UNPICKING DATA AT THE SEAMS: VAES, DISENTANGLEMENT AND INDEPENDENT COMPONENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Disentanglement, or identifying salient statistically independent factors of the data, is of interest in many areas of machine learning and statistics, with relevance to synthetic data generation with controlled properties, robust classification of features, parsimonious encoding, and a greater understanding of the generative process underlying the data. Disentanglement arises in several generative paradigms, including Variational Autoencoders (VAEs), Generative Adversarial Networks and diffusion models. Particular progress has recently been made in understanding disentanglement in VAEs, where the choice of diagonal posterior covariance matrices is proposed to promote mutual orthogonality between columns of the decoder's Jacobian. We continue this thread to show how this *linear* independence translates to *statistical* independence, completing the chain in understanding how the VAE's objective identifies independent components of, or disentangles, the data.

1 INTRODUCTION

Variational Autoencoders (**VAE**s, Kingma (2013); Rezende et al. (2014)) and a range of variants, e.g. β-VAE (e.g. Higgins et al., 2017) and Factor-VAE (Kim & Mnih, 2018), have been shown empirically to *disentangle* latent factors of variation in the data. For example, a trained VAE may generate face images that vary in distinct semantically meaningful ways, such as hair colour or facial expression, as individual latent variables are adjusted. This is both of practical use, e.g. for controlled generation of synthetic data with chosen properties, and intriguing as it is not knowingly designed into the training algorithm. A related phenomenon is observed in samples from a Generative Adversarial Network (GAN), which, in common with a VAE, applies a deterministic neural network function to samples of independently distributed latent variables, producing a *push-forward* distribution.

Understanding why disentanglement arises, seemingly "for free", is of interest since identifying and separating generative factors underlying the data goes to the heart of many aspects of machine learning, from classification to generation, interpretability to identifiability, right down to a fundamental understanding of the data itself. With a better appreciation of why disentanglement happens, we might be able to induce it more reliably, particularly in domains where we cannot easily perceive when features are disentangled, as we can for images and text.

Research into the cause of disentanglement has gradually led to a refined understanding of what is meant by "disentanglement", which typically refers to the separation of semantically meaningful generative factors (Bengio et al., 2013). Recent progress has been made towards understanding why disentanglement occurs in VAEs, tracing the root cause to the common use of *diagonal posterior covariance matrices*, a seemingly innocuous design choice made for computational efficiency (Rolinek et al., 2019; Kumar & Poole, 2020). Diagonal covariances are shown to promote orthogonality between columns of the Jacobian of the decoder, a property linked to disentangled features (Ramesh et al., 2018) and independent causal mechanisms (Gresele et al., 2021). We extend this line of work by providing a firmer basis for the covariance-orthogonality relationship together with theoretical analysis to show how orthogonality in the Jacobian translates to disentanglement in the push-forward distribution of a VAE, connecting *linear* independence of partial derivatives to *statistical* independence of components, or generative factors, of the data.

Interest in understanding how disentanglement arises in VAEs has increased as their generative quality has improved (e.g. Hazami et al., 2022) and they often a key component in state of the art diffusion models, where disentanglement is of great interest (Pandey et al., 2022; Zhang et al., 2022; Yang et al., 2023).

In this work, we analyse and extend recent advances in understanding disentanglement in VAEs by

- proving that orthogonality, or *linear independence*, between columns of the decoder Jacobian corresponds to identifying *statistically independent* components of the generative distribution with distinct latent variables (§4.1);
- providing conditions under which a VAE fully identifies the data distribution (§4.1); and
- presenting a novel interpretation of β in a β-VAE, as scaling the variance of the likelihood distribution, explaining why β affects both disentanglement and "posterior collapse" (§4.2).

2 BACKGROUND

Disentanglement: Disentanglement is not consistently defined in the literature, but typically refers to identifying salient, semantically meaningful features of the data with distinct latent variables, such that by varying a single variable, data can be generated that differ in a single aspect (Bengio et al., 2013; Higgins et al., 2017; Ramesh et al., 2018; Rolinek et al., 2019). Disentanglement has also been decomposed into necessary and sufficient-type concepts of *consistency* and *restrictiveness* (Shu et al., 2019). We show that disentanglement in a VAE relates to identifying *statistically independent components* of the data, comparable to independent component analysis (**ICA**).

Variational Autoencoder (VAE): A VAE is a latent generative model for data $x \in \mathcal{X} \doteq \mathbb{R}^m$, that models the data distribution by $p_\theta(x) = \int_z p_\theta(x|z)p(z)$ with parameters θ and latent variables $z \in \mathcal{Z} \doteq \mathbb{R}^d$. A VAE is trained by maximising a lower bound to the log likelihood (the **ELBO**),

$$\int_x p(x) \log p_\theta(x) \quad \geq \quad \int_x p(x) \int_z q_\phi(z|x) \big\{ \log p_\theta(x|z) - \beta \log \tfrac{q_\phi(z|x)}{p(z)} \big\} \quad \doteq \; \ell(\theta, \phi) \, , \tag{1}$$

where $q_\phi(z|x)$ learns to approximate the model posterior $p_\theta(z|x) \doteq \frac{p_\theta(x|z)p(z)}{p_\theta(x)}$; and $\beta = 1$. A VAE parameterises distributions by neural networks: $q_\phi(z|x) = \mathcal{N}(z; e(x), \Sigma_x)$ has mean $e(x)$ and diagonal covariance Σ_x output by an *encoder* network; and $p_\theta(x|z)$ is typically of exponential family form (e.g. Bernoulli or Gaussian) with natural parameter $\theta \doteq d(z)$ defined by a *decoder* network.[1] The prior is commonly a standard Gaussian $p(z) = \mathcal{N}(z; \mathbf{0}, \boldsymbol{I})$.

While samples generated from a VAE ($\beta = 1$) can exhibit disentanglement, setting $\beta > 1$ is found empirically to enhance the effect, typically with a cost to generative quality (Higgins et al., 2017).

Probabilistic Principal Component Analysis (PPCA): PPCA (Tipping & Bishop, 1999) considers a *linear* latent variable model with parameters $\boldsymbol{D} \in \mathbb{R}^{m \times d}$, $\sigma \in \mathbb{R}$ and noise $\epsilon \in \mathbb{R}^m$.[2]

$$x = \boldsymbol{D}z + \epsilon \qquad z \sim p(z) = \mathcal{N}(z;\, \mathbf{0}, \boldsymbol{I}) \qquad \epsilon \sim p_\sigma(\epsilon) = \mathcal{N}(\epsilon;\, \mathbf{0}, \sigma^2 \boldsymbol{I}) \, , \tag{2}$$

All distributions are Gaussian and known analytically, in particular the model posterior is given by

$$p_\theta(z|x) = \mathcal{N}(z;\, \tfrac{1}{\sigma^2}\boldsymbol{M}\boldsymbol{D}^\top x,\, \boldsymbol{M}) \qquad \text{where} \qquad \boldsymbol{M} = (\boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}^\top \boldsymbol{D})^{-1} \, . \tag{3}$$

