Disentanglement as Identifiable Pushforward Factorisation

Carl Allen
Centre for Data Science, École Normale Supérieure, Paris
carl_a@me.com

Abstract

We consider disentanglement in generative pushforward models, such as VAEs and GANs. For a generator/decoder $g: \mathcal{Z} \to \mathcal{X}$ with factorised prior $p(z) = \prod_i p(z_i)$, we define *independent-factor disentanglement* as statistical independence expressed over the data manifold where the pushforward density $p_\mu = g_\# p$ factorises into one-dimensional "seam" factors, each controlled by a distinct latent coordinate. We prove that p_μ factorises according to the SVD of g's Jacobian; that this notion of disentanglement equates to two conditions on g (C1-C2); and that, under those conditions and suitable assumptions, the seam factors are *identifiable* up to permutation and sign. For Gaussian (β-)VAEs, an optimal covariance identity suggests an implicit preference for C1-C2, justifying why disentanglement arises modulated by β. Experiments illustrate this mechanism on Gaussian data, dSprites, and CelebA.

1 Introduction

A generative latent variable model is said to be *disentangled* if each latent co-ordinate controls a distinct semantic aspect of generated samples, e.g. the position of an object or a facial expression in an image, and is thus useful for controlled generation. Disentanglement is observed in Variational Autoencoders (**VAE**s, Kingma & Welling (2014); Rezende et al. (2014)) and variants (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018), despite them not being deliberately designed to achieve it. Related phenomena are found in Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and diffusion models (Rombach et al., 2022; Pandey et al., 2022; Yang et al., 2023).

Disentanglement has been formalised in several ways, typically as recovery or alignment of learned latents with ground-truth factors. In genuinely unsupervised settings, however, those factors are not observed; the relevant question then is where independent factors are encoded in the observed distribution. We answer this for smooth pushforwards: independence appears as a factorisation of the manifold density in co-ordinates over the manifold.

Research into VAEs attributes their disentanglement to *diagonal* posterior covariances, commonly chosen for computational efficiency (Higgins et al., 2017; Burgess et al., 2018; Rolinek et al., 2019; Kumar & Poole, 2020). Approximate relationships suggest that diagonal covariances promote *column-orthogonality* in the decoder's Jacobian, a property empirically associated with disentanglement (Gresele et al., 2021). In GANs, disentangled features have been identified via the SVD of the generator's Jacobian (Ramesh et al., 2018), or induced by regularising generator derivatives: Wei et al. (2021) regularise Jacobian orthogonality, while Peebles et al. (2020) penalise mixed Hessian interactions. We connect these findings via a general distribution-level explanation of disentanglement.

Our key results are *model-agnostic*, general to smooth pushforwards under a deterministic function g:

- a distributional definition of disentanglement in terms of independent factors/components (D1),
- a canonical factorisation of the pushforward density via the SVD of g's Jacobian (Lem. 4.1),
- an *iff* characterisation of disentanglement via conditions on g's derivatives (C1-C2, Thm. 4.2), and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

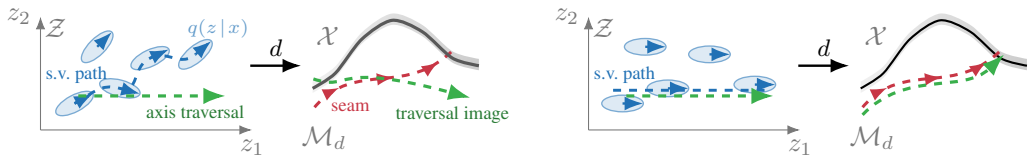


Figure 1: **Disentanglement illustrated for VAEs with full vs diagonal posteriors:** Right singular vectors $\boldsymbol{v}_i \in \mathcal{Z}$ of the decoder's Jacobian (blue arrows) define *singular vector (s.v.) paths* in latent space (dashed blue); left singular vectors $\boldsymbol{u}_i$ (red arrows) define *seams* in data space (dashed red). 1-D densities over seams factorise the manifold density p_μ. (***left***) For VAEs with full posteriors, s.v. paths are arbitrary, hence the image of an axis-traversal (green) need *not* follow a seam. (***right***) When the decoder satisfies C1-C2, as encouraged in VAEs with diagonal posteriors, paths are axis-aligned and seam factors are independent, i.e. *disentangled* (D1, §5).

- a proof that *independent factors are identifiable* up to natural symmetries (§6).

For Gaussian (β-)VAEs, an optimal covariance identity suggests an implicit preference, modulated by β, over decoders satisfying C1-C2 (§5). Overall, we extend understanding of disentanglement via an intuitive, distribution-level definition that formalises this well-known phenomenon and justifies well-known empirical findings.

2 Background

Notation: For data $x \in \mathcal{X} \doteq \mathbb{R}^m$ with density $p(x)$, and latent variables $z \in \mathcal{Z} \doteq \mathbb{R}^d$ ($d \leq m$), we consider functions $f : \mathcal{Z} \to \mathcal{X}$ that are continuous, injective and differentiable almost everywhere (***c.i.d.a.e.***), with image $\mathcal{M}_f = \{f(z) \mid z \in \mathcal{Z}\} \subseteq \mathcal{X}$. If f is differentiable at z, $\boldsymbol{J}_z$ denotes its Jacobian evaluated at z ($[\boldsymbol{J}_z]_{ij} = \frac{\partial x_i}{\partial z_j}$) with singular value decomposition (**SVD**) $\boldsymbol{J}_z = \boldsymbol{U}_z \boldsymbol{S}_z \boldsymbol{V}_z^\top$ ($\boldsymbol{U}_z^\top \boldsymbol{U}_z = \boldsymbol{I}$, $\boldsymbol{V}_z^\top \boldsymbol{V}_z = \boldsymbol{V}_z \boldsymbol{V}_z^\top = \boldsymbol{I}$).[1] Let $s_i \doteq \boldsymbol{S}_{ii}$ denote the i^{th} singular value, and $\boldsymbol{u}_i/\boldsymbol{v}_i$ the i^{th} left/right singular vectors (columns of $\boldsymbol{U}/\boldsymbol{V}$). Where $\boldsymbol{J}_z$ exists and has full column rank, $\mathcal{M}_f$ is locally a d-dimensional embedded manifold (see Fig. 5). By injectivity, f defines a bijection between $\mathcal{Z}$ and $\mathcal{M}_f$.

Latent Variable Model (LVM): We consider generative models with parameters θ and independent latent variables:

$$p_\theta(x) = \int_z p_\theta(x|z)p(z), \qquad p(z) = \prod_i p(z_i)\ . \tag{1}$$

The generative function $g : \mathcal{Z} \to \mathcal{X}$ is the map $g(z) = \mathbb{E}[x|z]$ from latent space to the *mean manifold* $\mathcal{M}_g = \{g(z)\} \subseteq \mathcal{X}$, with push-forward density $p_\mu \doteq g_\# p_z$ (defined relative to induced volume on the manifold). Eq. 1 admits the deterministic case $p_\theta(x|z) = \delta_{x - g(z)}$, where the data distribution is supported strictly on $\mathcal{M}_g$ with density p_μ (deterministic LVM).

Variational Autoencoder (VAE): A VAE learns parameters of an LVM by maximising the ELBO

$$\int_x p(x) \log p_\theta(x) \ \geq\ \int_x p(x) \int_z q_\phi(z|x) \Big(\log p_\theta(x|z) - \beta \log \tfrac{q_\phi(z|x)}{p(z)} \Big), \tag{2}$$

where $\beta = 1$. A *decoder* $d(z)$ parameterises likelihoods $p_\theta(x|z)$; an *encoder* parameterises Gaussian approximate posteriors $q_\phi(z|x) = \mathcal{N}(z; e(x), \Sigma_x)$, with *diagonal* covariances Σ_x. The prior $p(z)$ is standard Gaussian. We call a VAE with Gaussian likelihoods $p_\theta(x|z) \doteq \mathcal{N}(x; d(z), \sigma^2 \boldsymbol{I})$ a **Gaussian VAE**. A Gaussian VAE with a linear decoder $d(z) = \boldsymbol{D}z$, $\boldsymbol{D} \in \mathbb{R}^{m \times d}$, is called a **linear VAE (LVAE)**.

Generative Adversarial Network (GAN): A GAN learns a deterministic LVM parameterised by a *generator* network g, by minimising the Jensen-Shannon divergence between $p_\theta(x)$ and $p(x)$, as approximated via a classifier.

Disentanglement: Disentanglement commonly refers to assigning distinct generative factors to distinct latent variables, such that varying z_i changes generated samples in one feature while leaving others fixed (Bengio et al., 2013; Higgins et al., 2017; Ramesh et al., 2018; Rolinek et al., 2019). This is closely related to *Independent Component Analysis* (**ICA**), which seeks statistically independent components of data generated under Eq. 1. The issue is that ground-truth factors are typically unobserved; we therefore define disentanglement directly at the level of the pushforward density.

[1] To lighten notation, explicit dependence of $\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{S}, \boldsymbol{u}_i, \boldsymbol{v}_i, s_i$ on z is often suppressed where context is clear.

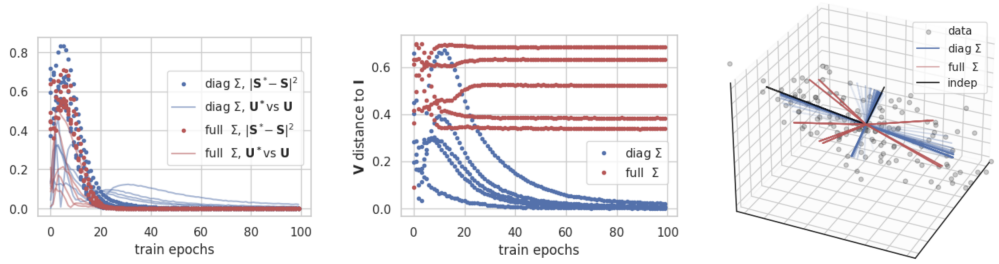


Figure 2: **An LVAE breaks rotational symmetry.** (***l***) both full and diagonal posterior VAEs fit the data, learning ground truth parameters $\boldsymbol{U}$, $\boldsymbol{S}$ (all losses$\to$0); (***c***) with diagonal Σ, right singular vectors $\boldsymbol{v}^i$ of $\boldsymbol{D}$ align with standard basis vectors $\boldsymbol{z}_i$ ($\boldsymbol{V}\to\boldsymbol{I}$, blue plots); (***r***) full-Σ VAEs map latent axes $\boldsymbol{z}_i$ to arbitrary directions (red), but diagonal-Σ VAEs learn (later epochs darker) to map $\boldsymbol{z}_i$ to *independent components* (blue $\to$ black).

Surprisingly, VAEs often exhibit disentanglement "for free", which is found to increase for $\beta>1$ albeit at a cost to generative quality, e.g. blurrier images (***β*-VAE**, Higgins et al., 2017). A standard GAN does not automatically disentangle, but has been induced to by diagonalising generator derivatives (Peebles et al., 2020; Wei et al., 2021). We theoretically justify and unify these observations.

Linear LVM: Probabilistic PCA (PPCA) (Tipping & Bishop, 1999) studies the linear case

$$p(x|z)=\mathcal{N}(x;\boldsymbol{W}z,\sigma^2\boldsymbol{I}),\qquad p(z)=\mathcal{N}(z;\,\boldsymbol{0},\boldsymbol{I})\,, \tag{3}$$

$\boldsymbol{W}\in\mathbb{R}^{m\times d}$, $\sigma\in\mathbb{R}$.[2] Exact posteriors $p(z|x)$ and MLE parameters $\boldsymbol{W}_*$ are known analytically:

$$p(z|x)=\mathcal{N}(z;\tfrac{1}{\sigma^2}\boldsymbol{M}\boldsymbol{W}^\top x,\boldsymbol{M}),\;\; \boldsymbol{M}^{-1}=\boldsymbol{I}+\tfrac{1}{\sigma^2}\boldsymbol{W}^\top\boldsymbol{W}\qquad \boldsymbol{W}_*=\boldsymbol{U}_{\boldsymbol{X}}(\boldsymbol{\Lambda}_{\boldsymbol{X}}-\sigma^2\boldsymbol{I})^{1/2}\boldsymbol{R} \tag{4}$$

where $\boldsymbol{\Lambda}_{\boldsymbol{X}}\in\mathbb{R}^{d\times d}$, $\boldsymbol{U}_{\boldsymbol{X}}\in\mathbb{R}^{m\times d}$ contain the largest eigenvalues and respective eigenvectors of 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 $\sigma^2\to 0$, $\boldsymbol{W}_*$ approaches the SVD of the data matrix $\boldsymbol{X}=\boldsymbol{U}_{\boldsymbol{X}}\boldsymbol{\Lambda}_{\boldsymbol{X}}^{1/2}\boldsymbol{V}_{\boldsymbol{X}}^\top\in\mathbb{R}^{m\times n}$, up to $\boldsymbol{V}_{\boldsymbol{X}}$ (i.e. classical PCA). The model is considered *unidentified* since $\boldsymbol{W}_*$ has uncountably infinite solutions (due to $\boldsymbol{R}$).

