Sprites` [23] and its recent extension (`C-Spr`) [20] that fills the sprites with some random color (regarded as noise), `Teapots` [10], and `Celeb-A` [19]. Also, we consider the subset of `Sprites` containing only the oval shape[13], denoted by `O-Spr`. The details of the datasets are described in the Supplement. All datasets provide ground-truth factor labels except for `Celeb-A`. For all datasets, the image sizes are normalized to 64 × 64, and the pixel intensity/color values are scaled to $[0, 1]$. We use cross entropy as the reconstruction loss.

2) Competing Approaches. We contrast our models

[10] See Supplement for the derivations.

[11] We empirically demonstrate this in Sec. 5.2 and Supplement.

[12] Our code is publicly available in https://seqam-lab.github.io/BFVAE/

[13] Since the shape factor is in nature a discrete variable, the underlying models that assume continuous latent variables would be suboptimal. Instead of explicitly modeling a combination of discrete/continuous latent variables as in the recent hybrid model [9], we eliminate this discrete factor by considering only the oval-shape images only.

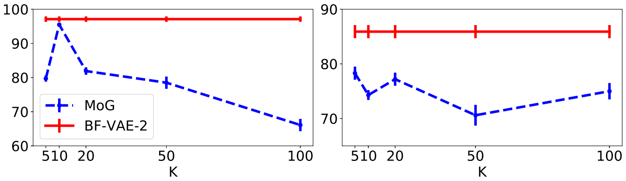

Figure 2. Disentanglement performance (Metric II) of F-VAE with MoG prior (Blue/Dashed) with different mixture orders (K) vs. BF-VAE-2 (Red/Solid) on `O-Spr` (Left) and `Sprites` (Right).

with **VAE** [17], **β-VAE** [13], and **F-VAE** (Factor-VAE) [15]. We also compare our BF-VAE models with the recent **RF-VAE** [16] that also considers differential treatment of relevant and nuisance latents.

3) Model Architectures and Optimization. We adopt the model architectures and optimization parameters similar to those in [15]. See Supplement for the details.

5.1. Quantitative Results

We consider three disentanglement metrics[14]: i) **Metric I** [15] collects data samples with one ground-truth factor fixed with the rest randomly varied, encodes them as $\boldsymbol{z}$, finds the index of the latent with the smallest variance, and measures the accuracy of classification from that index to the ID of the fixed factor (the higher the better), ii) **Metric II** [16] modifies Metric I by collecting samples of one factor varied with others fixed, and seeks the index of the largest latent variance. iii) **Metric III** [10] is based on regression from the latent vector to individual ground-truth factors, measuring three scores of prediction quality: *Disentanglement* for degree of dedication to each target, *Completeness* for degree of exclusive contribution by each covariate, and *Informativeness* for prediction error. Hence, higher scores are better for D and C, lower for I.

Tab. 1 summarizes all results, datasets and metrics. For all models across all datasets we use the latent dimension $d = 10$. Our models clearly outperform competing methods across all metrics in most instances. They are followed by RF-VAE, which also employs a notion of relevance, but not explicit non-Gaussianity.

Comparison w/ High Capacity Priors. Our analysis in Sec. 2 states that a relevant dimension prior $p(z_j)$ needs to be non-Gaussian, flexible enough to match the aggregate posterior $q(z_j)$. Here, we consider an alternative prior with those properties. Specifically, we use a F-VAE model with a Gaussian mixture prior $p(\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)$, with $\{(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_k$ the model parameters to be optimized in conjunction with the F-VAE's parameters. We contrast that model to our BF-VAE-2. The disentanglement performances (Metric II scores) on `O-Spr` and `Sprites` are summarized in Fig. 2, where we change the number of mixture components K to control the degree of flexibility of the F-VAE mixture prior. Results show the high capacity mixture consistently underperforms our Bayesian model; as K increases, it suffers from clear overfitting. This suggests the uncontrolled complex prior can be detrimental, in contrast to our controlled treatment of relevances.