The maximum likelihood solution is fully tractable: $\boldsymbol{D}_{\text{PPCA}} = \boldsymbol{U}(\boldsymbol{S} - \sigma^2 \boldsymbol{I})^{1/2}\boldsymbol{R}$, where $\boldsymbol{S} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{U} \in \mathbb{R}^{m \times d}$ contain the largest eigenvalues and corresponding eigenvectors of the data covariance $\boldsymbol{X}\boldsymbol{X}^\top$, and $\boldsymbol{R} \in \mathbb{R}^{d \times d}$ is orthonormal ($\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}$). As $\sigma^2 \to 0$, $\boldsymbol{D}_{\text{ML}}$ approaches the singular value decomposition (**SVD**) of the data $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ up to ambiguity in $\boldsymbol{V}$, as in classical PCA. Due to the ambiguity in $\boldsymbol{R}/\boldsymbol{V}$, the model is considered *unidentified*. While the exact solution is known, it can also be numerically approximated by optimising the ELBO, e.g. (Eq. 1) by *expectation maximisation* (**PPCA^{EM}**), where (**E**) sets $q_\phi(z|x)$ to its exact optimum $p_\theta(z|x)$ in Eq. 3; and (**M**) optimises w.r.t. θ.

Linear VAE (LVAE): A VAE with Gaussian likelihood $p_\theta(x|z) \doteq \mathcal{N}(x; d(z), \sigma^2 \boldsymbol{I})$ and linear decoder $d(x) = \boldsymbol{D}x$ (termed a *linear VAE* assumes the same underlying model as PPCA (2). Indeed, training an LVAE differs to PPCA^{EM} only in approximating the posterior by $q_\phi(z|x) = \mathcal{N}(z; \boldsymbol{E}x, \Sigma)$

[1]For Gaussian distributions, a fixed variance parameter σ^2 is also specified.
[2]Throughout, we assume data is centred which equates to including a mean parameter (Tipping & Bishop, 1999).

rather than computing its analytic optimum. While the latter may seem preferable, Lucas et al. (2019) showed that an LVAE with diagonal Σ *breaks the symmetry* of PPCA. This follows from Σ being both diagonal and optimal per Eq. 3,

$$\Sigma \;=\; \boldsymbol{M}_{\text{PPCA}} \;\doteq\; (\boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}_{\text{PPCA}}^\top \boldsymbol{D}_{\text{PPCA}})^{-1} \;=\; \sigma^2 \boldsymbol{R}^\top \boldsymbol{S}^{-1}\boldsymbol{R} \qquad \forall x, \tag{4}$$

(by definition of $\boldsymbol{D}_{\text{PPCA}}$). This requires $\boldsymbol{R} = \boldsymbol{I}$ and restricts the solution of an LVAE to $\boldsymbol{D}_{\text{LVAE}} = \boldsymbol{U}(\boldsymbol{S} - \sigma^2\boldsymbol{I})^{1/2}$ (*cf* $\boldsymbol{D}_{\text{PPCA}}$), up to trivial transformations (axis permutation and sign).

Orthogonality in a VAE Decoder's Jacobian: Beyond symmetry breaking in *linear* VAEs, diagonal posterior covariances are shown to promote disentanglement in *non-linear* VAEs by inducing columns of the decoder's Jacobian to be mutually orthogonal (Rolinek et al., 2019; Kumar & Poole, 2020). The generalised argument of Kumar & Poole (2020) reparameterises around the encoder mean, $z = e(x) + \epsilon$, $\epsilon \sim \mathcal{N}(\boldsymbol{0}, \Sigma_x)$, and Taylor expands to approximate a *deterministic* ELBO (**det-ELBO**):

$$\ell(x) = \mathbb{E}_{\epsilon|x}\Big[\log p_\theta(x|z = e(x) + \epsilon) \;-\; \beta \underbrace{\log \tfrac{p(\epsilon)}{p(z = e(x)+\epsilon)}}_{\text{KL}}\Big] \qquad \text{(Reparameterise)}$$

$$= \mathbb{E}_{\epsilon|x}\Big[\log p_\theta(x|z = e(x)) \;+\; \epsilon^\top \boldsymbol{j}_{e(x)}(x) \;+\; \tfrac{1}{2}\epsilon^\top \boldsymbol{H}_{e(x)}(x)\epsilon + O(\epsilon^3) \;-\; \beta\,\text{KL}\Big] \qquad \text{(Taylor)}$$

$$\approx \underbrace{\log p_\theta(x|z = e(x))}_{\text{AE}} + \tfrac{1}{2}\underbrace{\boldsymbol{H}_{e(x)}(x) \odot \Sigma_x}_{\text{gradient regularisation}} - \tfrac{\beta}{2}\Big(\underbrace{\|e(x)\|^2 + \text{tr}(\Sigma_x)}_{\textbf{prior}} - \underbrace{\log|\Sigma_x| - d}_{\textbf{entropy}}\Big). \tag{5}$$

Here, $\boldsymbol{j}_{z^*}(x) \doteq (\frac{\partial \log p_\theta(x|z)}{\partial z_i})_i$ and $\boldsymbol{H}_{z^*}(x) \doteq (\frac{\partial^2 \log p_\theta(x|z)}{\partial z_i \partial z_j})_{i,j}$ are the Jacobian and Hessian of $\log p_\theta$ evaluated at $z^* \in \mathcal{Z}$; and $\odot$ is the Frobenius (dot) product. Notably, $\mathbb{E}_{\epsilon|x}[O(\epsilon^3)]$ terms are dropped. Differentiating Eq. 5 w.r.t. Σ_x suggests a connection between the Hessian and encoder variance:

$$\nabla_{\Sigma_x}\ell(x) \approx \tfrac{1}{2}\big(\boldsymbol{H}_{e(x)}(x) - \beta(\boldsymbol{I} - \Sigma_x^{-1})\big) \quad \Rightarrow \quad \boxed{\Sigma_x^{-1} \approx \boldsymbol{I} - \tfrac{1}{\beta}\boldsymbol{H}_{e(x)}(x)}\,. \tag{6}$$

As in the linear case, the ELBO with diagonal Σ_x is thus minimised if the likelihood's Hessian is also diagonal. For exponential family $p_\theta(x|z)$ with decoder-defined natural parameter $\theta = d(z)$,

$$\boldsymbol{H}_{e(x)}(x) \;=\; -\boldsymbol{D}_{e(x)}^\top \boldsymbol{A}_{d\circ e(x)}^2 \boldsymbol{D}_{e(x)} \;+\; (x - \hat{x}_{d\circ e(x)})^\top \mathbf{D}_{e(x)}\,, \tag{7}$$

where $\boldsymbol{D}_{z^*}$ / $\mathbf{D}_{z^*}$ are the Jacobian / Hessian of the decoder evaluated at z^*; $\hat{x}_{\theta^*} = \mathbb{E}[x|\theta^*]$; and $\boldsymbol{A}_{\theta^*}^2 = -\frac{d^2}{d\theta^2}\log p_{\theta^*}(x|z) = \text{Var}[x|\theta^*]$. Note, $\boldsymbol{A}_\theta$ is diagonal if dimensions x_i of x are conditionally independent given θ, e.g. $\boldsymbol{A}_\theta = \frac{1}{\sigma}\boldsymbol{I}$ for Gaussian $p_\theta(z|x)$. The key conclusion is that for Gaussian likelihoods and commonly used decoders where the last term in Eq. 7 is small (e.g. ReLU networks or similar), $\Sigma_x^{-1} \approx \boldsymbol{I} + \frac{1}{\beta\sigma^2}\boldsymbol{D}_{e(x)}^\top \boldsymbol{D}_{e(x)}$, and so *columns of the decoder's Jacobian are orthogonal.*