Given the same linear LVM, an LVAE approximates the solution by maximising the ELBO with $p_\theta(x|z)=\mathcal{N}(x;\boldsymbol{D}z,\sigma^2\boldsymbol{I})$ and $q_\phi(z|x)=\mathcal{N}(z;\boldsymbol{E}x,\Sigma)$. Interestingly, an LVAE with *diagonal* Σ breaks the rotational symmetry of PPCA (Lucas et al., 2019). From Eq. 4 we have

$$\Sigma_*\overset{(4i)}{=}(\boldsymbol{I}+\tfrac{1}{\sigma^2}\boldsymbol{W}_*^\top\boldsymbol{W}_*)^{-1}\overset{(4ii)}{=}\sigma^2\boldsymbol{R}^\top\boldsymbol{\Lambda}_{\boldsymbol{X}}^{-1}\boldsymbol{R}\,, \tag{5}$$

so for Σ to be both optimal *and* diagonal, $\boldsymbol{R}$ must belong to a finite set of signed permutations, thus the optimal decoder $\boldsymbol{D}_*=\boldsymbol{U}_{\boldsymbol{X}}(\boldsymbol{\Lambda}_{\boldsymbol{X}}-\sigma^2\boldsymbol{I})^{1/2}$ is unique *up to permutation and sign* (see Fig. 2). We claim that this result of diagonal covariances is in fact *disentanglement* in the linear case.

3 Defining Disentanglement

In Euclidean co-ordinates, $p(x)=\prod_i p_i(x_i)$ if and only if random variables x_i are independent. For generated images, however, the relevant co-ordinates are generally not pixels but intrinsic co-ordinates on the generated manifold $\mathcal{M}_g$. We define disentanglement as this standard independence property over $\mathcal{M}_g$: the manifold density factorises into independent one-dimensional factors, each obtained by varying one latent co-ordinate.

Definition D 1 (Independent-factor disentanglement). For $g:\mathcal{Z}\to\mathcal{X}$ *c.i.d.a.e.*, we say p_μ is *disentangled with respect to the latent co-ordinates* if, on $\mathcal{Z}_{\text{reg}}$, there exist 1-D densities $\{f_i\}$ such that

$$p_\mu\big(g(z)\big)=\prod_{i=1}^{d}f_i\big(u_i(z)\big)\,, \tag{6}$$

where: (a) each factor f_i is the *1-D push-forward* of $p(z_i)$ along the line obtained by moving in the i^{th} latent coordinate z_i (with all others fixed); (b) u_i is the co-ordinate of $g(z)$ along the image of that line; and (c) for $z\sim p(z)$, the random variables $\{u_i(z)\}$ are mutually independent.

[2]For simplicity we assume that data is centred, equivalent to including a mean parameter (Tipping & Bishop, 1999).

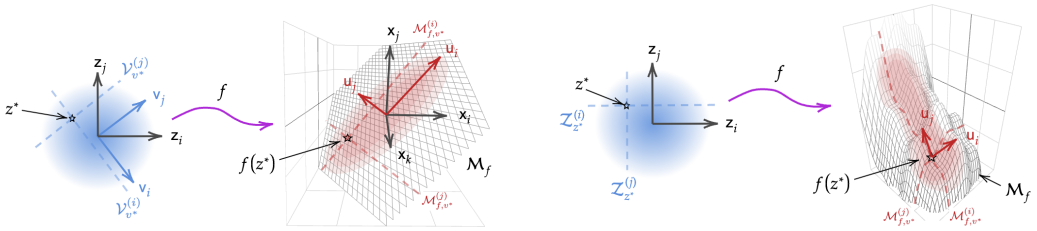


Figure 3: **Pushing forward $p(z)$, from singular vector paths to seams**: 1-D marginals $p_i(z_i)$ over *s.v. paths* $\mathcal{V}^i_{z^*}$ (dashed blue) factorise $p(z)$; and push-forward to 1-D seam densities over *seams* $\mathcal{M}^i_{f,z^*}$ (dashed red) that factorise p_μ (Lemma 4.1). (***l***) For linear f *without* **C1** (e.g. full-Σ_x LVAE), $\mathcal{V}^i_{z^*}$ are straight lines but need not axis-align (as observed in Fig. 2(*r*)). (***r***) For c.i.d.a.e. f satisfying **C1-C2**, $\mathcal{V}^i_{z^*}$ are axis-aligned (by C1) and seam densities are independent components (by C2) that factorise p_μ, as required for disentanglement (D 1).

D1 requires independent factors and independent latent control, which need not be human-semantic labels, such as colour or shape. It differs from the group-action definition of **?**, which formalises independent transformations without requiring statistical independence.

We illustrate this density-level definition in the linear case of an LVAE ($d(z) = \boldsymbol{D}z$) with diagonal posteriors fitted to data from a linear LVM where $p_\mu = \mathcal{N}(0, \boldsymbol{W}\boldsymbol{W}^\top)$ ($\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$). We claim that an optimal decoder $\boldsymbol{D}_* = \boldsymbol{U}\boldsymbol{S}$ (see §2) disentangles p_μ per Definition D1.

Being Gaussian, p_μ factorises as a product of 1-D Gaussians $p_\mu(x) = \prod_i \mathcal{N}(u_i; 0, s_i^2)$ along eigenvectors $\boldsymbol{u}_i$ of its covariance $\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{U}\boldsymbol{S}^2\boldsymbol{U}^\top$, with co-ordinates $u_i \doteq \boldsymbol{u}_i^\top x \in \mathbb{R}$. If $x = \boldsymbol{D}_* z$, then $u_i = \boldsymbol{u}_i^\top \boldsymbol{D}_* z = s_i z_i$. Thus p_μ factorises as a product of independent 1-D push-forward densities $f_i = \mathcal{N}(0, s_i^2)$ (as in Eq. 6); and the decoder maps axis-aligned directions z_i to corresponding co-ordinates u_i with density f_i – meeting all requirements of D 1. As expected, samples generated by re-sampling z_i, holding $z_{j \neq i}$ constant, differ in a single component/feature u_i. The role of *diagonal* posteriors is discussed in Appendix A.

4 Disentanglement ⇔ Decoder Constraints

In the linear case, disentanglement arises by constraining right singular vectors of the generative function $\boldsymbol{D}_* = \boldsymbol{U}\boldsymbol{S}$. In the *non-linear* case, we must consider the generative function's Jacobian.

For the Jacobian $\boldsymbol{J}_z = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, singular vectors $\boldsymbol{v}_i$, $\boldsymbol{u}_i$ (columns of $\boldsymbol{V}$, $\boldsymbol{U}$) respectively define local orthonormal bases for $\mathcal{Z}$ at z (**$\boldsymbol{V}$-basis**) and the tangent space to $\mathcal{M}_g$ at $x = g(z)$ (**$\boldsymbol{U}$-basis**). At each point, $\boldsymbol{J}_z \boldsymbol{v}_i(z) = s_i(z)\boldsymbol{u}_i(z)$: a small displacement along a right singular vector maps to a displacement along the corresponding left singular vector, scaled by s_i. The frames may vary with z.

Singular Vector Paths and Seams: Directional derivatives $\boldsymbol{J}_z \boldsymbol{v}^i = s^i \boldsymbol{u}^i$ show that a small perturbation in direction $\boldsymbol{v}^i$ at $z \in \mathcal{Z}$ maps to a small perturbation in direction $\boldsymbol{u}^i$ at $g(z) \in \mathcal{M}_g$. By extension, if a path in $\mathcal{Z}$ follows $\boldsymbol{v}^i$ at each point (as a vector field) its image is a path following $\boldsymbol{u}^i$ over $\mathcal{M}_g$. For such paths to be well defined, singular values must be distinct and their order must be fixed.

Definition D 2 (Regular set and continuous SVD). For *c.i.d.a.e.* $g : \mathcal{Z} \to \mathcal{X}$, the *regular set* is defined as $\mathcal{Z}_{\text{reg}} \doteq \{z \in \mathcal{Z} \mid \boldsymbol{J}_z \text{ exists, has full column rank, and } s_1(z) > \cdots > s_d(z) > 0\}$. Vector fields $z \mapsto \boldsymbol{v}_i(z)$, $z \mapsto \boldsymbol{u}_i(z)$ and singular values $z \mapsto s_i(z)$ can be made continuous on each connected component of $\mathcal{Z}_{\text{reg}}$ by fixing the SVD $\boldsymbol{J}_z = \boldsymbol{U}_z \boldsymbol{S}_z \boldsymbol{V}_z^\top$.[3,4]

Definition D 3 (Singular-vector path). For $z^* \in \mathcal{Z}_{\text{reg}}$, the i-th singular-vector path is the curve z_t^i satisfying $z_0^i = z^*$ and $\dot{z}_t^i = \boldsymbol{v}_i(z_t^i)$ on its maximal regular interval $I_{z^*,i}$. Write $\mathcal{V}^i_{z^*} = \{z_t^i : t \in I_{z^*,i}\}$. Reversing the vector-field sign reverses orientation.

Definition D 4 (Seam). The corresponding seam is $x_t^i = g(z_t^i)$, with image $\mathcal{M}^i_{g,z^*}$. The chain rule gives $\dot{x}_t^i = s_i(z_t^i)\boldsymbol{u}_i(z_t^i)$. Its signed arc-length coordinate is $u_i(t) = \int_0^t s_i(z_\tau^i)\, d\tau$, so $\dot{u}_i = s_i > 0$.

[3] e.g. start from an arbitrary point, order singular values strictly decreasing and choose signs continuously.

[4] Restricting to $\mathcal{Z}_{\text{reg}}$ only excludes points where $\boldsymbol{J}_z$ is undefined or has repeated singular values; these edge cases can be avoided without affecting the results. All statements are made on a fixed connected component of $\mathcal{Z}_{\text{reg}}$.

Notice that if s.v. paths follow latent axes, and seam densities are independent factors of p_μ, we have disentanglement (D 1). We thus state two conditions:

C1: Right singular vectors of $\boldsymbol{J}_z$ are standard basis vectors, i.e. after axis relabeling/sign flips, $\boldsymbol{V}_z = \boldsymbol{I}$.

C2: The matrix of partial derivatives of singular values $(\frac{\partial s_i}{\partial z_j})_{i,j}$ is diagonal, i.e. $\frac{\partial s_i}{\partial z_j} = 0$ for all $i \neq j$.

Lemma 4.1 (Factorisation over seams). *Let $g : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e. *and the prior factorise as $p(z) = \prod_{i=1}^{d} p_i(z_i)$. Then, the manifold density p_μ on $\mathcal{M}_g$ factorises as*

$$p_\mu\big(g(z)\big) \;=\; \prod_{i=1}^{d} \tfrac{p_i(z_i)}{s_i(z)}\,, \quad \textit{for every } z \in \mathcal{Z}_{\mathrm{reg}}. \tag{7}$$

Moreover, under C1, each factor $\frac{p_i(z_i)}{s^i(z)}$ is the 1-D density over the i-th seam $\mathcal{M}^i_{g,z}$ at $x = g(z)$, obtained by pushing forward the 1-D marginal $p_i(z_i)$ over the i-th s.v. path $\mathcal{V}^i_z$ through z parallel to the i^{th} latent axis.

Proof. See Appendix B (by standard change-of-variables).

Lemma 4.1 states that p_μ admits a canonical product decomposition, which, under C1, are 1-D densities over seams, each the push–forward of the marginal over an *axis-aligned* latent s.v. path. This is not yet disentanglement: the factors become independent seam factors only when C2 removes cross-dependence of $s_i(z)$ on $z_{j\neq i}$.

Theorem 4.2 (Disentanglement $\Leftrightarrow$ C1-C2). *Let $g : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e. *and the prior factorise as $p(z) = \prod_{i=1}^{d} p_i(z_i)$. The push-forward density p_μ on the manifold $\mathcal{M}_g$ is* disentangled *(D 1) if and only if g satisfies **C1** and **C2** a.e..*

Proof. See Appendix C.