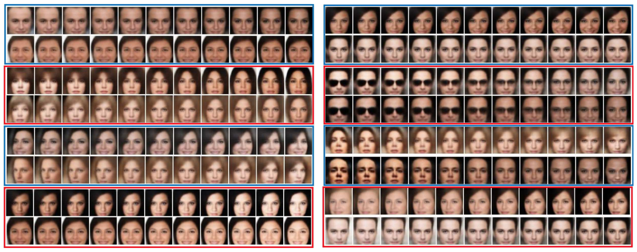

Figure 3. Latent space traversal in BF-VAE-2 on `Celeb-A`. We train two BF-VAE-2 models with two different η values ($\eta = \eta_S = \eta_H$ large and small). **(Left panel: strong factors)** contains latent traversal results with four latent variables (two subjects for each) that are detected (according to high r_j) by both η small and large models. They correspond to (from top to bottom): gender, frontal hair, azimuth, and brightness, which are considered as strong/major factors. **(Right panel: weak factors)** shows traversal with four other latent variables that are detected (according to high r_j) only by the small η model. They correspond to: smiling, sunglasses, elevation, and baldness, which are considered as weak/minor factors. See Supplement for the enlarged images and further details.

5.2. Qualitative Results

In this section we investigate qualitative performance of our BF-VAE approaches. We focus on: i) *Latent space traversal*: We depict images synthesized by traversing a single latent variable at a time while fixing the rest, and ii) *Accuracy of variable relevance indicator*: As discussed in Sec. 3, our models have implicit/explicit indicators that point to relevant and nuisance variables. Specifically, i) BF-VAE-0 (learned α_j): j relevant if α_j is away from 1, while j is nuisance if $\alpha_j \approx 1$, ii) BF-VAE-1 (DOF of the corrected prior $\overline{p}(z_j)$, equal to $2\hat{a}_j$): j is relevant if $\hat{a}_j$ is small (distant from Gaussian), and vice versa, iii) BF-VAE-2 (learned relevance indicator variable r_j): j is relevant if r_j is large, and vice versa.

Due to the lack of space, we report selected results in this section, with more extensive results in Supplement. Results are shown for `3D-Face` (Fig. 4), `O-Spr` (Fig. 5), and `Teapots` (Fig. 6). The latent space traversal demonstrates that variation of each latent variable while the others are held fixed, visually leads to change in one of the ground-truth factors exclusively (except for the `Teapots`). Also, these visually identified factors indeed correspond to those variables indicated as relevant by our models. See the figure captions for details.

Control of Cardinality of Relevant Factors. One of the distinguishing benefits of our BF-VAE-2 (and also BF-VAE-0) is that the trade-off parameter(s) η can control the

[14] More details can be found in the Supplement.

Tab. 1. Disentanglement metrics for benchmark datasets. For Metric III, the three figures in each cell indicate Disentanglement / Completeness / Informativeness (top row based on the LASSO regressor, the bottom on the Random Forest. Note that the higher the better for D and C, while the lower the better for I. The best scores for each metric (within the margin of significance) among the competing models are shown in red and second-best in blue.