3 From Diagonal Posterior Covariance to Orthogonality

Before building on it, we first make *precise* the relationship between posterior covariance and the log likelihood's Hessian in Eq. 6. Maximising the ELBO with a Gaussian posterior approximation is equivalent to an *averaged Laplace approximation* (Opper & Archambeau, 2009). Hence at optimality,

$$\Sigma_x^{-1} \;=\; -\tfrac{d}{d\Sigma_x}\mathbb{E}_{q(z|x)}[\log p_\theta(x,z)] \;=\; -\mathbb{E}_{q(z|x)}[\tfrac{d^2}{dz^2}\log p_\theta(x,z)] \;=\; \boldsymbol{I} - \tfrac{1}{\beta}\mathbb{E}_{q(z|x)}[\boldsymbol{H}_z(x)]\,, \tag{8}$$

which follows from: (i) differentiating the ELBO w.r.t. Σ_x; (ii) the link to Laplace approximation; and (iii) the Gaussian prior. (This also follows from the deterministic ELBO (Eq. 5) by differentiating the *full* Taylor series w.r.t. Σ_x.) Thus, Eq. 6 holds *in expectation*. Accordingly, for Gaussian likelihoods and small higher decoder derivatives (e.g. for ReLU networks), columns of the decoder Jacobian are orthogonal in expectation over each posterior $q_\phi(z|x)$. Although weaker than Eq. 6 suggests, the relationship between column-orthogonality of the decoder Jacobian and disentanglement has been verified empirically (Rolinek et al., 2019; Kumar & Poole, 2020). We conjecture that orthogonality holds more consistently the more posteriors overlap and regions of overlap are subject to multiple simultaneous orthogonality constraints (see §4.2). This less rigid relationship may also partly justify why disentanglement is observed variably in practice (Locatello et al., 2019).

Comparing optimal covariances for PPCA (Eq. 3) and Gaussian VAE (dropping β for clarity),

$$\hat{\Sigma}^{-1}_{\text{PPCA}} = \boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}^\top\boldsymbol{D} \qquad\qquad \hat{\Sigma}^{-1}_{x,\text{VAE}} = \mathbb{E}_{q(z|x)}[\boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}^\top_{e(x)}\boldsymbol{D}_{e(x)}]\ , \tag{9}$$

reveals how the optimal posterior covariance for a non-linear VAE generalises the result for the classical linear case. This is insightful, since it is this relationship that enables an LVAE to break the rotational symmetry of PPCA (Eq. 4) and, more pertinently, results in standard basis vectors $\boldsymbol{z}_i \in \mathcal{Z}$ corresponding to, or *identifying*, independent principal axes of variance of the data, $\boldsymbol{D}_{\text{LVAE}}\boldsymbol{z}_i = \boldsymbol{U}(\boldsymbol{S} - \sigma^2\boldsymbol{I})^{1/2}\boldsymbol{z}_i \propto \boldsymbol{u}_i$. In other words, the relationship that *disentangles* independent factors of variation in the linear case, is mirrored in the non-linear case.

We briefly consider the linear det-ELBO (Eq. 5), i.e. for an LVAE (dropping β for clarity),

$$\begin{aligned} 2\ell^{\text{LVAE}} &= \mathbb{E}_x\big[-\tfrac{1}{\sigma^2}\|x - \boldsymbol{D}\boldsymbol{E}x\|^2 \;-\; (\boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}^\top\boldsymbol{D}) \odot \Sigma \;+\; \log|\Sigma| \;-\; \|\boldsymbol{E}x\|^2 \;+\; d\big] \\ &= \mathbb{E}_x\big[-\tfrac{1}{\sigma^2}\|x - \boldsymbol{D}\boldsymbol{E}x\|^2 \;-\; \|\boldsymbol{E}x\|^2\big] \;-\; \log|\boldsymbol{I} + \tfrac{1}{\sigma^2}\boldsymbol{D}^\top\boldsymbol{D}|\ , \end{aligned} \tag{10}$$

where the last line assumes optimal posterior covariance (Eq. 9). This gives a fully deterministic objective to solve PPCA that can also be viewed as (theoretically-driven) regularised deterministic PCA, variants of which have been widely studied in terms of their optima (Kunin et al., 2019; Bao et al., 2020) and learning dynamics (Saxe et al., 2014; Bao et al., 2020).

Lastly, note that higher $\text{Var}[x|z] = \sigma^2$ corresponds to higher $\text{Var}[z|x] = \Sigma$ and vice versa, i.e. intuitively, uncertainty in one domain goes hand in hand with uncertainty in the other (see §4.2).

4 From Orthogonality to Disentanglement

Having clarified the connection between diagonal posterior covariance and column-orthogonality of the Jacobian, we look to deepen understanding of why such orthogonality causes disentanglement.

Note, that this relates two very different notions of "independence": orthogonality pertains to geometric *linear* independence, while disentanglement relates to *statistical* independence of factors, or components, of a distribution. These concepts do not necessarily imply one another and it is not immediate that a column-orthogonal Jacobian implies that different latent dimensions correspond to statistically independent, often semantically meaningful, factors of variation in the data.

4.1 From Linear Independence to Statistical Independence

We now develop our main result to show how optimising the ELBO with diagonal posterior covariance, a seemingly innocuous design choice, leads to disentanglement of generative factors of the data, relating *linearly independent* columns of the decoder's Jacobian to *statistically independent* components of the data.

The generative model of a VAE can be decomposed into stochastic and deterministic steps: sample the prior $z \sim p(z)$; apply a deterministic function $\hat{x} = \mu \circ d(z)$ (composing the decoder $d: z \mapsto \theta$ and mean function $\mu: \theta \mapsto \hat{x} \doteq \mathbb{E}[x|z]$); and add element-wise noise, $x \sim p(x|\hat{x})$. For continuous data, e.g. images, element-wise noise often serves only as "blur" and is omitted when generating synthetic data, which are thus sampled from the "push-forward" distribution of mean parameters $p(\hat{x})$. We consider theses distributions more generally for $\mathcal{X} = \mathbb{R}^m$, $\mathcal{Z} = \mathbb{R}^d$, $d \leq m$.

Definition 1 (push-forward distribution)**.** *For a function $f: \mathcal{Z} \to \mathcal{X}$ and prior distribution over $\mathcal{Z}$, $p_Z(z)$, the* push-forward distribution *$p^{\#}_{f,\,p_Z}(x)$ is defined implicitly over $\{x = f(z) \mid z \sim p(z)\} \subseteq \mathcal{X}$.*

Unless stated otherwise, we assume:

A#1. *Latent variables are sampled from independent standard normals, $p_Z(z) = \prod_{i=1}^d \mathcal{N}(z_i; 0, 1)$.*[3]

A#2. *$f: \mathcal{Z} \to \mathcal{X}$ is injective and differentiable.*[4]

Note that under A#2, f:

[3]Since we assume A#1 throughout, $p_Z(z)$ is generally dropped from the subscript of $p^{\#}_f$ to lighten notation.