Thus C1-C2 are *precisely* the conditions by which the canonical pointwise factorisation (Eq. 7) becomes disentanglement, bringing axis-alignment and factor independence, respectively (see Fig. 3).

Geometric interpretation: Under the ambient Euclidean metric in $\mathcal{X}$, the pullback metric in latent space is $G = \boldsymbol{J}_z^\top \boldsymbol{J}_z$; and Eq. 7 divides the prior density by the induced volume factor $\sqrt{\det G}$. Under C1-C2, $G = \mathrm{diag}(s_i(z_i)^2)$ and we can define co-ordinates over the manifold

$$u_i = \int_{z_i^0}^{z_i} s_i(t)\, dt, \qquad \sum_i s_i(z_i)^2\, dz_i^2 = \sum_i du_i^2. \tag{8}$$

Thus the metric in u is the identity and the regular region is intrinsically flat, although its image may bend extrinsically in ambient space. As such, C1-C2 cannot hold throughout an intrinsically curved region, such as a sphere patch (extending the theory to such geometries is future work). We note that flatness alone implies neither C1-C2 nor a factorised density.

5 How Diagonal VAE Posteriors Encourage C1-C2

Having defined disentanglement and characterised it in terms of conditions on the generative function, we now consider why it arises in Gaussian VAEs with diagonal posteriors. Prior work attributes VAE disentanglement to diagonal posteriors via approximate relationships (Rolinek et al., 2019; Kumar & Poole, 2020, Eq. 11), which can be stated *exactly* from the Price/Bonnet Theorem (e.g. Opper & Archambeau, 2009). For a given decoder, an (unrestricted) optimal Gaussian covariance satisfies

$$\Sigma_x^{-1} \;=\; \boldsymbol{I} - \tfrac{1}{\beta}\mathbb{E}_{q(z|x)}[\mathbf{L}_z(x)] \;\overset{*}{=}\; \boldsymbol{I} + \tfrac{1}{\beta\sigma^2}\mathbb{E}_{q(z|x)}[\boldsymbol{J}_z^\top \boldsymbol{J}_z - (x - d(z))^\top \mathbf{H}_z]\,, \tag{9}$$

where $\mathbf{L}_z(x) = \nabla_z^2 \log p_\theta(x|z)$ is the log likelihood Hessian; $\boldsymbol{J}_z \doteq \frac{dx}{dz}$ and $\mathbf{H}_z \doteq \frac{d^2x}{dz^2}$ are the Jacobian and Hessian of the decoder (all evaluated at $z \in \mathcal{Z}$); and (*) assumes a Gaussian VAE. It can be seen that Eq. 9 generalises the linear result ($\boldsymbol{M}^{-1}$, Eq. 4) and relates $\sigma^2 \doteq \mathrm{Var}[x|z]$ to $\Sigma_x \doteq \mathrm{Var}[z|x]$, showing that (un)certainty in x and z go hand in hand.

Notably, Eq. 9 implies that a diagonal posterior (LHS) can attain the unrestricted optimum (RHS) only if the latter is itself diagonal. Hence, under joint training, diagonal posteriors encourage decoders with diagonalised derivatives (in aggregate/expectation) and thus do not guarantee disentanglement, but bias trained decoders toward the derivative structure that Thm. 4.2 shows is equivalent to D 1 disentanglement.

Property P1. For fixed x and z, matrices $\boldsymbol{J}_z^\top \boldsymbol{J}_z$ and $(x - d(z))^\top \mathbf{H}_z$ (as in Eq. 9) are each diagonal.

Lemma 5.1 (**VAE Disentanglement**). *If a trained Gaussian VAE with diagonal posteriors induces **P1** for z in a neighbourhood of $e(x)$, then it induces **C1** on that neighbourhood and **C2** to a first order approximation (Appendix D bounds the residual Taylor error).*

Proof. See Appendix D. (C1 follows from SVD of $\boldsymbol{J}_z$; C2 follows since directions $r(z) = x - d(z) \in \mathcal{X}$ of the directed Hessian are, to a first order approximation, tangent to $\mathcal{M}_d$.)

Thus, **Eq. 9 suggests a mechanism for disentanglement in VAEs**: diagonal posteriors encourage diagonality of decoder derivatives on posterior-supported regions, yielding **C1**, and **C2** to first order, biasing the push-forward density toward disentanglement. The expectation, posterior-support and local-approximation gaps are also one plausible source of reported variability (Locatello et al., 2019).

Noting that disentanglement is encouraged specifically over posterior supports, it can be seen that β implicitly affects the variance of such posteriors (explicitly so for Gaussian VAEs), whereby *higher β widens posteriors* (see Appendix E). Thus, **Eq. 9 suggests why increasing β enhances disentanglement** (Higgins et al., 2017; Burgess et al., 2018): it broadens where disentanglement is encouraged and increases posterior overlap where simultaneous constraints apply (see Fig. 6).

6 Identifiability

We now investigate whether a model capable of fitting data generated under the model class, will learn the *true* generative factors, up to a particular symmetry say, or if it might find a spurious factorisation.

Theorem 6.1 (**LVAE Identifiability**). *Let data be generated under the linear Gaussian LVM (Eq. 3) with ground-truth $g(z) = \boldsymbol{W}z$, $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top \in \mathbb{R}^{m \times d}$ of full column rank and* distinct *singular values. Let an LVAE with* diagonal *posteriors attain a global population ELBO optimum such that its learned parameters yield $p_\mu^{(d)} \equiv p_\mu^{(g)}$ on the mean manifold. Then the LVAE achieves disentanglement (D1) and identifies ground-truth independent components on $\mathcal{M}_g$ up to permutation and sign.*

Proof. See Appendix F.1 (Follows SVD uniqueness). □

Thus, if an LVAE fits the data, latent axes identify the ground truth independent factors.

Remark 6.2 (*$\boldsymbol{V}$ immaterial*). Ground-truth right singular vectors $\boldsymbol{V}$ (of $\boldsymbol{W}$) are not recoverable from $p(x)$ under the PPCA/LVAE model; this *is not* a lack of identification. With a standard Gaussian prior, any orthonormal change of basis of z preserves independence and leaves $p(x)$ unchanged. The only data-relevant object is $\boldsymbol{U}\boldsymbol{S}$; the arbitrary basis ($\boldsymbol{V}$) in which $\boldsymbol{W}$ was written has no bearing on $p(x)$.[5]

To generalise the linear result, we first isolate the part that depends only on the manifold density itself: if p_μ admits a seam factorisation in local orthonormal seam coordinates, then those seams are *intrinsic* to p_μ and unique (up to permutation and sign). We then show that any matching pushforward satisfying C1–C2 necessarily recovers those seams. (All proofs in Appendix F.2)

Lemma 6.3 (**Seams are Intrinsic**). *Let $\mathcal{M} \subseteq \mathcal{X}$ carry a manifold density p_μ. Assume that on a regular set $\mathcal{M}_{\text{reg}}$ there exist local coordinates $u = (u_1, \ldots, u_d)$ (see Eq. 8), whose coordinate lines are 1-D seams and directions form an orthonormal basis of $T_x\mathcal{M}$ at each $x \in \mathcal{M}_{\text{reg}}$, together with 1-D densities $\{f_i\}$ such that*

$$p_\mu(x) \;=\; \prod_{i=1}^{d} f_i\big(u_i(x)\big), \qquad x \in \mathcal{M}_{\text{reg}}. \tag{10}$$

Let $\mathbf{H}_x \doteq \nabla^2_{\mathcal{M}} \log p_\mu(x)$ denote the intrinsic Hessian on the manifold, and assume its eigenvalues are pairwise distinct a.e.. Then wherever the eigenvalues are distinct, the d seam directions are determined intrinsically by p_μ, as eigenvectors of $\mathbf{H}_x$, unique up to permutation and sign.

Lemma 6.4 (**A matching pushforward finds seams**). *Under assumptions of Lemma 6.3, let $d : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e. *with factorised prior $p(z) = \prod_i p_i(z_i)$ and push-forward density $p_\mu^{(d)} \equiv p_\mu$* matching *on $\mathcal{M}_d \doteq \{d(z)\} = \mathcal{M}$. If d satisfies **C1–C2** a.e., then for any z and $x = d(z)$:*

- *left singular vectors $\boldsymbol{U}_z$ of $\boldsymbol{J}_z$ coincide with the intrinsic seam directions in Lemma 6.3 (up to permutation and sign);*

[5] This is analogous to the difference between a linear transformation and its matrix representation, requiring a specific basis.

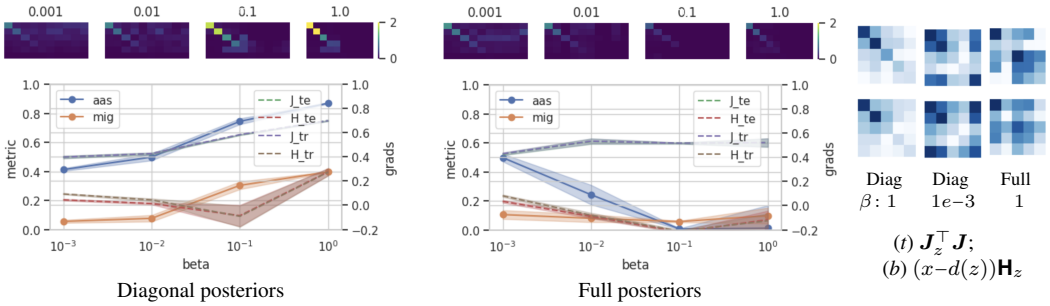


Figure 4: **Diagonal vs Full Posteriors**: (***left***) (*bottom*) disentanglement metrics and estimated diagonality of Eq. 9 terms (see Appendix G). With diagonal posteriors, disentanglement and diagonality are correlated (supporting P1), relative to full posteriors; (***top***) heatmaps of mutual information between model latents and ground truth factors. (***right***) derivatives in Eq. 9. terms are less diagonal for lower beta or full posteriors. (All results averaged over multiple runs)

- *the images under d of singular-vector paths (D3) are exactly the intrinsic seams through x;*
- *along the i-th intrinsic seam, the factor f_i is the 1-D push-forward of $p_i(z_i)$ (as in Lemma 4.1).*

Theorem 6.5 (Gaussian VAE Identifiability). *Let data be generated by* c.i.d.a.e. $g : \mathcal{Z} \to \mathcal{X}$ *with standard Gaussian prior. Let a Gaussian VAE with* diagonal *posteriors learn a decoder $d : \mathcal{Z} \to \mathcal{X}$. Suppose both g and d satisfy **C1–C2** and manifold densities match: $p_\mu^{(d)} \equiv p_\mu^{(g)}$ on $\mathcal{M} = \{g(z)\} = \{d(z)\}$. If the eigenvalues of the intrinsic log-density Hessian $\nabla^2_{\mathcal{M}} \log p_\mu$ are distinct a.e., then d identifies ground-truth independent components on $\mathcal{M}_g$, up to permutation and sign.*

The theorem assumes the product-seam structure exists and that the density Hessian selects its axes, not that those are known. Lemma 6.4 proves frame alignment under C1-C2, whereas metric and density-Hessian frames differ in general. We note that these exact statements are model-agnostic and do not require the VAE mechanism in §5.

Remark 6.6. Permutation and sign symmetry is optimal since singular vectors $\boldsymbol{u}_i$ have no inherent order/orientation.

We can also consider fitting a Gaussian VAE to data sampled from the push-forward of other priors.

Corollary 6.7 (Blind source separation). *In Theorem 6.5, if priors $p^{(g)}(z)$ and $p^{(d)}(z)$ factorise and $p_\mu^{(g)} \equiv p_\mu^{(d)}$ with C1-C2 holding* a.e.*, then the seam decomposition p_μ on $\mathcal{M}$ is unique, up to permutation and sign, and g and $p^{(g)}(z)$ are recoverable up to an axis-aligned diffeomorphism $\phi \doteq g^{-1} \circ d : \mathcal{Z} \to \mathcal{Z}$.*

Proof. Immediate from first part of Theorem 6.5 proof, which does not depend on the form of $p(z)$. The diffeomorphism follows since $\mathcal{M}_g = \mathcal{M}_d$ and by injectivity of d, g. □

Remark 6.8 (*ICA unidentifiability*). Classical nonlinear ICA aims to identify ground truth factors of data generated by the LVM in Eq. 1, which is impossible in general due to density-preserving transformations, not only Gaussian rotations (Hyvärinen & Pajunen, 1999). Positive results require additional restrictions (e.g. Horan et al., 2021; Buchholz et al., 2022). We show a further positive result: where the generative function satisfies C1-C2 and the intrinsic log-density Hessian has simple spectrum, the seam frame is identifiable.