Datasets/Metrics		VAE	β-VAE	F-VAE	RF-VAE	BF-VAE-0	BF-VAE-1	BF-VAE-2
3D-Face	I	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	99.9 ± 0.1	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	II	93.4 ± 0.7	95.5 ± 0.6	92.8 ± 1.1	95.2 ± 0.5	95.6 ± 0.5	97.2 ± 0.5	97.5 ± 0.5
	III	.96 / .81 / .37 .99 / .84 / .26	.96 / .78 / .40 .98 / .86 / .31	1.0 / .82 / .36 .96 / .83 / .25	1.0 / 1.0 / .48 1.0 / .93 / .37	1.0 / 1.0 / .45 1.0 / .90 / .33	1.0 / 1.0 / .45 1.0 / .90 / .34	1.0 / 1.0 / .44 1.0 / .88 / .41
Sprites	I	80.2 ± 0.3	80.8 ± 0.8	81.9 ± 1.0	85.4 ± 1.2	87.9 ± 0.9	93.8 ± 0.6	85.5 ± 0.8
	II	58.2 ± 1.4	76.8 ± 0.9	77.6 ± 1.4	79.1 ± 1.3	82.7 ± 1.1	82.2 ± 0.6	85.9 ± 1.2
	III	.59 / .68 / .52 .57 / .69 / .46	.67 / .69 / .53 .72 / .84 / .40	.84 / .84 / .53 .73 / .82 / .41	.85 / .87 / .53 .73 / .83 / .41	.89 / 1.0 / .60 .75 / .83 / .44	.92 / .90 / .54 .75 / .83 / .34	.88 / 1.0 / .58 .75 / .86 / .48
C-Spr	I	79.8 ± 0.6	81.2 ± 0.4	85.6 ± 0.8	80.7 ± 0.9	87.7 ± 0.5	93.2 ± 0.6	94.7 ± 0.8
	II	61.2 ± 1.5	74.3 ± 1.7	76.2 ± 0.8	81.4 ± 1.1	83.0 ± 1.4	84.2 ± 1.1	83.5 ± 0.7
	III	.52 / .55 / .54 .58 / .62 / .51	.77 / .82 / .53 .73 / .83 / .39	.79 / .76 / .52 .75 / .83 / .42	.87 / .91 / .54 .64 / .72 / .30	1.0 / .95 / .56 .88 / .83 / .47	.95 / .95 / .58 .79 / .88 / .42	.86 / .91/ .56 .84 / .85 / .45
O-Spr	I	97.2 ± 0.4	75.3 ± 0.6	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	II	53.2 ± 1.5	70.2 ± 1.2	80.6 ± 1.1	95.4 ± 0.5	97.8 ± 0.7	99.8 ± 0.2	97.1 ± 0.8
	III	.42 / .43 / .54 .32 / .55 / .46	.58 / .49 / .49 .56 / .58 / .36	1.0 / .88 / .33 .81 / .84 / .24	1.0 / .99 / .49 .93 / .87 / .22	1.0 / 1.0 / .42 .99 / .93 / .22	1.0 / .97 / .40 .99 / .92 / .21	1.0/ .93 / .42 .98 / .91 / .23
Teapots	I	90.1 ± 0.9	56.9 ± 1.1	91.9 ± 0.8	98.7 ± 0.4	94.8 ± 1.2	97.6 ± 0.3	97.9 ± 0.4
	II	77.7 ± 1.3	47.3 ± 0.9	74.6 ± 1.8	83.1 ± 1.2	90.4 ± 1.0	82.7 ± 1.3	88.9 ± 0.8
	III	.60 / .53 / .40 .81 / .72 / .31	.31 / .27 / .72 .45 / .61 / .52	.63 / .61 / .46 .75 / .78 / .29	.63 / .56 / .37 .90 / .79 / .27	.72 / .61 / .34 .89 / .80 / .25	.70 / .65 / .48 .78 / .80 / .50	.67 / .62 / .41 .87 / .80 / .32