[4]Note that continuous piece-wise linear functions are the limit of a sequence of differentiable functions, i.e. differentiable functions can be found that are arbitrarily close to a continuous piece-wise linear function.

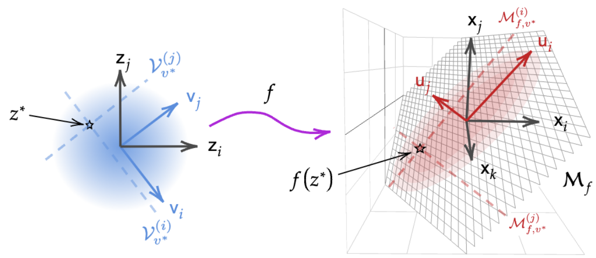

Figure 1: Translating linear independence to statistical independence (**linear $\boldsymbol{f}$**, with Jacobian $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ and manifold $\mathcal{M}_f \subseteq \mathcal{X}$): A point in $\mathcal{Z}$ is denoted z when in the standard basis $\{\boldsymbol{z}_i, \boldsymbol{z}_j\}$ and v in the $\boldsymbol{V}$-basis (columns of $\boldsymbol{V}$: $\boldsymbol{v}_i, \boldsymbol{v}_j$, solid blue). $\mathcal{V}^{(i)}_{v^*} \subseteq \mathcal{Z}$ are lines passing through v^* as each co-ordinate v_i varies (dashed blue). Each $\mathcal{V}^{(i)}_{v^*}$ maps to a sub-manifold $\mathcal{M}^{(i)}_{f,v^*} \subseteq \mathcal{M}_f$ (dashed red) passing through $f(z^*)$ parallel to the $\boldsymbol{U}$-basis (columns of $\boldsymbol{U}$: $\boldsymbol{u}_i, \boldsymbol{u}_j$, solid red). $\mathcal{V}^{(i)}_{v^*}$ are statistically independent: $p(v^*) = \prod_i p(v^*_i)$. f induces push-forward distributions $p^{\#(i)}_{f,v^*}$ over $\mathcal{M}^{(i)}_{f,v^*}$ and for $x = f_{\boldsymbol{V}}(v^*) \equiv f(z^*)$, the density over the manifold $p(x) = \prod_i p^{\#(i)}_{f,v^*}(x)$ factorises as a product of independent components.

(i) defines a d-dimensional *manifold* $\mathcal{M}_f = \{f(z) \,|\, z \in \mathcal{Z}\} \subseteq \mathcal{X}$ embedded in $\mathbb{R}^m$ supporting $p^\#_f$;
(ii) is bijective between $\mathcal{Z}$ and $\mathcal{M}_f$; and
(iii) has, by injectivity, full-rank Jacobian $\boldsymbol{J}$ evaluated at z^*, $\forall z^* \in \mathcal{Z}$.

Letting $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ denote the SVD of $\boldsymbol{J}$ ($\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$, $\boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I}$), we note that columns $\boldsymbol{v}_i \in \mathcal{Z}$ of $\boldsymbol{V}$ define a (local) orthonormal basis for $\mathcal{Z}$, while columns $\boldsymbol{u}_i \in \mathcal{X}$ of $\boldsymbol{U}$ define a basis for the tangent space to $\mathcal{M}_f$ at $f(z^*)$. Since $\boldsymbol{J}\boldsymbol{v}_i = s_i\boldsymbol{u}_i$ (where $s_i \doteq \boldsymbol{S}_{i,i}$, the i^{th} singular value), at each $z^* \in \mathcal{Z}$ the Jacobian identifies (local) basis vectors in $\mathcal{Z}$ with (local) basis vectors in $\mathcal{X}$.

4.1.1 Linear f

For intuition, we first consider the linear case $f(z) = \boldsymbol{D}z$ (satisfying A#2): $\mathcal{M}_f$ is a d-dimensional subspace (hyperplane through the origin), the Jacobian $\boldsymbol{J} = \boldsymbol{D}$ is constant $\forall z^* \in \mathcal{Z}$, and

$$p^\#_f(x = f(z)) \;=\; |\boldsymbol{D}|^{-1} p(z) \;=\; \prod_i |s_i|^{-1} p(z_i) \tag{11}$$

factorises. We now interpret factors $|s_i|^{-1}p(z_i)$. We express z in the $\boldsymbol{V}$-basis as $v = \boldsymbol{V}^\top z$, a basis transformation with Jacobian $\frac{\partial v}{\partial z} = \boldsymbol{V}^\top$. Thus $p(v) = |\boldsymbol{V}|p(z) = p(z)$, as expected given only the basis/perspective has changed. By rotational symmetry of the Gaussian prior, $p(v) = \prod_i p(v_i)$ and $p(v_i) = \mathcal{N}(0, 1)$. We can consider x as a function of z expressed in the $\boldsymbol{V}$-basis (i.e. of v), $x = f(z) = \boldsymbol{U}\boldsymbol{S}v \doteq f_{\boldsymbol{V}}(v)$, for which the partial derivatives (columns of $f_{\boldsymbol{V}}(v)$'s Jacobian) are orthogonal: $\frac{\partial x}{\partial v_i}^\top \frac{\partial x}{\partial v_j} = s_i s_j \boldsymbol{u}_i^\top \boldsymbol{u}_j = \{s_i^2 \text{ if } i = j, \text{ o/w } 0\}$; and $\|\frac{\partial x}{\partial v_i}\| = |s_i|$. Eq. 11 thus becomes

$$p^\#_f(x = f_{\boldsymbol{V}}(v)) \;=\; \prod \|\tfrac{\partial x}{\partial v_i}\|^{-1} p(v_i)\;, \tag{12}$$

where factors now have the form of uni-variate probability distributions under a change of variables. With reference to Figure 1, let $v^* = \boldsymbol{V}^\top z^*$ be an evaluation point expressed in the $\boldsymbol{V}$-basis and let $\{\mathcal{V}^{(i)}_{v^*} \subseteq \mathcal{Z}\}_i$ be lines passing through v^* as co-ordinate i in the $\boldsymbol{V}$-basis varies (dashed blue), i.e. $\mathcal{V}^{(i)}_{v^*} = \{(v^*_1, ..., v_i, ..., v^*_d) \,|\, v_i \in \mathbb{R}\}$. The image of each $\mathcal{V}^{(i)}_{v^*}$ under $f_{\boldsymbol{V}}$ forms a linear sub-manifold (or line) $\mathcal{M}^{(i)}_{f,v^*} = \{f_{\boldsymbol{V}}(v) \,|\, v \in \mathcal{V}^{(i)}_{v^*}\} \subseteq \mathcal{M}_f$ (dashed red). Evaluating at different v^*, $\mathcal{M}^{(i)}_{f,v^*}$ are always *parallel* to the (local) $\boldsymbol{U}$-basis at $f_{\boldsymbol{V}}(v) \in \mathcal{X}$, i.e. to the left singular vectors of $\boldsymbol{D}$ (solid red).