In summary, we have defined independent-factor disentanglement, shown that it holds if and only if C1-C2 hold, and that, under the stated density-Hessian spectrum assumption, seam factors are unique and hence identifiable up to natural symmetry.

7 Empirical Support

The empirical question is whether the VAE mechanism above pushes trained decoders toward C1-C2, and whether this structure aligns with ground truth factors when known. From our analysis, we test whether (1) diagonal posteriors promote diagonalised decoder derivatives, especially $J_z^\top J_z$; (2) derivative diagonality correlates with benchmark factor recovery; and (3) in the linear case, diagonal posteriors break the rotational symmetry exactly.

Both (1) and (2) are clearly illustrated in the linear case where ground truth factors are known analytically. Fig. 2 shows results for diagonal and full covariance LVAEs learning Gaussian parameters (see caption for details). (*left*) Both models learn optimal parameters as expected. (*centre*) However, only diagonal covariances cause right singular vectors $\boldsymbol{V}$ of $\boldsymbol{J}_z$ to converge to standard basis vectors, i.e. $\boldsymbol{V} \to \boldsymbol{I}$, inducing **C1** (prediction 2). (*right*) Thus latent traversals map to independent components along left singular vectors $\boldsymbol{u}^i$, yielding disentanglement (D1) (prediction 1). This provides compelling evidence that diagonal covariances "break the rotational symmetry" of a Gaussian prior.

Though simplified, the linear case is fundamental since, from our analysis, disentanglement under general pushforward functions follows a similar rationale. Interestingly, Fig. 2 shows that learning parameters of a disentangled model is notably slower (*left*), due to the rate $\boldsymbol{V} \to \boldsymbol{I}$ (*centre*). A diagonal covariance model must find one of a finite subset of the infinite solutions set of a full covariance model.

Various studies show empirical support for (1) and (2). At fixed β, Rolinek et al. (2019, Fig. 4) report a correlation between decoder orthogonality and disentanglement across independent runs trained for different numbers of epochs, supporting the predicted relationship. They also show that columns of the decoder's Jacobian are more orthogonal (i.e. $\boldsymbol{V} \to \boldsymbol{I}$, **C1**) in VAEs with diagonal posteriors than those with full posteriors (Fig. 7), and that diagonality correlates with disentanglement. Supporting (1), Kumar & Poole (2020) show that *directly* inducing column-orthogonality in the decoder Jacobian promotes disentanglement.

Complementary evidence comes from GANs, where the same decoder-side picture appears without any posterior model. Ramesh et al. (2018) trace locally disentangled directions by following Jacobian singular vectors of a GAN generator. Wei et al. (2021) directly regularise orthogonality of Jacobian-induced output variations across latent co-ordinates, closely aligned with diagonalising $J_z^\top J_z$, hence with C1. Peebles et al. (2020) directly penalise mixed second derivatives of the generator, a Hessian-side regulariser closely related to the mechanism in Eq. 9, hence to the emergence of C2. Together these results support the view that disentanglement is controlled by derivative structure of a general pushforward map, not specific to the VAE objective.

dSprites: Further to prior evidence, we train diagonal and full posterior Gaussian VAEs ($d = 10$) on the *dSprites* dataset with 5 known generative factors (results averaged over 5 runs). How well each latent co-ordinate identifies a ground truth factor can be estimated explicitly from their *mutual information*, and a function of mutual information is often used to evaluate a model's overall disentanglement, e.g. the *mutual information gap* (MIG, Chen et al. (2018)).

Fig. 4 (*main plots*) reports MIG, AAS, an auxiliary score measuring closeness of the latent-factor mutual-information matrix to a permuted diagonal form, and derivative diagonality estimates plotted against β (see Appendix G for details). For diagonal posterior VAEs (*left*), disentanglement and diagonality both broadly increase with β. For full posteriors (*right*) no clear trend is observed. The heatmaps (top, aligned with main plots by β) show mutual information between each latent co-ordinate and ground truth factor (ordered greedily to put highest scores along the diagonal). We see that for diagonal posterior VAEs, disentanglement increases with β and that individual latent co-ordinates (horizontal) correlate with distinct ground truth factors (vertical), whereas that trend is not observed for full posteriors. For illustration, Fig. 4 (right) shows heatmaps of the $d \times d$ derivative terms in Eq. 9 ($\boldsymbol{J}_z^\top \boldsymbol{J}_z$ term top, Hessian term bottom), each for diagonal covariances, $\beta = 1$ (*l*); diagonal covariances, $\beta = 0.001$ (*c*); and full covariances, $\beta = 1$ (*r*) (each averaged over a batch).

We see that diagonalisation of the Jacobian term is more evident than that of the Hessian, consistent with our analysis of Eq. 9 vs disentanglement conditions C1-C2 (§5).

CelebA: We include comparable analysis on the CelebA dataset of natural face images in Appendix I.

Varying $\boldsymbol{\beta}$: Our understanding of the interplay between β and disentanglement is that higher β implies higher expected noise, "blurring" reconstructions, but enhancing disentanglement; while lower β (less assumed noise) tightens reconstructions but limits disentanglement to concentrated, potentially disconnected regions of $\mathcal{Z}$ (§5). This suggests that the heuristic of starting with high β and reducing it during training may retain axis alignment while sharpening reconstructions. We run experiments on the dSprites dataset and report results in Appendix J for constant β baselines (1, 0.001), and exponentially reducing β ($1 \to 0.001$) over training. As predicted, the results show that annealing β gives both sharp reconstructions and good disentanglement. We note the resemblance of

this to "de-noising" in autoencoders and diffusion models, suggesting that dynamically varying β is an interesting, potentially principled, direction for future research.

8 Related Work

Higgins et al. (2017) first showed that disentanglement in VAEs is enhanced by setting $\beta > 1$ in the ELBO (Eq. 2). Burgess et al. (2018) conjectured that diagonal posterior covariances may cause disentanglement. Rolinek et al. (2019) showed supporting empirical evidence (Fig. 5) and derived an approximate relationship between diagonal posteriors and Jacobian orthogonality, conjectured to cause disentanglement. Kumar & Poole (2020) generalised the argument, reaching an approximation to the identity in Eq. 9. We give the unrestricted covariance identity in Eq. 9 and characterise independent-factor disentanglement by C1-C2. Their connection in a trained VAE remains an implicit, conditional mechanism.

Lucas et al. (2019); Bao et al. (2020) and Koehler et al. (2022) study properties of linear VAEs. Notably Lucas et al. (2019) show the equivalence of β and $\mathrm{Var}[x|z]$ in Gaussian VAEs (whose scope we discuss in Appendix E); and prove identifiability of LVAEs (which we generalise to the non-linear case). Zietlow et al. (2021) show that disentanglement can be sensitive to perturbing the data. Reizinger et al. (2022) seek to relate the ELBO to *independent mechanism analysis* (Gresele et al., 2021), which encourages column-orthogonality in the mixing function of ICA.[6] Disentanglement relates closely to (noisy) ICA (§2, e.g. Hyvarinen, 1998). ICA assumes the same generative LVM but does not model the posterior, which we show is critical to disentanglement in VAEs, or otherwise explicitly induce C1-C2, necessary for disentanglement under D1, rather asymmetric priors or side information are typically required.

Our pushforward perspective unifies several GAN disentanglement results. Ramesh et al. (2018) identify disentangled directions from singular vectors of the generator Jacobian, close to our seam/s.v. path view. Wei et al. (2021) encourage output variations induced by different latent co-ordinates to be orthogonal, close to our C1. Peebles et al. (2020) suppress mixed second derivatives of the generator, related to the mechanism behind C2. We show, for a general pushforward density, why such derivative-side regularisers are not merely heuristic, but target structure that characterises D1 disentanglement in smooth pushforward models.

Chadebec & Allassonnière (2022) and Arvanitidis et al. (2018) consider paths in latent space defined by the inverse image of paths over the data manifold (our *s.v. paths*). Syrota et al. (2025) identify invariant metric quantities, such as distances, despite latent-coordinate ambiguity; our question is what additional density and decoder structure selects factor axes. Horan et al. (2021) prove identifiability using local isometry and non-Gaussianity. At the metric level, $J^\top J = I$ is the $s_i = 1$ special case of C1-C2; our density-Hessian spectrum assumption selects axes differently and neither theorem subsumes the other. Pan et al. (2023) claim that the data manifold is identifiable from a geometric perspective assuming Jacobian-orthogonality, differing to our probabilistic factorisation view. Bhowal et al. (2024) consider linear and non-linear components of the encoder/decoder, loosely resembling our Jacobian SVD view. However, dissecting a function into linear/non-linear components is not well defined, whereas distinct singular values determine unique singular directions, up to permutation/sign. Buchholz et al. (2022) analyse identifiability of function classes, e.g. proving that *conformal maps* are identifiable and *orthogonal coordinate transformations* (satisfying C1) are not. Our theoretically derived disentanglement constraints C1-C2 define a function class corresponding to a density factorisation that is unique and hence identifiable, up to symmetry (Thm. 6.5). Brady et al. (2023) and Lachapelle et al. (2023) analyze *additive/compositional* decoders with block-diagonal Hessians in pixel space, strictly stronger than our Thm. 4.2, which allows, e.g., rotations/scale/colour changes that would not be block-diagonal in pixel space.

9 Conclusion

Disentanglement is often evaluated by comparing learned latents to known ground-truth factors, but those factors are unavailable in genuinely unsupervised settings. We instead show that disentanglement can be defined at the level of the learned pushforward density. Independent factors can be encoded as one-dimensional seam factors of the manifold density, each controlled by a latent

[6] We report discrepancies in Reizinger et al. (2022) in Appendix K.

co-ordinate. Here, the latent factorisation $p(z) = \prod_i p(z_i)$ can be considered *projected / lifted* onto the manifold, with independent densities over *singular vector paths* in latent space pushed-forward to independent densities over manifold *seams*.

Our main results are not specific to any particular methodology, e.g. VAEs, but general to smooth push-forwards of a factorised prior, which include GANs and flows. For such pushforwards, D1 is equivalent to C1-C2 and, with the stated simple density-Hessian spectrum, seam factors are identifiable up to natural symmetries. The C1-C2 restrictions allow this result, despite the known *unidentifiability* of unrestricted nonlinear ICA (Hyvärinen & Pajunen, 1999).

In Gaussian VAEs, diagonal posteriors can favour the C1-C2 decoder structure necessary for this disentanglement through the variational gap, while the expectation, term-wise and approximation qualifications give a potential explanation of the variability in the observed effect. We also see that β of a β-VAE indirectly controls posterior covariance, explaining the empirical finding that $\beta > 1$ promotes disentanglement while degrading generative quality (and $\beta < 1$ mitigates *posterior collapse*). Prior works also show that generative factors can be identified in the GAN latent space via closely related generator derivatives (Ramesh et al., 2018; Wei et al., 2021; Peebles et al., 2020).

VAEs and their variants form part of many state-of-the-art modelling pipelines, e.g. latent diffusion (e.g. Rombach et al., 2022; Pandey et al., 2022; Yang et al., 2023; Zhang et al., 2022) and LLMs. Other recent works show that supervised learning (Dhuliawala et al., 2024) and self-supervised learning (Bizeul et al., 2024) can be viewed as latent models. In future work we will look to extend our findings to these other latent paradigms.

Acknowledgments

Carl is grateful to Gabriel Peyré, Maarten de Hoop, Eddie Aamari and Ivana Balazevic for helpful discussions over the course of this work; and to CFM for funding of the Laplace Fellowship. This work was performed using HPC resources from GENCI–IDRIS (Grant 2025-AD011015749R1).

References

Arvanitidis, G., Hansen, L. K., and Hauberg, S. Latent space oddity: On the curvature of deep generative models. In *ICLR*, 2018.

Bao, X., Lucas, J., Sachdeva, S., and Grosse, R. B. Regularized linear autoencoders recover the principal components, eventually. In *NeurIPS*, 2020.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013.

Bhowal, P., Soni, A., and Rambhatla, S. Why do variational autoencoders really promote disentanglement? In *ICML*, 2024.

Bizeul, A., Schölkopf, B., and Allen, C. A Probabilistic Model to explain Self-Supervised Representation Learning. In *TMLR*, 2024.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. In *Conference on Computational Natural Language Learning*, 2015.