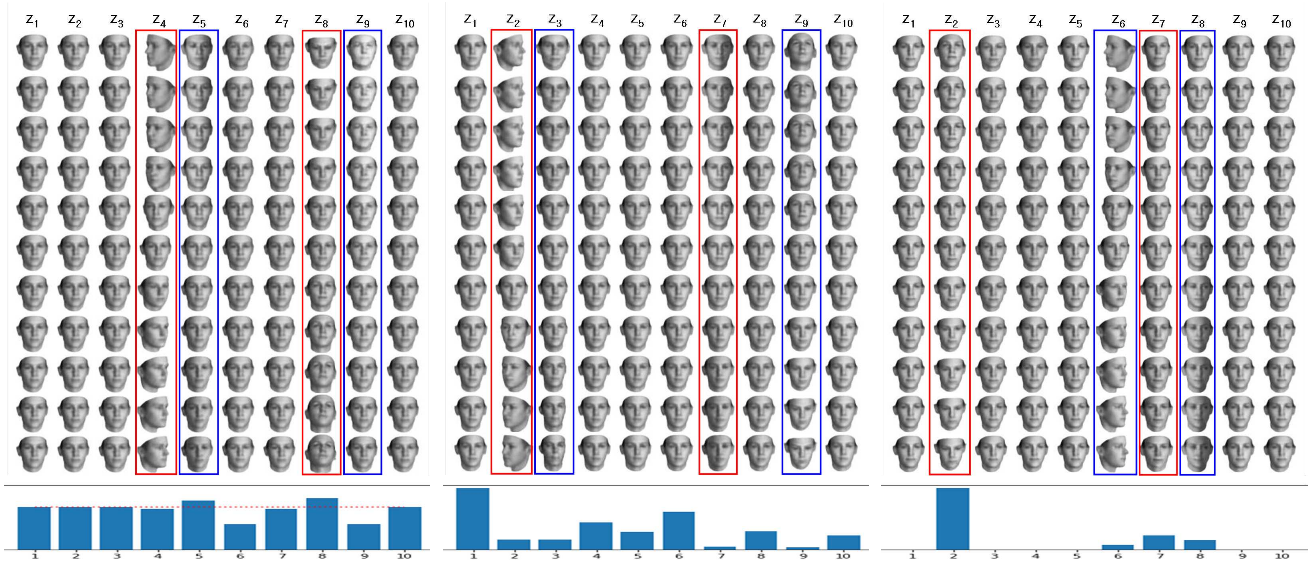

Figure 4. Latent space traversal in our three BF-VAE models on the 3D-Face dataset. (Left) BF-VAE-0 with the learned prior variances $\boldsymbol{\alpha}^{-1}$ at the bottom (the value 1.0 depicted as the red dotted line), (Middle) BF-VAE-1 with the DOF ($2\hat{a}_j$) of the corrected prior $\overline{p}(z_j)$ at the bottom, and (Right) BF-VAE-2 with the learned relevance vector $\mathbf{r}$ at the bottom. **(Left: BF-VAE-0)** The four visually evident dimensions of variability (z_4, z_5, z_8, z_9) are highlighted within colored boxes, where each exactly matches one of the four ground-truth factors ($z_4 =$ azimuth, $z_5 =$ lighting, $z_8 =$ elevation, and $z_9 =$ subject ID). The learned α_j for all these four dims are away from 1. **(Middle: BF-VAE-1)** The four recovered, highlighted, dimensions match the ground-truth factors, and their $\overline{p}(z_j)$'s also have relatively small DOFs, as expected. **(Right: BF-VAE-2)** Again the four factors are nearly correctly identified, corresponding to the high values in the indicator variables r_j's.

number of relevant factors to be detected by the model. We visually verify this on Celeb-A dataset. As shown in Fig. 3 (detailed in the caption), adopting large η leads only strong factors to be detected, while having small η allows many weak factors identified.

6. Conclusion

This work demonstrates that, for recovery of disentangled factors of variation in data, it is essential to embrace and model the non-Gaussian nature of relevant factors while, at the same time, discerning them from Gaussian nuisances, in