We can consider the push-forward distributions of $f_{\boldsymbol{V}}$ restricted to each line $\mathcal{V}^{(i)}_{v^*}$, which are supported on $x \in \mathcal{M}^{(i)}_{f,z^*}$ with density $p^{\#(i)}_{f,v^*}(f_{\boldsymbol{V}}(v)) = \|\frac{\partial x}{\partial v_i}\|^{-1} p(v_i)$, hence matching the factors in Eq. 12.

Thus the probability density of points on the manifold $x \in \mathcal{M}_f$,

$$p_f^{\#}(x) = \prod_i p_{f,v^*}^{\#(i)}(x)\ . \tag{13}$$

factorises as a product of densities over 1-D sub-manifolds in $\mathcal{X}$, analogously to how $p(z)$ factorises as product of 1-D Gaussians in $\mathcal{Z}$.

Since a perturbation to $x = f_{\boldsymbol{V}}(v^*)$ in basis vector $\boldsymbol{u}_i$ in $\mathcal{X}$ corresponds (exclusively) to a perturbation to v^* in basis vector $\boldsymbol{v}_i$ in $\mathcal{Z}$, only $p(v_i)$ can vary, hence $p_{f,v^*}^{\#(i)}(x)$ may also vary but no other component distribution changes. Overall, linearly independent basis vectors $\boldsymbol{v}_i$ in $\mathcal{Z}$ are projected by f to linearly independent basis vectors $\boldsymbol{u}_i$ in $\mathcal{X}$, and statistical independence over sub-manifolds defined by $\boldsymbol{v}_i$ is preserved over sub-manifolds defined by $\boldsymbol{u}_i$. To summarise, this proves:

Theorem 1. *Assuming A#1 (independent Gaussian latent variables) and a linear function $f : \mathcal{Z} \to \mathcal{X}$, $f(z) = \boldsymbol{D}z$, the push-forward distribution $p_f^{\#}$ factorises as a product of statistically independent components in $\mathcal{X}$. Statistically independent vectors in $\mathcal{Z}$ parallel to right singular vectors of $\boldsymbol{D}$ map to statistically independent vectors in $\mathcal{X}$ parallel to left singular vectors of $\boldsymbol{D}$.*

Remark 1. *The probability density over each $\mathcal{V}_{v^*}^{(i)} \subseteq \mathcal{Z}$ is a standard Gaussian $\mathcal{N}(v; 0, 1)$. The density $p_{f,v^*}^{\#(i)}(x) = s_i^{-1} p(f_V^{-1}(x)_i)$ over each sub-manifold, $\mathcal{M}_{f,z^*}^{(i)} \subseteq \mathcal{X}$, is also Gaussian $\mathcal{N}(x; 0, s_i^2)$.*

Remark 2. *The SVD of the Jacobian $\boldsymbol{J} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ can be interpreted in terms of the chain rule $\boldsymbol{J} = \frac{\partial x}{\partial u}\frac{\partial u}{\partial v}\frac{\partial v}{\partial z}$; and as $\boldsymbol{U}, \boldsymbol{V}^\top$ transforming the basis in each domain (termed the* independent bases *of f), and diagonal $\boldsymbol{S} = \frac{\partial u}{\partial v}$ is the Jacobian of f for elements expressed in the independent bases. A basis vector in one domain uniquely affects one basis vector in the other: $\frac{\partial u_i}{\partial v_j} = \{s_i \text{ if } i = j;\ 0 \text{ o/w}\}$.*

Remark 3. *Since right singular vectors $\boldsymbol{V}$ are a basis, or matter of perspective, they have no effect on the model and cannot be identified. Component order and direction (positive/negative sign) similarly have no effect on the data distribution.*

Corollary 1.1. *For data generated under the linear PPCA model (Eq. 2), an LVAE identifies statistically independent components of the data. If singular values of ground truth $\boldsymbol{D}$ are distinct, the model is fully identified.*

Proof. The PPCA model satisfies the assumptions of Theorem 1. Columns of $\boldsymbol{D}_{\text{LVAE}}$ are scalar multiples of left singular vectors of ground truth $\boldsymbol{D}$, which, by Theorem 1, define statistically independent components of the data. Identifiability follows from uniqueness of $p_{f,v^*}^{\#(i)}(x)$. □

4.1.2 Non-linear f, column-orthogonal Jacobian

Theorem 1 for a linear function may not seem surprising, but notably its proof does not rely on linearity of f. We now follow a similar argument for f that may be non-linear, assuming instead

A#3. *($\forall\, z^* \in \mathcal{Z}$) columns of $\boldsymbol{J}$ are mutually orthogonal, i.e. $\frac{\partial x}{\partial z_i}^\top \frac{\partial x}{\partial z_j} = 0, i \neq j$; equivalently $\boldsymbol{V} = \boldsymbol{I}$.*

Theorem 2. *Assuming A#1-3, the push-forward distribution $p_f^{\#}$ factorises as a product of statistically independent components in $\mathcal{X}$. At each point z^*, statistically independent vectors in $\mathcal{Z}$ parallel to the standard basis map to statistically independent vectors in $\mathcal{X}$ parallel to left singular vectors of the Jacobian $\boldsymbol{J}$ evaluated at z^*.*

Proof. The push-forward distribution of f satisfies

$$p_f^{\#}(f(z)) \ = \ |\boldsymbol{J}|^{-1} p(z) \ = \ \prod_i |s_i|^{-1} p(z_i) \ = \ \prod_i \|\tfrac{\partial x}{\partial z_i}\|^{-1} p(z_i)\ , \tag{14}$$

equivalent to Eq. 12 but without the need for a change of basis. As illustrated in Figure 2 and analogously to the linear case, let $\mathcal{Z}_{z^*}^{(i)} \subset \mathcal{Z}$ denote orthogonal lines passing through z^* parallel to the standard basis, $\mathcal{Z}_{z^*}^{(i)} = \{(z_1^*, ..., z_i, ..., z_d^*) \mid z_i \in \mathbb{R}\}$ (dashed blue). To isolate the action of f over each $\mathcal{Z}_{z^*}^{(i)}$, we define $f_{z^*}^{(i)} : \mathcal{Z}_{z^*}^{(i)} \to \mathcal{M}_f$, $f_{z^*}^{(i)}(z_i) \doteq f(z_1^*, \ldots, z_i, \ldots, z_d^*)$, which each map the line $\mathcal{Z}_{z^*}^{(i)} \subseteq \mathcal{Z}$ to a 1-D sub-manifold $\mathcal{M}_{f,z^*}^{(i)} \doteq \{f_{z^*}^{(i)}(z_i) \mid z_i \in \mathbb{R}^d\} \subset \mathcal{M}_f$ passing through $f(z^*)$ (dashed