Brady, J., Zimmermann, R. S., Sharma, Y., Schölkopf, B., Von Kügelgen, J., and Brendel, W. Provably learning object-centric representations. In *ICML*, 2023.

Buchholz, S., Besserve, M., and Schölkopf, B. Function classes for identifiable nonlinear independent component analysis. In *NeurIPS*, 2022.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Chadebec, C. and Allassonnière, S. A geometric perspective on variational autoencoders. In *NeurIPS*, 2022.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Dhuliawala, S., Sachan, M., and Allen, C. Variational Classification. In *TMLR*, 2024.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NeurIPS*, 2014.

Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanism analysis, a new concept? In *NeurIPS*, 2021.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*, 2017.

Horan, D., Richardson, E., and Weiss, Y. When is unsupervised disentanglement possible? In *NeurIPS*, 2021.

Hyvarinen, A. Noisy independent component analysis, maximum likelihood estimation, and competitive learning. In *IEEE International Joint Conference on Neural Networks Proceedings*, 1998.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. In *Neural Networks*, volume 12, pp. 429–439, 1999.

Kim, H. and Mnih, A. Disentangling by factorising. In *ICML*, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *ICLR*, 2014.

Koehler, F., Mehta, V., Zhou, C., and Risteski, A. Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias. In *ICLR*, 2022.

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

Lachapelle, S., Mahajan, D., Mitliagkas, I., and Lacoste-Julien, S. Additive decoders for latent variables identification and cartesian-product extrapolation. In *NeurIPS*, 2023.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Don't Blame the ELBO! a Linear VAE Perspective on Posterior Collapse. In *NeurIPS*, 2019.

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

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

Pandey, K., Mukherjee, A., Rai, P., and Kumar, A. Diffusevae: Efficient, controllable and high-fidelity generation from low-dimensional latents. In *TMLR*, 2022.

Peebles, W., Peebles, J., Zhu, J.-Y., Efros, A., and Torralba, A. The hessian penalty: A weak prior for unsupervised disentanglement. In *ECCV*, 2020.

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

Reizinger, P., Gresele, L., Brady, J., Von Kügelgen, J., Zietlow, D., Schölkopf, B., Martius, G., Brendel, W., and Besserve, M. Embrace the gap: Vaes perform independent mechanism analysis. In *NeurIPS*, 2022.

Rezende, D. J. and Viola, F. Taming vaes. *arXiv preprint arXiv:1810.00597*, 2018.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Rolinek, M., Zietlow, D., and Martius, G. Variational Autoencoders Pursue PCA Directions (by Accident). In *CVPR*, 2019.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Syrota, S., Zainchkovskyy, Y., Xi, J., B., B.-R., and Hauberg, S. Identifying metric structures of deep latent variable models. In *ICML*, 2025.

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

Wei, Y., Shi, Y., Liu, X., Ji, Z., Gao, Y., Wu, Z., and Zuo, W. Orthogonal jacobian regularization for unsupervised disentanglement in image generation. In *ICCV*, 2021.

Yang, T., Wang, Y., Lu, Y., and Zheng, N. Disdiff: unsupervised disentanglement of diffusion probabilistic models. In *NeurIPS*, 2023.

Zhang, Z., Zhao, Z., and Lin, Z. Unsupervised representation learning from pre-trained diffusion probabilistic models. In *NeurIPS*, 2022.

Zietlow, D., Rolinek, M., and Martius, G. Demystifying inductive biases for (beta-) vae based architectures. In *ICML*, 2021.

A Dropping posterior diagonality in the linear case

To highlight the role of *diagonal* posteriors in linear disentanglement (§3) we consider *full* posterior LVAEs whereby $\boldsymbol{R} \neq \boldsymbol{I}$ in general in Eq. 4. The broad argument in §3 follows except that (a) columns $\boldsymbol{r}_i \in \mathcal{Z}$ of $\boldsymbol{R}$ map to independent $\boldsymbol{u}_i$ directions in $\mathcal{X}$; whereas (b) standard basis vectors in $\mathcal{Z}$ map to $\boldsymbol{u}_i^\top \boldsymbol{R}$ directions, which are arbitrary with respect to the independent features along $\boldsymbol{u}_i$ directions. Axis-aligned traversals in latent space thus correspond to varying *entangled* combinations of (not distinct) features. This is shown empirically in Fig. 2 (see caption for details).

B Singular Vector Paths & Seams

Definition D3 (Singular-vector path). For $z^* \in \mathcal{Z}_{\text{reg}}$, the i-th singular-vector path is the curve z_t^i satisfying $z_0^i = z^*$ and $\dot{z}_t^i = \boldsymbol{v}_i(z_t^i)$ on its maximal regular interval $I_{z^*,i}$. Write $\mathcal{V}_{z^*}^i = \{z_t^i : t \in I_{z^*,i}\}$. Reversing the vector-field sign reverses orientation.

Definition D4 (Seam). The corresponding seam is $x_t^i = g(z_t^i)$, with image $\mathcal{M}_{g,z^*}^i$. The chain rule gives $\dot{x}_t^i = s_i(z_t^i)\boldsymbol{u}_i(z_t^i)$. Its signed arc-length coordinate is $u_i(t) = \int_0^t s_i(z_\tau^i)\, d\tau$, so $\dot{u}_i = s_i > 0$.

Lemma B.1 (Paths $\mapsto$ seams). *Let $g : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e.*, $z^* \in \mathcal{Z}_{\text{reg}}$, $i \in \{1, \dots, d\}$, and let $\mathcal{V}_{z^*}^i$ be the i-th s.v. path through z^*. Then the image of $\mathcal{V}_{z^*}^i$ under g is the i-th seam through $g(z^*)$: $\mathcal{M}_{g,z^*}^i = \{g(z) : z \in \mathcal{V}_{z^*}^i\}$.*

Proof. For $x_t^i \doteq g(z_t^i)$, by the chain rule and SVD: $\frac{dx_t^i}{dt} = \boldsymbol{J}_{z_t^i}\frac{dz_t^i}{dt} = \boldsymbol{J}_{z_t^i}\boldsymbol{v}^i(z_t^i) = s^i(z_t^i)\,\boldsymbol{u}^i(z_t^i)$, so x_t^i satisfies Def. 4. □

Lemma 4.1 (Factorisation over seams). *Let $g : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e. *and the prior factorise as $p(z) = \prod_{i=1}^d p_i(z_i)$. Then, the manifold density p_μ on $\mathcal{M}_g$ factorises as*

$$p_\mu\big(g(z)\big) \;=\; \prod_{i=1}^{d} \tfrac{p_i(z_i)}{s_i(z)}\,, \quad \textit{for every } z \in \mathcal{Z}_{\text{reg}}. \tag{7}$$

Moreover, under C1, each factor $\frac{p_i(z_i)}{s^i(z)}$ is the 1-D density over the i-th seam $\mathcal{M}_{g,z}^i$ at $x = g(z)$, obtained by pushing forward the 1-D marginal $p_i(z_i)$ over the i-th s.v. path $\mathcal{V}_z^i$ through z parallel to the i^{th} latent axis.

Proof. By a standard change–of–variables (on embedded manifolds) and $|\boldsymbol{J}_z^\top \boldsymbol{J}_z| = \prod_{i=1}^d s^i(z)^2$,

$$p_\mu\big(g(z)\big) \;=\; \det(\boldsymbol{J}_z^\top \boldsymbol{J}_z)^{-1/2}\, p(z) \;=\; \tfrac{\prod_{i=1}^d p_i(z_i)}{\prod_{i=1}^d s^i(z)}\,,$$

yielding Eq. 7. Under C1, s.v.-paths are axis aligned in latent space and map to seams. For each i and $x = g(z)$, the change–of–variables formula along $\mathcal{M}_{g,z}^i$ (i-th seam through x with coordinate u_i, Def. 4) at $t = 0$ gives the local pushed 1-D *seam–density*

$$f_i^{(z)}\big(u_i(t)\big) \;\doteq\; \tfrac{p_i([z_t^i]_i)}{s^i(z_t^i)}\,, \qquad \tfrac{d}{dt}u_i(t) = s^i(z_t^i),\;\; u_i(0) = 0, \tag{11}$$

where $t \mapsto z_t^i$ is the i-th singular–vector path through z (Def. 3) with co-ordinate $z_i^i(t)$. Evaluating at $t = 0$: $f_i^{(z)}\big(u_i(0)\big) = \frac{p_i(z_i)}{s^i(z)}$, shows that seam-densities are the factors of Eq. 7. □

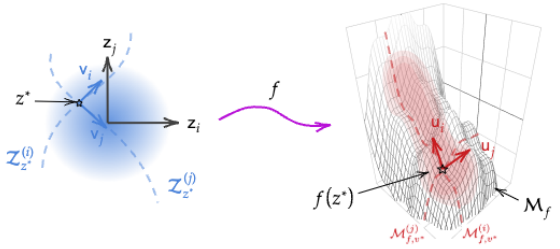


Figure 5: **Seam factorisation**: For c.i.d.a.e. $f : \mathcal{Z} \to \mathcal{X}$, with manifold $\mathcal{M}_f \subseteq \mathcal{X}$, *s.v. paths* $\mathcal{V}_{z^*}^k \subseteq \mathcal{Z}$ (dashed blue) following right singular vectors $\boldsymbol{v}_k$ of Jacobian $\boldsymbol{J}_{z^*}$ at z^* (solid blue), map to *seams* $\mathcal{M}_{f,z^*}^k \subseteq \mathcal{M}_f$ (dashed red) following left singular vectors $\boldsymbol{u}_k$ at $f(z^*) \in \mathcal{X}$ (solid red). $\boldsymbol{z}_k \in \mathcal{Z}$ are standard basis vectors.

C Disentanglement

Theorem 4.2 (**Disentanglement ⇔ C1-C2**). *Let* $g : \mathcal{Z} \to \mathcal{X}$ *be* c.i.d.a.e. *and the prior factorise as* $p(z) = \prod_{i=1}^{d} p_i(z_i)$. *The push-forward density* p_μ *on the manifold* $\mathcal{M}_g$ *is* disentangled *(D1) if and only if* g *satisfies* ***C1*** *and* ***C2*** *a.e..*

Proof. (C1/2 ⇒D1) By Thm. 4.1, the manifold density factorises pointwise as $p_\mu(g(z)) = \prod_{i=1}^{d} \frac{p_i(z_i)}{s^i(z)}$. By **C1**, $\boldsymbol{V}_z = \boldsymbol{I}$ for all z, so the i-th singular–vector path through z is exactly the axis–aligned line $\{z' : [z']_i \text{ varies}, [z']_{\neg i} = [z]_{\neg i}\}$; by Lemma B.1 its image is the i-th seam through $x = g(z)$ following $\boldsymbol{u}^i$. By **C2**, $s^i(z)$ depends only on z_i. Define the seam coordinate u_i along the i-th seam as in D4; then u_i is a strictly monotone function of z_i, hence the 1-D push–forward of p_i along that seam is $f_i(u_i) = \left|\frac{du_i}{dz_i}\right|^{-1} p_i(z_i) = \frac{p_i(z_i)}{s^i(z_i)}$, Thus $p_\mu(g(z)) = \prod_i f_i(u_i(z))$ with each f_i evaluated on the i-th seam. Finally, since u_i is monotone in z_i and $\{z_i\}$ are independent, the random variables $\{u_i\}$ are independent; hence factors $\{f_i(u_i)\}$ are *statistically independent* as required by D 1.

(D1⇒ C1/2) Assume p_μ is disentangled under g. By D 1, each factor f_i is obtained by pushing forward $p(z_i)$ over an axis-aligned line in direction $\boldsymbol{z}^i$, and that line's image follows $\boldsymbol{J}_z \boldsymbol{z}^i = \boldsymbol{J}_z^i$ at $g(z)$ (column i of $\boldsymbol{J}_z$) with 1-D density $f_i = |\boldsymbol{J}_z^i|^{-1} p(z_i)$. Thus, $p(g(z)) = |\boldsymbol{J}_z|^{-1} p(z)$ by standard change-of-variables but also $p(g(z)) = \prod_i f_i = \prod_i |\boldsymbol{J}_z^i|^{-1} p(z_i)$, by factor independence. Hence $|\boldsymbol{J}_z| = \prod_i |\boldsymbol{J}_z^i|$, which occurs if and only if columns $\boldsymbol{J}_z^i$ are orthogonal, i.e. $\boldsymbol{J}_z^\top \boldsymbol{J}_z$ is diagonal or in SVD terms $\boldsymbol{J}_z = \boldsymbol{U}_z \boldsymbol{S}_z$ (**C1**).