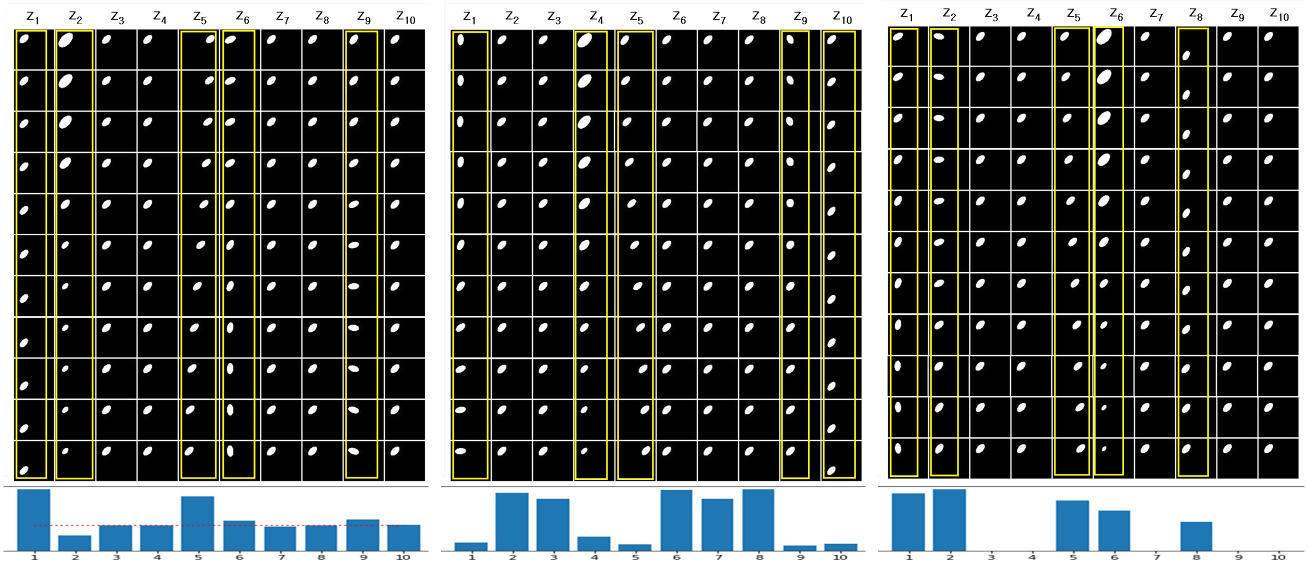

Figure 5. Latent traversal on `O-Spr`. The same interpretation as Fig. 4. **(Left: BF-VAE-0)** Those five highlighted dimensions of major variability (z_1, z_2, z_5, z_6, z_9), match the four ground-truth factors (scale, X-, Y-pos, rotation), while the rotation is spread across z_6 and z_9. These factors also exactly correspond to the learned α_j's that are distant from 1, as we anticipated. **(Middle: BF-VAE-1)** Similar to BF-VAE-0, it identifies five variables with the rotation spread across z_1 and z_9. These relevant variables, as expected, have small DOFs in $\overline{p}(z_j)$'s. **(Right: BF-VAE-2)** Again very similar to the previous two models. The learned $\mathbf{r}$ accurately indicates the relevant dimensions.