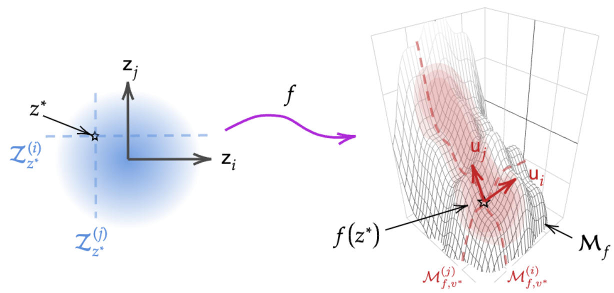

Figure 2: Translating linear independence to statistical independence (**non-linear $\boldsymbol{f}$**, orthogonal Jacobian $\boldsymbol{J}=\boldsymbol{USV}^\top$ (evaluated at z^*), manifold $\mathcal{M}_f \subseteq \mathcal{X}$): $\mathcal{Z}^{(i)}_{z^*} \subseteq \mathcal{Z}$ are lines passing through z^* by varying co-ordinate i in the standard basis (dashed blue). Each $\mathcal{Z}^{(i)}_{z^*}$ maps to a sub-manifold $\mathcal{M}^{(i)}_{f,z^*} \subseteq \mathcal{M}_f$ (dashed red) passing through $f(z^*)$ parallel to the local $\boldsymbol{U}$-basis (columns of $\boldsymbol{U}$: $\boldsymbol{u}_i, \boldsymbol{u}_j$, red). $\mathcal{Z}^{(i)}_{z^*}$ are statistically independent: $p(z^*) = \prod_i p(z^*_i)$. f induces push-forward distributions $p^{\#(i)}_{f,z^*}$ over $\mathcal{M}^{(i)}_{f,z^*}$ and for $x = f(z^*)$, the density over the manifold $p(x) = \prod_i p^{\#(i)}_{f,z^*}(x)$ factorises as a product of independent components.

red). Given $f^{(i)}_{z^*}$ is a restriction of f, $\frac{d}{dz_i} f^{(i)}_{z^*} = \frac{\partial x}{\partial z_i}$ is tangent to the manifold and sub-manifold $\mathcal{M}^{(i)}_{f,z^*}$ (solid red), and since $\frac{\partial x}{\partial z_i}$ are assumed orthogonal, sub-manifolds $\mathcal{M}^{(i)}_{f,z^*}$ are orthogonal at z^*. Considering how density is mapped (from dashed blue to dashed red), $f^{(i)}_{z^*}$ and the marginal $p(z_i)$ over its domain $\mathcal{Z}^{(i)}_{z^*}$ define a push-forward distribution $p^{\#(i)}_{f,z^*}(x) = \|\frac{\partial x}{\partial z_i}\|^{-1} p(z_i)$, over $x \in \mathcal{M}^{(i)}_{f,z^*}$, where $z = f^{-1}(x)$. Thus by Eq. 14, for $x \in \mathcal{M}_f$,

$$p^{\#}_f(x) \;=\; \prod_i p^{\#(i)}_{f,v^*}(x) \tag{15}$$

factorises into uni-variate distributions supported over sub-manifolds $\mathcal{M}^{(i)}_{f,z^*}$ that are mutually orthogonal where they meet. By orthogonality, traversing sub-manifold $\mathcal{M}^{(i)}_{f,z^*}$ corresponds to varying a single latent factor z_i, hence all other probability factors remain constant and $\{\mathcal{M}^{(i)}_{f,z^*}\}_i$ serve as a (non-linear) co-ordinate system in $\mathcal{X}$ corresponding to *statistically independent components*. □

Remark 4. *The independent basis of f is the standard basis of $\mathcal{Z}$ and $\|\frac{\partial x}{\partial z_i}\| = |s_i|$.*

Remark 5. *The density restricted to $\mathcal{Z}^{(i)}_{v^*}$ is a standard Gaussian $\mathcal{N}(z;0,1)$. The density $p^{\#(i)}_{f,v^*}(x) = |s_i|^{-1} p(f^{-1}(x)_i)$ over $\mathcal{M}^{(i)}_{f,z^*}$ is not Gaussian in general since s_i can vary arbitrarily over $x \in \mathcal{M}^{(i)}_{f,z^*}$.*

Corollary 2.1. *For data generated from a push-forward distribution $p^{\#}_f$, where $p(z)$ satisfies A#1 and f satisfies A#2 and A#3, a Gaussian VAE identifies statistically independent components of the data with distinct latent dimensions. If ground truth singular values s_i as a function of $z \in \mathcal{Z}^{(i)}_{v^*}$ are unique, the model is fully identified (the analogue of distinct singular values).*

Proof. The data distribution and an optimised Gaussian VAE each satisfy A#1, A#2 and (from §2) A#3, so by Theorem 2 data lie on a manifold with statistically independent sub-manifolds. The VAE defines a similar manifold and its objective is maximised *iff* $p_\theta(x) = p(x)$, hence when distributions over VAE sub-manifolds match those over ground truth sub-manifolds. VAE sub-manifolds map 1-to-1 to latent dimensions by Theorem 2. Identifiability follows from uniqueness of $p^{\#(i)}_{f,v^*}(x)$. □

4.1.3 NON-LINEAR f

Having seen that column orthogonality (A#3) is *sufficient* for independent factors to manifest in $\mathcal{X}$, we consider if it is *necessary*. We relax A#3 and consider a general push-forward distribution under A#1 and A#2 (independent latent variables and injective, differentiable f).

In the two previous scenarios (linear, column-orthogonal Jacobian), sub-manifolds in $\mathcal{Z}$ ($\mathcal{V}^{(i)}_{v^*}$, $\mathcal{Z}^{(i)}_{z^*}$) are linear, defined by the right singular vectors of the Jacobian, and *constant* $\forall z \in \mathcal{Z}$. Those sub-manifolds can also be defined parametrically, as continuous paths that follow right singular vectors at each point (*cf* integrating over a vector field).[5] In our now relaxed scenario, singular vectors can vary over $\mathcal{Z}$. Since the SVD of a matrix $\boldsymbol{M}$ is continuous w.r.t. $\boldsymbol{M}$ (Papadopoulo & Lourakis, 2000), if $\boldsymbol{J}$ is continuous w.r.t. z (by A#2) then right singular vectors $\boldsymbol{v}_i$ (as a function of z) trace continuous sub-manifolds $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ that are orthogonal everywhere. By definition of the SVD, mutually orthogonal $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ map to mutually orthogonal sub-manifolds $\mathcal{M}^{(i)}_f \subseteq \mathcal{M}_f$ (as previously), thus the push-forward distribution over the manifold $\mathcal{M}_f$ again factorises as a product of component push-forward distributions over each $\mathcal{M}^{(i)}_f$. However, sub-manifolds $\mathcal{V}^{(i)} \in \mathcal{Z}$ need not now be linear and *are not statistically independent in general*, i.e. the density at $z^* \in \mathcal{Z}$ may not equal the product of densities over $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ passing through z^*.

Thus, either: (1) sub-manifolds $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ are not statistically independent and $p(x)$ does not factorise as a product of *independent* components ("f *entangles* z_i"); or (2) sub-manifolds $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ are statistically independent (e.g. an arbitrary orthonormal basis) and map under f to independent factors in $\mathcal{X}$. In case (2), since an optimal Gaussian VAE maps independent components $\mathcal{M}^{(i)} \subseteq \mathcal{X}$ to the standard basis in $\mathcal{Z}$, independent factors in $\mathcal{X}$ are identified, but sub-manifolds $\mathcal{V}^{(i)} \subseteq \mathcal{Z}$ are *unidentifiable*, analogous to $\boldsymbol{V}$ in PCA.