Since $s_i \doteq [\boldsymbol{S}_z]_{ii} = |\boldsymbol{J}_z^i|$ then $f_i = p(z_i)/s_i$. Each factor f_i is a function of z_i, so by independence, for $i \neq j$, $0 = \frac{\partial \log f_i}{\partial z_j} = \frac{\partial \log p_i}{\partial z_j} - \frac{\partial \log s_i}{\partial z_j} = -\frac{\partial \log s_i}{\partial z_j}$ (since z_i are independent), implying $\frac{\partial s_i}{\partial z_j} = 0$ for $i \neq j$ (**C2**). □

D Proof of Decoder Derivative Constraints

Property P1. For fixed x and z, matrices $\boldsymbol{J}_z^\top \boldsymbol{J}_z$ and $(x - d(z))^\top \mathbf{H}_z$ (as in Eq. 9) are <u>each</u> diagonal.

Lemma 5.1 (VAE Disentanglement). *If a trained Gaussian VAE with diagonal posteriors induces **P1** for z in a neighbourhood of $e(x)$, then it induces **C1** on that neighbourhood and **C2** to a first order approximation (Appendix D bounds the residual Taylor error).*

Proof. ***(Preliminaries)***: Recall $p(z) = \mathcal{N}(0, \boldsymbol{I})$ and $p(x \mid z) = \mathcal{N}(x; d(z), \sigma^2 \boldsymbol{I})$. Let $q(z \mid x) = \mathcal{N}(z; e(x), \Sigma_x)$ be the trained posterior with Σ_x diagonal (by assumption). Denote the SVD of the decoder Jacobian by

$$\begin{aligned} \boldsymbol{J}_z = \boldsymbol{U}_z \, \boldsymbol{S}_z \, \boldsymbol{V}_z^\top, \quad \boldsymbol{U}_z \in \mathbb{R}^{m \times d}, \quad \boldsymbol{V}_z \in \mathbb{R}^{d \times d}, \\ \boldsymbol{S}_z = \mathrm{Diag}(s_1(z), \ldots, s_d(z)), \quad s_i(z) > 0. \end{aligned} \tag{12}$$

Assume full column rank on the manifold ($s_i(z) > 0$). For Gaussian likelihood with variance σ^2, the Hessian of the log-likelihood w.r.t. z can be written

$$\nabla_z^2 \log p(x \mid z) \;=\; -\tfrac{1}{\sigma^2} \big(\boldsymbol{J}_z^\top \boldsymbol{J}_z - \sum_{\ell=1}^{n} r(z)_\ell \, \boldsymbol{H}_\ell(z) \big),$$

where $r(z) = x - d(z)$ and $\boldsymbol{H}_\ell(z) \in \mathbb{R}^{d \times d}$ is the Hessian of the ℓ-th decoder coordinate, $[\boldsymbol{H}_\ell]_{pq} = \partial^2 d_\ell / \partial z_p \, \partial z_q$. Combined with the Opper–Archambeau fixed-point yields

$$\begin{aligned} \Sigma_x^{-1} \;&=\; -\,\mathbb{E}_q\big[\nabla_z^2 \log p(x, z)\big] \\ &=\; -\,\mathbb{E}_q\big[\nabla_z^2 \log p(z) + \nabla_z^2 \log p(x|z)\big] \\ &=\; \boldsymbol{I} - \mathbb{E}_q\big[\nabla_z^2 \log p(x \mid z)\big] \\ &=\; \boldsymbol{I} + \tfrac{1}{\sigma^2} \mathbb{E}_q\big[\boldsymbol{J}_z^\top \boldsymbol{J}_z\big] - \tfrac{1}{\sigma^2} \mathbb{E}_q\big[\sum_{\ell=1}^{n} r(z)_\ell \, \boldsymbol{H}_\ell(z)\big]. \end{aligned} \tag{13}$$

(C1): For diagonal Σ_x in Eq. 13 and z concentrated around $e(x)$ under P1 ("no cancellation"), we have

$$\boldsymbol{J}_z^\top \boldsymbol{J}_z \;\text{ is diagonal,} \quad \sum_{\ell=1}^{n} r(z)_\ell \, \boldsymbol{H}_\ell(z) \;\text{ is diagonal.} \tag{14}$$

Diagonal $\boldsymbol{J}_z^\top \boldsymbol{J}_z = \mathrm{Diag}(s_1^2, \ldots, s_k^2)$ implies right singular vectors $\boldsymbol{v}^i$ are the standard basis, up to signed permutations. By relabelling latent axes and absorbing signs, $\boldsymbol{V}_z = \boldsymbol{I}$ for all z visited by the encoder, establishing **C1**.

(Directed Hessian is Tangent to Manifold): For a trained model, $d(e(x)) \approx x$ is assumed to be small. For z concentrated around $e(x)$, i.e. $z = e(x) + \delta$ with $\delta > 0$ small, a first-order Taylor expansion gives

$$\begin{aligned} d(z) \;&=\; d(e(x)) + \boldsymbol{J}_{e(x)} \, \delta + O(\|\delta\|^2) \\ \Rightarrow \quad r(z) \;&\doteq\; x - d(z) \;\approx\; -\boldsymbol{J}_{e(x)} \, \delta + O(\|\delta\|^2). \end{aligned} \tag{15}$$

Thus, to first order (we quantify the bound below), for z concentrated around $e(x)$, $r(z)$ lie in the column space of $\boldsymbol{J}_{e(x)}$, i.e. in the span of the left singular vectors $\{u_i(z)\}_{i=1}^k$ (columns of $\boldsymbol{U}_z$):

$$r \;=\; \boldsymbol{U}_z a, \qquad a \in \mathbb{R}^k \tag{16}$$

Hence, Eq. 14 implies that $\sum_{\ell=1}^n r_\ell \, \boldsymbol{H}_\ell(z)$ is diagonal for all $r \in \mathrm{span}(\boldsymbol{U}_z)$. In particular, $\sum_{\ell=1}^n u_{i\ell}^\top \, \boldsymbol{H}_\ell(z)$ is diagonal for all rows u_i of $\boldsymbol{U}_z$ and, by definition of slices $\boldsymbol{H}_\ell$ and $\boldsymbol{J}_z$,

$$[\sum_{\ell=1}^{n} u_{i\ell}^\top \, \boldsymbol{H}_\ell(z)]_{pq} \;=\; \sum_{\ell=1}^{n} u_{i\ell}^\top \, \tfrac{\partial^2 d_\ell}{\partial z_p z_q} \;=\; (\boldsymbol{U}_z^\top \tfrac{\partial \boldsymbol{J}_z}{\partial z_p})_{iq} \tag{17}$$

(Diagonality of $(\frac{\partial \boldsymbol{s}_i}{\partial \boldsymbol{z}_j})_{i,j}$):

Differentiating $\boldsymbol{U}_z^\top \boldsymbol{U}_z = \boldsymbol{I}$ to get $\frac{\partial \boldsymbol{U}_z^\top}{\partial z_j}\boldsymbol{U}_z + \boldsymbol{U}_z^\top \frac{\partial \boldsymbol{U}_z}{\partial z_j} = \boldsymbol{0}$ shows $\Omega_j(z) := \boldsymbol{U}_z^\top \frac{\partial \boldsymbol{U}_z}{\partial z_j} \in \mathbb{R}^{d\times d}$ is skew-symmetric. Differentiating $\boldsymbol{J}_z = \boldsymbol{U}_z \boldsymbol{S}_z$ w.r.t. z_j and premultiplying by $\boldsymbol{U}_z^\top$ then gives

$$\boldsymbol{U}_z^\top \frac{\partial \boldsymbol{J}_z}{\partial z_j} \;=\; \Omega_j(z)\,\boldsymbol{S}_z \;+\; \frac{\partial \boldsymbol{S}_z}{\partial z_j}. \tag{18}$$

Since $\Omega_j(z)$ is skew-symmetric, all diagonal entries are zero; since $\frac{\partial \boldsymbol{S}_z}{\partial z_j}$ is diagonal, all non-diagonal entries are zero. Thus, of respective entries, only one is non-zero and can be considered separately.

From Eq. 17, only $(\Omega_j(z)\,\boldsymbol{S}_z)_{:j}$ elements can be non-zero, thus *all* elements of $\Omega_j(z)$ must be zero (by skew-sym.), **ruling out rotation in the tangent plane**[7]

For $\frac{\partial \boldsymbol{S}_z}{\partial z_j}$, diagonality (only $(\frac{\partial \boldsymbol{S}_z}{\partial z_j})_{kk}$ elements non-zero) and Eq. 17 (only $(\frac{\partial \boldsymbol{S}_z}{\partial z_j})_{:j}$ elements non-zero) imply that only elements $(\frac{\partial \boldsymbol{S}_z}{\partial z_j})_{jj} = \frac{\partial s_j}{\partial z_j}$ can be non-zero, eliminating mixed partials

$$\frac{\partial s_i}{\partial z_j}(z) \;=\; 0 \quad \text{for all } i \neq j,$$

i.e. the Jacobian of the singular-value map $s(z) = (s_1(z), \ldots, s_k(z))$ is *diagonal*, proving **C2**. □

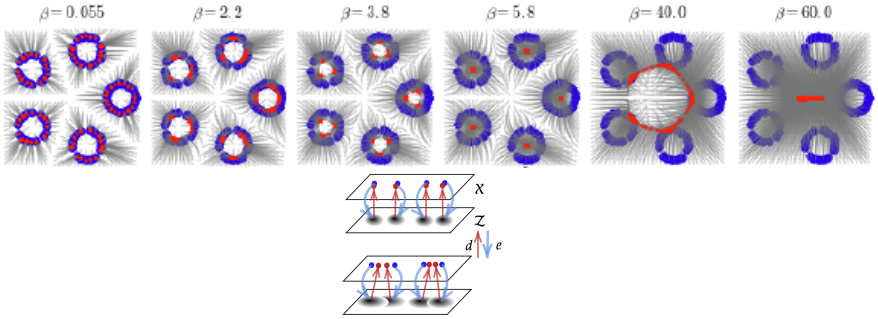


Figure 6: Illustrating $\beta \propto \mathrm{Var}[x|z]$ (blue=data, red=reconstruction): (*l*) For low β ($\beta = 0.55$), $\mathrm{Var}[x|z]$ is low (by Eq. 9), and data must be well reconstructed (right, top). As β increases, $\mathrm{Var}[x|z]$ and so $\mathrm{Var}[z|x]$ increase, and posteriors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

Taylor-remainder bound. Let $\mu = e(x)$, $\delta = z - \mu$, $\eta = x - d(\mu)$ and suppose $\|D^2 d(w)[a,b]\| \leq M\|a\|\|b\|$ on every segment $[\mu, z]$ considered. Taylor expansion about z gives

$$r(z) = -J_z\delta + \varepsilon(z), \qquad \|\varepsilon(z)\| \leq \|\eta\| + \tfrac{M}{2}\|\delta\|^2.$$

For the directed-Hessian matrices this implies

$$\|r(z)^\top H_z - (-J_z\delta)^\top H_z\|_{\mathrm{op}} \leq M\|\eta\| + \tfrac{M^2}{2}\|\delta\|^2.$$

If the bound holds over the full Gaussian posterior support, the expected bound (RHS) is $M\|\eta\| + \frac{M^2}{2}\,\mathrm{tr}\,\Sigma_x$. For a bounded neighbourhood the same statement holds conditionally, with $\mathbb{E}[\|\delta\|^2 \mid z \text{ in the neighbourhood}]$ instead. The bound concerns ambient Taylor error, not intrinsic curvature. It is small with accurate reconstruction, concentrated posteriors and bounded decoder curvature.

E β controls Noise Variance

Choosing $\beta > 1$ in Eq. 2 can enhance disentanglement (Higgins et al., 2017; Burgess et al., 2018) and has been viewed as re-weighting ELBO components or as a Lagrange multiplier. We show that β implicitly controls the likelihood's variance and that the "β-ELBO" remains a valid objective.

Dividing the ELBO by a constant and suitably adjusting the learning rate leaves the VAE training algorithm unchanged, hence consider Eq. 2 divided through by β. For a Gaussian VAE with fixed variance $\mathrm{Var}[x|z] = \sigma^2$, this equates to a standard VAE with variance $\beta\sigma^2$ (Lucas et al., 2019). More generally, this β^{-1}-scaling acts as a *temperature* parameter giving an implied likelihood $p_\theta(x|z)^{1/\beta}$:

[7] we have: $\Omega_j(z)_{kj} = -\Omega_j(z)_{jk} = 0$, if $j \neq k$; and $\Omega_j(z)_{jj} = 0$ (skew-sym.).