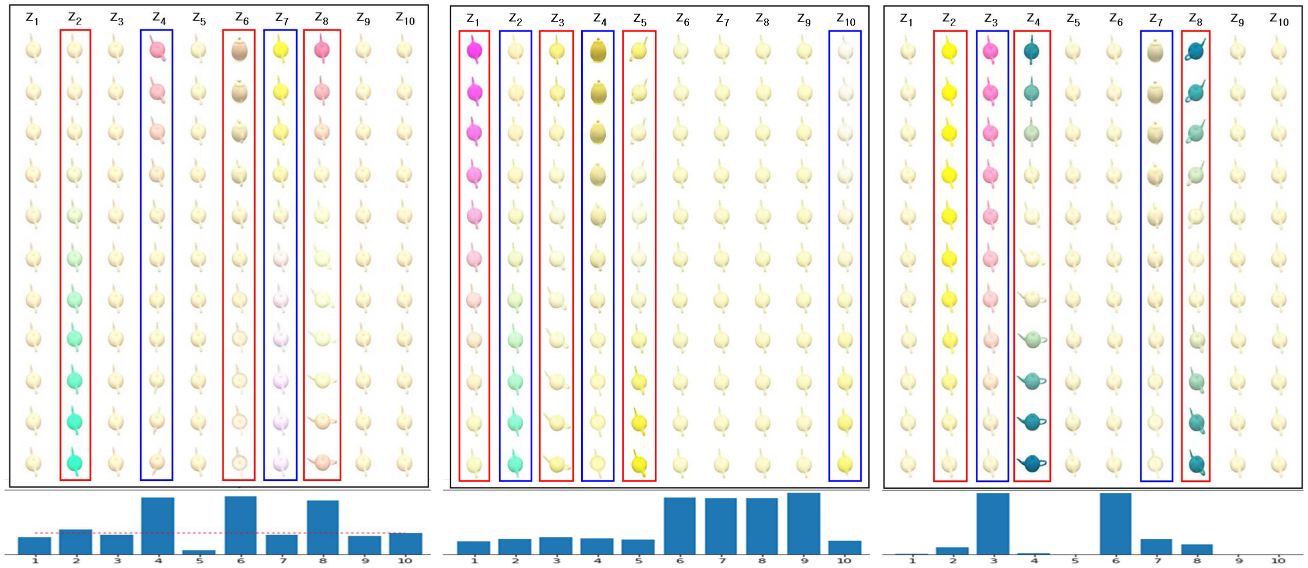

Figure 6. Latent traversal on `Teapots`. The same interpretation as Fig. 4. Note that the five ground-truth factors in this dataset are: two pose variations (azimuth and elevation) and three color changes (R,G,B). **(Left: BF-VAE-0)** The five variables that explain the major variability in images, (z_2, z_4, z_6, z_7, z_8), do not perfectly match the true factors one by one, and two or more factors are entangled in some variables (e.g., z_8 explains both color R and azimuth. Note that a similar failure was also observed in [10] with complex ResNet models. **(Middle: BF-VAE-1)** and **(Right: BF-VAE-2)** Overall similar behaviors as BF-VAE-0, but the relevance indicators (implicit DOF in BF-VAE-1 and the explicit relevance vector $\mathbf{r}$ in BF-VAE-2) correctly identify the dimensions of major variability.

contrast to traditional prior assumptions used for VAEs. We showed that a VAE endowed with a hierarchical Bayesian prior, the BF-VAE, can effectively model both aspects of this task. Empirical evaluation on benchmark datasets validates this ability of the BF-VAE family, showing consistently leading performance across three disentanglement metrics. We also demonstrated the models' ability to recover strong indicators of data variability, with clear qualitative effects observed through traversals in the learned factor space and re-synthesis of data via the models' learned decoding stage.

References

[1] Abdul Fatir Ansari and Harold Soh. Hyperprior induced unsupervised disentanglement of latent representations, 2018. arXiv:1809.04497. 5

[2] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002. 2

[3] 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. 1

[4] Peter J. Bickel, Chris A.J. Klaassen, Ya'acov Ritov, and Jon August Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York, 1998. 2

[5] Philemon Brakel and Yoshua Bengio. Learning independent features with adversarial nets for non-linear ICA. In *arXiv preprint*, 2017. 1, 5

[6] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, 2018. 1, 5

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing Generative Adversarial Nets, 2016. In Advances in Neural Information Processing Systems. 1, 5

[8] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical Variational Auto-Encoders, 2018. Uncertainty in AI. 5

[9] Emilien Dupont. Learning disentangled joint continuous and discrete representations, 2018. In Advances in Neural Information Processing Systems. 5

[10] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations, 2018. In Proceedings of the Second International Conference on Learning Representations, ICLR. 5, 6, 8

[11] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations, 2018. arXiv:1804.02086v4. 5

[12] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018. arXiv:1812.02230. 1, 2

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. 1, 5, 6

[14] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, New York, 2001. 2

[15] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. International Conference on Machine Learning, 2018. 1, 3, 4, 5, 6

[16] Minyoung Kim, Yuting Wang, Pritish Sahu, and Vladimir Pavlovic. Relevance Factor VAE: Learning and Identifying Disentangled Factors, 2019. arXiv:1902.01568. 6

[17] Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes, 2014. In Proceedings of the Second International Conference on Learning Representations, ICLR. 1, 2, 6

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variation inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. 1, 5

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 5

[20] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018. arXiv:1811.12359. 5

[21] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, and Ian Goodfellow. Adversarial autoencoders. In *International Conference on Learning Representations*, 2016. 1, 5

[22] Michael Mathieu, Junbo Zhao, Pablo Sprechmann, Aditya Ramesh, and Yann LeCun. Disentangling factors of variation in deep representations using adversarial training, 2016. In Advances in Neural Information Processing Systems. 5

[23] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset, 2017. 5

[24] Robert Brian O'Hara and Mikko J. Sillanpää. A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis*, 4(1):85–117, 2009. 4

[25] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition, 2009. Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance. 5

[26] Jakub M. Tomczak and Max Welling. VAE with a VampPrior, 2018. Artificial Intelligence and Statistics. 3, 5