4.2 Interpreting β of β-VAE

We now consider the role of β parameter in the β-VAE objective (Eq. 1), which is empirically observed to affect disentanglement (Higgins et al., 2017). Previous works interpret β as re-weighting the KL and reconstruction components of the ELBO, or serving as a Lagrange multiplier for a KL "constraint". We provide an interpretation more in keeping with the original ELBO.

To model data from a given domain, a (β-)VAE requires a suitable likelihood $p_\theta(x|z)$, e.g. a Gaussian likelihood for coloured images, and a Bernoulli for black and white images where pixel values $x^k \in [0,1]$ are bounded (Higgins et al., 2017) . In the Gaussian case, dividing Eq. 1 by β shows that training a β-VAE with encoder variance $\mathrm{Var}[x|z] = \sigma^2$ is *equivalent to a VAE* with $\mathrm{Var}[x|z] = \beta\sigma^2$ and adjusted learning rate (Lucas et al., 2019). We now interpret β for other likelihoods.

In the Bernoulli example mentioned above, black and white image pixels are not strictly black *or* white ($x^k \in \{0,1\}$) and may lie between ($x^k \in [0,1]$), hence the Bernoulli distribution appears invalid as it does not sum to 1 over the domain of x^k. That is, unless each sample is treated as the mean $\bar{x}$ of multiple (true) Bernoulli samples. Multiplying the likelihood by a factor $\kappa > 1$ is then tantamount to scaling the number of observations as though each were made κ times, lowering the variance of the "mean" observation, $\mathrm{Var}[\bar{x}] \overset{\kappa\to\infty}{\longrightarrow} 0$.[6] Thus, multiplying the KL term by β in a β-VAE, or equivalently dividing the likelihood by β, amounts to scaling the likelihood's variance by β, just as in the Gaussian case: **higher $\boldsymbol{\beta}$** corresponds to lower κ ("fewer observations") and so **higher likelihood variance**. Since the argument holds for any exponential family likelihood, we have proved

Theorem 3 (β-$\mathrm{VAE}_{\sigma^2} \equiv \mathrm{VAE}_{\beta\sigma^2}$)**.** *If the likelihood $p_\theta(x|z)$ is of exponential family form, a $\boldsymbol{\beta}$-VAE with $\mathrm{Var}[x|z] = \sigma^2$ is equivalent to a VAE with $\mathrm{Var}[x|z] = \beta\sigma^2$.*

In the most general case, the β-ELBO (Eq. 1) is maximised if $q(z|x) \propto p_\theta(x|z)^{1/\beta} p(z)$, and β can be interpreted as a *temperature* parameter: high β dilates the likelihood towards a uniform distribution (high $\mathrm{Var}[x|z]$), low β concentrates it towards a delta distribution (low $\mathrm{Var}[x|z]$).

Figure 3, from Rezende & Viola (2018), nicely illustrates the effect of varying β and empirically demonstrates the relationship to $\mathrm{Var}[x|z]$. As variance increases, posteriors of nearby data points $\{x_i\}$ (blue) increasingly overlap (by Eq. 6/9) and the decoder maps latents in regions of overlap to weighted averages of x_i (red). Since $\mathrm{Var}[x|z]$ governs how close data points need to be for this effect, it acts as a "glue" over $x \in \mathcal{X}$ (see caption for details).

In §3, we saw that optimising the ELBO encourages Jacobian orthogonality, on which disentanglement relies, in expectation over posteriors (Eq. 8). We conjecture that this justifies why increased β

[5]e.g. $v^{(i)}(t) \in \mathcal{V}^{(i)}_{v^*},\ v^{(i)}(t) = v^* + \int_0^t \frac{dv}{dt} dt = v^* + t\boldsymbol{v}_i$ (where $\frac{dv}{dt} = \frac{\partial v_i}{\partial z}(v(t)) = \boldsymbol{v}_i$)

[6]A mode-parameterised Beta distribution could also be considered, but we keep to a more general argument.

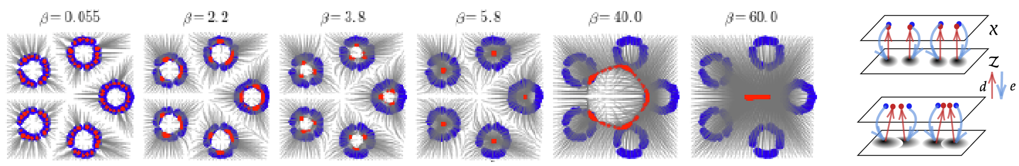

Figure 3: Effect of $\mathrm{Var}[x|z]$, or β, on reconstruction (blue = data, red = reconstruction): (*l*) For low β ($\beta = 0.55$), $\mathrm{Var}[x|z]$ is low, by Eq. 6 & 9, and data are closely reconstructed (see right, top). As β increases, $\mathrm{Var}[x|z]$ and so $\mathrm{Var}[z|x]$ increase and posteriors of nearby data points $\{x_i\}$ increasingly overlap (see right, bottom). For z in overlap of $\{q(z|x_i)\}$, the decoder $\mathbb{E}[x|z]$ maps to a weighted average of $\{x_i\}$. Initially, close neighbours map to their mean ($\beta = 2.2, 3.8$), then small circles "become neighbours" and map to their centroids, until finally all samples map to the global centroid ($\beta = 60$). (reproduced with permission from Rezende & Viola, 2018) (*r*) illustrating posterior overlap, (*t*) low β, (*b*) higher β.

enhances disentanglement (Higgins et al., 2017; Burgess et al., 2018): increasing β increases $\mathrm{Var}[x|z]$ and so $\mathrm{Var}[x|z]$ (Eq. 8), which (i) encourages orthogonality over a broader region of $\mathcal{Z}$; and (ii) increases posterior overlap where multiple orthogonality constraints apply simultaneously (Fig. 3).

We note that Theorem 3 also allows clearer interpretation of other works that vary β. While setting $\beta > 1$ can enhance disentanglement, setting $\beta < 1$ is found to mitigate "posterior collapse" (**PC**), which describes when a VAE's likelihood is sufficiently expressive such that it learns to directly model the data distribution, $p(x|z) = p(x)$, leaving latent variables redundant (Bowman et al., 2015).

Corollary 3.1 ($\beta < 1$). *Setting $\beta < 1$ is expected to mitigate posterior collapse.*

Proof. From Theorem 3, $\beta < 1$ reduces $\mathrm{Var}[x|z]$, constraining the distributional family that $p_\theta(x|z)$ can describe. For some β, $\mathrm{Var}[x|z] < \mathrm{Var}[x]$ and so $p(x) \neq p_\theta(x|z)$, $\forall \theta$, making PC impossible. ☐

5 Related Work

Many works study aspects or variants of VAEs, or disentanglement in other modelling paradigms. Here, we review those that offer insight into understanding the underlying cause of disentanglement in VAEs. Higgins et al. (2017) first showed that disentanglement is enhanced by increasing β in Eq. 1, and Burgess et al. (2018) hypothesised that diagonal posterior covariances may be the cause, encouraging latent dimensions to align with generative factors of the data. Rolinek et al. (2019) empirically showed and theoretically supported a link between diagonal posterior covariances and orthogonality in the decoder Jacobian, deemed responsible for disentanglement. Kumar & Poole (2020) simplified and generalised the argument. These works demonstrate that diagonal posteriors provide an inductive bias that breaks the rotational symmetry of an isometric Gaussian prior, side-stepping impossibility results related to independent component analysis (e.g. Locatello et al., 2019).