- as $\beta \to \infty$, the implied entropy tends to uniform (as though there were more noise in the data, so the model fits more loosely), and
- as $\beta \to 0$, the implied entropy reduces to a delta (so reconstructions should be tight).

Optimal posteriors fit to the implicit likelihood, $q_\phi(z|x) \propto p_\theta(x|z)^{1/\beta} p(z)$, which thus dilate ($\beta > 1$) or concentrate ($\beta < 1$), generalising the known Gaussian result (Lucas et al., 2019).[8]

Empirical support: The above claim is well illustrated on synthetic data in Fig. 6 (Rezende & Viola, 2018, see caption for details). It also explains the *blur in β-VAEs*, since $\beta > 1$ effectively assumes more noise the model fits more loosely. It also explains why $\beta < 1$ helps mitigate *posterior collapse* (Bowman et al., 2015), i.e. when a VAE's likelihood is sufficiently expressive that it can model the full data distribution (i.e. $p_\theta(x|z) = p(x)$ for some θ), leaving latent variables redundant and the posterior "collapses" to prior. As $\beta \to 0$, the implied likelihood entropy, and so the range of distributions that can be described, reduces. Thus for some $\beta < 1$, the effective variance falls below $\mathrm{Var}[x]$, rendering posterior collapse impossible as some variance in x can only be explained by some latent z. Our claim that β controls effective variance thus explains well-known empirical observations, which in turn provide empirical support for the claim.

F Identifiability Proofs

F.1 Proof of Linear VAE Identifiability

Theorem 6.1 (**LVAE Identifiability**). *Let data be generated under the linear Gaussian LVM (Eq. 3) with ground-truth $g(z) = \boldsymbol{W}z$, $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top \in \mathbb{R}^{m \times d}$ of full column rank and* distinct *singular values. Let an LVAE with* diagonal *posteriors attain a global population ELBO optimum such that its learned parameters yield $p_\mu^{(d)} \equiv p_\mu^{(g)}$ on the mean manifold. Then the LVAE achieves disentanglement (D1) and identifies ground-truth independent components on $\mathcal{M}_g$ up to permutation and sign.*

Proof. ***(Ground truth)*** If $x = \boldsymbol{W}z$, then letting $u \doteq \boldsymbol{U}^\top x$ and $v \doteq \boldsymbol{V}^\top z$, gives $u = \boldsymbol{S}v$, $u_i = s_i v_i$. Since $z \sim \mathcal{N}(0, \boldsymbol{I})$ and $\boldsymbol{V}$ is orthonormal, $v \sim \mathcal{N}(0, \boldsymbol{I})$, hence $\{v_i\}$, and $\{u_i\}$ are mutually independent: $u_i \sim \mathcal{N}(0, s_i^2)$ and $p_\mu^{(\mathrm{g})}(x) = \prod_{i=1}^d \mathcal{N}(u_i; 0, s_i^2)$.

(Model) Let $d(z) = \boldsymbol{D}z$ with SVD $\boldsymbol{D} = \boldsymbol{U}_{\boldsymbol{D}}\boldsymbol{S}_{\boldsymbol{D}}\boldsymbol{V}_{\boldsymbol{D}}^\top$. For a global population optimum, rotating the decoder on the right preserves its generated Gaussian distribution and can diagonalise its exact posterior covariance. The corresponding posterior mean is representable by the encoder, so this choice attains zero variational gap. Any globally optimal matching decoder must also attain zero gap; hence its exact posterior precision $I + \sigma^{-2} D^\top D$ is diagonal, giving C1. Since C2 is vacuously satisfied, Theorem 4.2 implies that $p_\mu^{(\mathrm{d})}$ is disentangled and factorises into statistically independent components along the decoder's seams (columns of $\boldsymbol{U}_{\boldsymbol{D}}$). Since $u_{\boldsymbol{D},i} = s_{\boldsymbol{D},i} z_i$ and $z_i \sim \mathcal{N}(0, 1)$, each seam factor is Gaussian with variance $s_{\boldsymbol{D},i}^2$, i.e. $p_\mu^{(\mathrm{d})} = \prod_{i=1}^d \mathcal{N}(u_{\boldsymbol{D},i}; 0, s_{\boldsymbol{D},i}^2)$.

(Matching) Equality $p_\mu^{(\mathrm{d})} \equiv p_\mu^{(\mathrm{g})}$ and *distinct* $\{s_i\}$ imply uniqueness of the Gaussian product decomposition, up to permutation. Thus the LVAE's independent components (seam factors) match ground-truth components up to permutation/sign, i.e. identifiability and disentanglement on $\mathcal{M}_g$. □

F.2 Proof of Gaussian VAE Identifiability

Lemma 6.3 (**Seams are Intrinsic**). *Let $\mathcal{M} \subseteq \mathcal{X}$ carry a manifold density p_μ. Assume that on a regular set $\mathcal{M}_{\mathrm{reg}}$ there exist local coordinates $u = (u_1, \ldots, u_d)$ (see Eq. 8), whose coordinate lines are 1-D seams and directions form an orthonormal basis of $T_x\mathcal{M}$ at each $x \in \mathcal{M}_{\mathrm{reg}}$, together with 1-D densities $\{f_i\}$ such that*

$$p_\mu(x) \;=\; \prod_{i=1}^{d} f_i\big(u_i(x)\big), \qquad x \in \mathcal{M}_{\mathrm{reg}}. \tag{10}$$

[8] Technically, the β-ELBO's value is incorrect without renormalising the implicit likelihood, but that is typically irrelevant, e.g. for commonly used Gaussian likelihoods, only the quadratic "MSE" term appears in the loss.

Let $\mathbf{H}_x \doteq \nabla^2_{\mathcal{M}} \log p_\mu(x)$ denote the intrinsic Hessian on the manifold, and assume its eigenvalues are pairwise distinct a.e.. Then wherever the eigenvalues are distinct, the d seam directions are determined intrinsically by p_μ, as eigenvectors of $\mathbf{H}_x$, unique up to permutation and sign.

Proof. Fix $x \in \mathcal{M}_{\text{reg}}$, and let $\boldsymbol{u}^1, \dots, \boldsymbol{u}^d$ be the unit seam directions at x, stacked as columns of $\boldsymbol{U}_x \in \mathbb{R}^{m \times d}$. By assumption, these directions form an orthonormal basis of $T_x\mathcal{M}$, with local seam coordinates $u = (u_1, \dots, u_d)$.

Since

$$\log p_\mu(x) = \sum_{i=1}^{d} \log f_i(u_i(x)),$$

its mixed second derivatives in the seam coordinates vanish:

$$\left[\nabla^2_u \log p_\mu(x)\right]_{ij} = \begin{cases} \frac{\partial^2}{\partial u_i^2} \log f_i(u_i(x)) & (i = j), \\ 0 & (i \neq j). \end{cases}$$

Because the seam coordinates are orthonormal, the manifold metric in these coordinates is the identity, so the intrinsic Hessian equals the ordinary Hessian in the u-coordinates at x.[9] Hence

$$\mathbf{H}_x \;=\; \boldsymbol{U}_x \left[\nabla^2_u \log p_\mu(x)\right] \boldsymbol{U}_x^\top$$

is an eigendecomposition of $\mathbf{H}_x$.

Therefore the seam directions are eigenvectors of $\mathbf{H}_x$. Since the eigenvalues are pairwise distinct a.e., these directions are unique up to permutation and sign. □

Implication for identifiability. Lemma 6.3 isolates the intrinsic geometry of p_μ: once p_μ is fixed, the seams and their directions are fixed (P&S). Any Gaussian VAE decoder d matching p_μ and satisfying **C1–C2** must therefore align its singular paths with those seams and inherit the same seam factors.

Lemma 6.4 (A matching pushforward finds seams). *Under assumptions of Lemma 6.3, let $d : \mathcal{Z} \to \mathcal{X}$ be* c.i.d.a.e. *with factorised prior $p(z) = \prod_i p_i(z_i)$ and push-forward density $p_\mu^{(d)} \equiv p_\mu$* matching *on $\mathcal{M}_d \doteq \{d(z)\} = \mathcal{M}$. If d satisfies **C1–C2** a.e., then for any z and $x = d(z)$:*

- *left singular vectors $\boldsymbol{U}_z$ of $\boldsymbol{J}_z$ coincide with the intrinsic seam directions in Lemma 6.3 (up to permutation and sign);*
- *the images under d of singular-vector paths (D3) are exactly the intrinsic seams through x;*
- *along the i-th intrinsic seam, the factor f_i is the 1-D push-forward of $p_i(z_i)$ (as in Lemma 4.1).*

Proof. Since $p_\mu^{(d)} \equiv p_\mu$, both define the same manifold density on the common manifold $\mathcal{M}$.

Under **C1–C2**, Lemma 4.1 gives

$$p_\mu(x) = \prod_{i=1}^{d} \frac{p_i(z_i)}{s_i(z)}, \qquad x = d(z),$$

with each $s_i(z)$ depending only on z_i by **C2**. Define local seam coordinates by

$$u_i(z_i) \doteq \int_0^{z_i} s_i(\zeta)\, d\zeta.$$

Then each factor depends only on the corresponding seam coordinate:

$$p_\mu(x) = \prod_{i=1}^{d} f_i(u_i(x)), \qquad f_i(u_i) = \frac{p_i(z_i)}{s_i(z_i)}.$$

[9] Here "orthonormal seam coordinates" means that the seam coordinate vector fields form an orthonormal coordinate frame on the local patch.

Moreover, by the chain rule and **C1**,

$$\frac{\partial x}{\partial u_i} = \frac{\partial x}{\partial z_i}\frac{\partial z_i}{\partial u_i} = \frac{\boldsymbol{J}_z \boldsymbol{z}_i}{s_i(z_i)} = \boldsymbol{u}_i(z).$$

Hence the seam coordinates induced by d are orthonormal, since

$$\Big\langle \frac{\partial x}{\partial u_i}, \frac{\partial x}{\partial u_j} \Big\rangle = \boldsymbol{u}_i(z)^\top \boldsymbol{u}_j(z) = \delta_{ij}.$$

Thus the assumptions of Lemma 6.3 hold, with seam frame $\boldsymbol{U}_z$. By that lemma, the columns of $\boldsymbol{U}_z$ are exactly the intrinsic seam directions, up to P&S. By Lemma B.1, singular-vector paths map to seams; and by Lemma 4.1, the factor along seam i is the 1-D push-forward of $p_i(z_i)$. □

Note that the two proofs above adopt a similar technique, but Lemma 6.3 is entirely intrinsic to the manifold (hence no mention of a Jacobian), whereas Lemma 6.4 is with reference to a parameterisation of the manifold by a function d.

Theorem 6.5 (**Gaussian VAE Identifiability**). *Let data be generated by* c.i.d.a.e. $g : \mathcal{Z} \to \mathcal{X}$ *with standard Gaussian prior. Let a Gaussian VAE with* diagonal *posteriors learn a decoder* $d : \mathcal{Z} \to \mathcal{X}$. *Suppose both* g *and* d *satisfy **C1–C2** and manifold densities match:* $p_\mu^{(d)} \equiv p_\mu^{(g)}$ *on* $\mathcal{M} = \{g(z)\} = \{d(z)\}$. *If the eigenvalues of the intrinsic log-density Hessian* $\nabla^2_{\mathcal{M}} \log p_\mu$ *are distinct a.e., then* d *identifies ground-truth independent components on* $\mathcal{M}_g$*, up to permutation and sign.*

Proof. (**Matching** $\boldsymbol{U}$) Equality $p_\mu^{(g)} \equiv p_\mu^{(d)}$ implies the same H_x. By Lemma 6.3, its eigenvectors are intrinsic and unique (P&S), so $\boldsymbol{U}_z^{(d)} = \boldsymbol{U}_z^{(g)}$ (P&S, herein assume indices are relabelled to match).

(**Matching** $\boldsymbol{S}$) In this common basis, under **C1–C2**, on-manifold scores are equal:

$$\begin{aligned}\big[\nabla_u \log p_\mu^{(g)}(x)\big]_i &= \tfrac{1}{s_i^{(g)}(z)} \tfrac{\partial}{\partial z_i}\Big(\log p_i(z_i) - \log s_i^{(g)}(z)\Big) \\ &= \big[\nabla_u \log p_\mu^{(d)}(x)\big]_i,\end{aligned}$$

hence integrating along seam i (with $z_{\neg i}$ fixed, using equality of p_μ at a reference point to fix the integration constant), 1-D seam factors $\frac{p_i(z_i)}{s_i}$ match. Since p_i is fixed, $s_i^{(d)} = s_i^{(g)}$, i.e. $\boldsymbol{S}_z^{(d)} = \boldsymbol{S}_z^{(g)}$.