Several works investigate analytically tractable linear VAEs (Lucas et al., 2019; Bao et al., 2020; Koehler et al., 2022). Zietlow et al. (2021) show that disentanglement is sensitive to perturbations to the data distribution. Reizinger et al. (2022) relate the VAE objective to *independent causal mechanisms* (Gresele et al., 2021) which consider *non-statistically independent* sources that contribute to a mixing function by orthogonal columns of the Jacobian. This directly relates to the orthogonal Jacobian bias of VAEs, but differs to our approach that identifies statistically independent components/sources. Ramesh et al. (2018) trace independent factors by following leading left singular vectors of the Jacobian of a GAN generator. In the opposite direction, Chadebec & Allassonnière (2022) trace manifolds in latent space by following a locally averaged metric derived from VAE posterior co-variances. Pan et al. (2023) claim that the data manifold is identifiable from a geometric perspective assuming Jacobian-orthogonality, which differs to our focus on statistical independence. More recently, Bhowal et al. (2024) consider the encoder/decoder dissected into linear and non-linear aspects, loosely resembling our view of the Jacobian in terms of its SVD. However, the decoder function is quite different to its Jacobian and dissecting a function into linear and non-linear components is not well defined whereas an SVD is unique.

Recently, Buchholz et al. (2022) analysed several function classes identifiable by Independent Component Analysis (ICA), including conformal maps. This relates closely to our analysis of a decoder with column-orthogonal Jacobian (§4.1.2), which is a conformal map (see Def. 2, Buchholz et al., 2022). Conformal maps are proved to be identifiable in abstract via Moebius transforms, whereas we give a constructive proof for VAEs in terms of the SVD of the decoder's Jacobian. Combining these appears a promising direction to better understand the interplay between stochastic and deterministic approaches to learning latent generative factors.

6 CONCLUSION

Unsupervised disentanglement of independent factors of the data is a fundamental aim of machine learning and significant recent progress has been made in the case of VAEs. We extend that work by showing: (i) that the previously proposed approximate relationship can be defined precisely; and (ii) that the choice of diagonal posterior covariances in a VAE causes statistically independent components of the data to align with distinct latent variables of the model, i.e. disentanglement. In the process, we provide a novel yet straightforward interpretation of β in a β-VAE, which plausibly explains why increasing β promotes disentanglement but degrades generation quality; and why decreasing β mitigates posterior collapse. We also supplement the proof of orthogonality by showing that the likelihood's Hessian is necessarily encouraged to be diagonal and giving a detailed analysis of the Jacobian's optimal singular values.

Neural networks are often considered too complex to explain, yet recent advances make their deployment in everyday applications all but inevitable. Improved theoretical understanding is therefore essential to be able to confidently take full advantage of machine learning progress in non-trivial and potentially critical systems, and we believe that the body of work that we add to here is a useful step. Interestingly, our approach rests on the fact that, regardless of the model's complexity, its Jacobian, which transforms the density of the prior, can be considered in relatively simple terms.

Not only is a better understanding of VAEs of interest in itself, VAEs are often part of the pipeline in recent diffusion models that achieve state-of-the-art generative performance (e.g. Pandey et al., 2022; Yang et al., 2023; Zhang et al., 2022). Other recent works show that supervised learning (Dhuliawala et al., 2023) and self-supervised learning (Bizeul et al., 2024) can be viewed from a latent variable perspective and trained under a suitable variant of the ELBO, connecting VAEs to other learning paradigms in a common mathematical language.

One limitation of our work and of current understanding more generally is that disentanglement is observed in VAEs with non-Gaussian likelihoods (Higgins et al., 2017), whereas current work, including ours, focus predominantly on the Gaussian case. We plan to address this in future work.

REFERENCES

Xuchan Bao, James Lucas, Sushant Sachdeva, and Roger B Grosse. Regularized linear autoencoders recover the principal components, eventually. In *NeurIPS*, 2020.

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.

Pratik Bhowal, Achint Soni, and Sirisha Rambhatla. Why do variational autoencoders really promote disentanglement? In *ICML*, 2024.

Alice Bizeul, Bernhard Schölkopf, and Carl Allen. A probabilistic model to explain self-supervised representation learning. *arXiv preprint arXiv:2402.01399*, 2024.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Simon Buchholz, Michel Besserve, and Bernhard Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *NeurIPS*, 2022.

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

Clément Chadebec and Stéphanie Allassonnière. A geometric perspective on variational autoencoders. In *NeurIPS*, 2022.

Shehzaad Dhuliawala, Mrinmaya Sachan, and Carl Allen. Variational classification. *TMLR*, 2023.

Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *NeurIPS*, 2021.

Louay Hazami, Rayhane Mama, and Ragavan Thurairatnam. Efficientvdvae: Less is more. *arXiv preprint arXiv:2203.13751*, 2022.

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

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *ICML*, 2018.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Frederic Koehler, Viraj Mehta, Chenghui Zhou, and Andrej Risteski. Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias. In *ICLR*, 2022.

Abhishek Kumar and Ben Poole. On Implicit Regularization in β-VAEs. In *ICML*, 2020.

Daniel Kunin, Jonathan Bloom, Aleksandrina Goeva, and Cotton Seed. Loss landscapes of regularized linear autoencoders. In *ICML*, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

James Lucas, George Tucker, Roger B Grosse, and Mohammad Norouzi. Don't Blame the ELBO! a Linear VAE Perspective on Posterior Collapse. In *NeurIPS*, 2019.

Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural computation*, 21(3):786–792, 2009.

Ziqi Pan, Li Niu, and Liqing Zhang. Geometric inductive biases for identifiable unsupervised learning of disentangled representations. In *AAAI*, 2023.

Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. *TMLR*, 2022.

Théodore Papadopoulo and Manolis IA Lourakis. Estimating the jacobian of the singular value decomposition: Theory and applications. In *ECCV*, 2000.

Aditya Ramesh, Youngduck Choi, and Yann LeCun. A spectral regularizer for unsupervised disentanglement. *arXiv preprint arXiv:1812.01161*, 2018.

Patrik Reizinger, Luigi Gresele, Jack Brady, Julius Von Kügelgen, Dominik Zietlow, Bernhard Schölkopf, Georg Martius, Wieland Brendel, and Michel Besserve. Embrace the gap: Vaes perform independent mechanism analysis. In *NeurIPS*, 2022.

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

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

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational Autoencoders Pursue PCA Directions (by Accident). In *CVPR*, 2019.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*, 2014.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *ICLR*, 2019.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. Disdiff: unsupervised disentanglement of diffusion probabilistic models. In *NeurIPS*, 2023.

Zijian Zhang, Zhou Zhao, and Zhijie Lin. Unsupervised representation learning from pre-trained diffusion probabilistic models. *NeurIPS*, 2022.

Dominik Zietlow, Michal Rolinek, and Georg Martius. Demystifying inductive biases for (beta-) vae based architectures. In *ICML*, 2021.