With $\boldsymbol{V}_z^{(d)} = \boldsymbol{V}_z^{(g)} = \boldsymbol{I}$ by **C1**, it follows that $\boldsymbol{J}_z^{(d)} = \boldsymbol{J}_z^{(g)}$ and the seam decomposition is identified up to P&S. □

G Disentanglement Metrics

Axis alignment score (AAS): AAS is an auxiliary benchmark diagnostic, not a definition of disentanglement. Given mutual information matrix $M_{ji} = I(z_j; y_i)$ between learned latents and known factors, AAS measures closeness to a permuted diagonal:

$$\mathrm{AAS}(M) = \frac{1}{2}\frac{\sum_j \max_i M_{ji} + \sum_i \max_j M_{ji}}{\sum_{j,i} M_{ji}}.$$

The row maxima penalise latents that spread mass across factors; the column maxima penalise factors spread across latents. Thus AAS complements factor-centric metrics such as MIG by measuring approximate one-to-one axis alignment.

Normalised off diagonal: For gradient terms (here, a $d \times d$ matrix M) we compute a measure of diagonality by computing the ratios of normalised off-diagonal absolute values to on-diagonal values.

```
d = M.shape[1]
num_off_diag = d * (d - 1)
M = abs(M)
M = diag(M)^(-0.5) * M * diag(M)^(-0.5)       # normalise
mean_off_diag = (sum(M) - sum(diag(M))) / num_off_diag
```

H Empirical Relationship between Diagonal Posteriors and Diagonalised Decoder Derivatives.

For ease of reference we show results of Rolinek et al. (2019) showing that VAEs with diagonal posteriors tend to have higher column-orthogonality in the Jacobian (equivalent to C1).

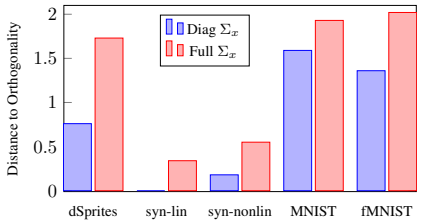


Figure 7: **Empirical support for P1**: VAEs with diagonal Σ_x show higher Jacobian column-orthogonality (Rolinek et al., 2019).

I Empirical results on Natural Data (CelebA)

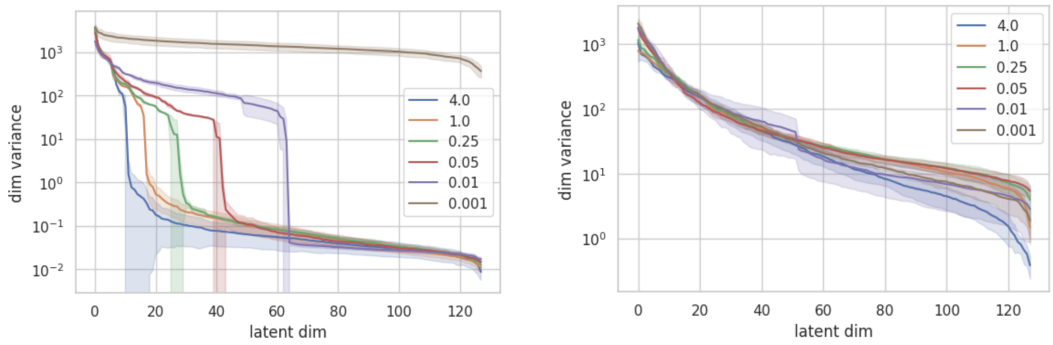
This appendix complements §7, applying the same architecture and training regime used for dSprites (and as in Burgess et al. (2018)) to CelebA, a more complex natural dataset without introducing new confounds. We vary β and compare *diagonal* vs *full* posterior covariances, reporting: (i) latent traversals showing the dependence of disentanglement on posterior structure (Fig. 8); (ii) how utilisation of the latent space varies with β and posterior structure (Fig. 9); (iii) diagonality of the Price/Bonnet derivative terms (Fig. 10); and (iv) reconstruction/sampling quality (11). See captions for details.



Diagonal posteriors | Full posteriors

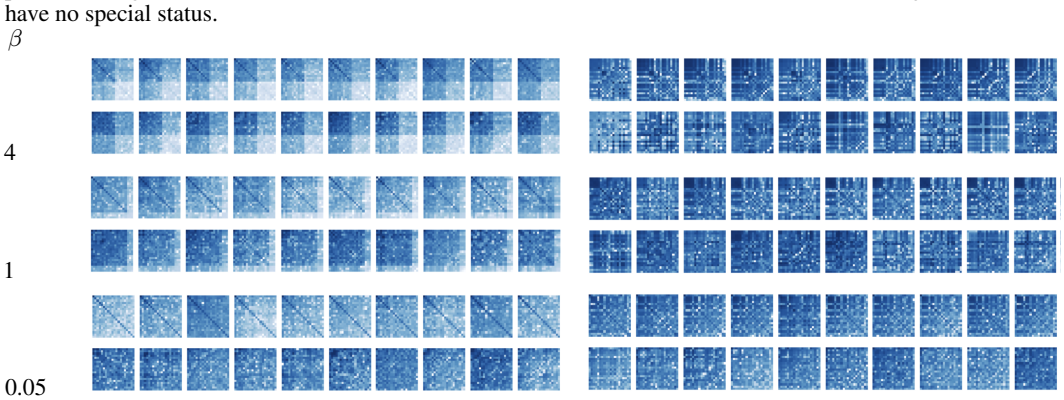
Figure 8: **Traversals over dimensions of highest variance** ($\beta = 4$): Each row shows images generated as individual latent dimensions z_i are varied with rows ordered by latent activity (See Fig. 9 caption). For diagonal posteriors (left), traversals more clearly demonstrate disentanglement, i.e. identification of distinct semantic features with distinct latent dimensions, e.g. background shade (row 1), facial orientation (row 3), lighting (row 5), background colour (row 7). For full covariance posteriors, independent features are less clearly assigned to distinct latent dimensions, i.e. less disentangled.

Summary: For *diagonal* posteriors, we observe: (a) a small set of *active* latents whose number increases as β decreases; (b) stronger Jacobian orthogonality among active dimensions; and (c) reconstructions/samples of comparable quality to the full-covariance model across β (Figures 8–11). These match the predictions of our theory and mirror the synthetic/dSprites trends, supporting



Diagonal posteriors Full posteriors

Figure 9: **Active latent dimension depend on β (denoted by colour) and posterior covariance structure (left/right)**: For a trained model and for each latent dimension z_i, the variance $x|z_i$ is estimated by taking equidistant traversals in latent space and computing the Euclidean distance between samples at each end of the traversal. The plot shows the (estimated) variance, or *latent activity*, per dimension ordered by magnitude (log scale, mean over 5 runs, standard deviation indicated by shaded areas). With diagonal covariances (left), there are relatively sharp cliff-edges implying that dimensions are (broadly) *active* or *inactive* and that axis-aligned directions are preferred. The number of active dimensions increases as β reduces and reconstructions get "sharper", requiring more information to be captured and thus greater capacity (i.e. latent dimensions). For full posteriors (right), the distribution of variance over dimensions is much smoother and axis-aligned directions have no special status.



(a) Diagonal posteriors (b) Full posteriors

Figure 10: **Heatmaps of Derivative Terms in Eq. 9 for CelebA for Diagonal and Full Posterior Covariances**: For each value of β (indicated on left), there are two rows for $\boldsymbol{J}^\top \boldsymbol{J}$ (upper) and the directed Hessian (lower) over 10 random test samples. Colour intensity indicates log magnitude of matrix entries. For each heatmap, rows and columns correspond to a latent dimension z_i, ordered by latent activity (See Fig. 9 caption). We show the top 20 most active dimensions. For diagonal covariances (left), the active dimensions are visible as a darker block in the upper left, which grows as β reduces (matching Fig. 9), and diagonal structure is visible for active dimensions of $\boldsymbol{J}^\top \boldsymbol{J}$. Such structure is not visible for full covariances (right). The Hessians show less discernable structure and we suspect that such a higher-order derivative requires more samples to be well estimated for a complex distribution.

on natural data the claim that diagonal posteriors favour Jacobian orthogonality and latent-factor alignment.



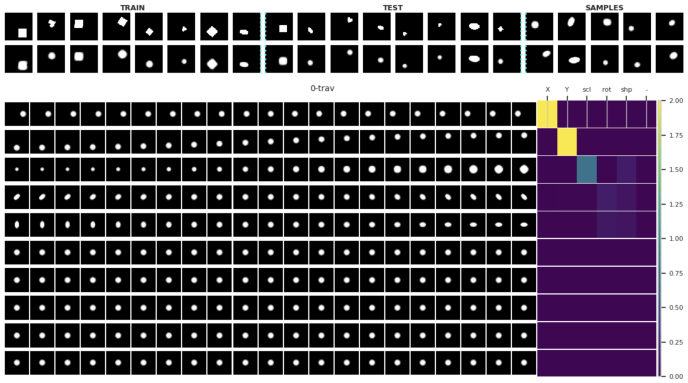
(a) Diagonal posteriors



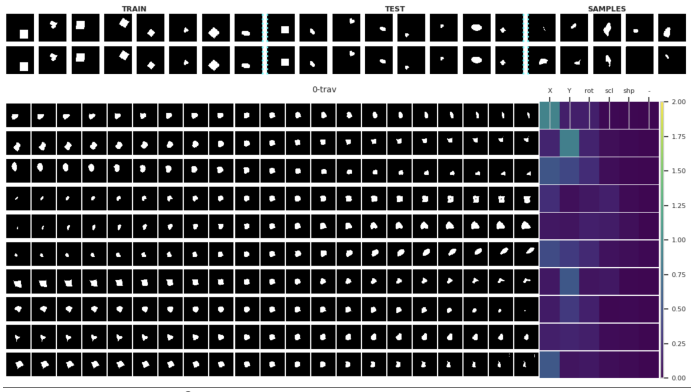
(b) Full posteriors

Figure 11: **Reconstructions and Samples for CelebA for a range of β values and Diagonal and Full Posterior Covariances**: For each value of β (left), there are two rows in three sections: (left) train samples (upper) and reconstructions (lower); (mid) test samples (upper) and reconstructions (lower); and (right) samples (both rows). As β reduces, reconstruction quality and samples improve (i.e. blur reduces). Reconstruction and sample quality is broadly comparable for diagonal and full covariances, indicating that the latent space is reoriented towards axis-alignment without necessarily impacting performance.

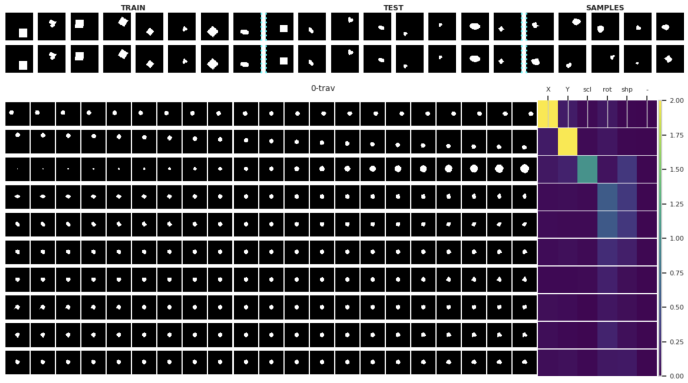
J Reducing β over training



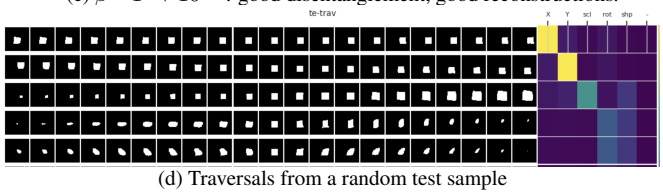
(a) $\beta = 1$: good disentanglement, blurry reconstructions.



(b) $\beta = 10^{-3}$: no clear disentanglement, good reconstructions.



(c) $\beta = 1 \rightarrow 10^{-3}$: good disentanglement, good reconstructions.



(d) Traversals from a random test sample

Figure 12: **Testing the β-hypothesis:** (top) high β (1) gives best disentanglement (see heatmap) but blurry images (see top rows); (mid) low β (0.001) gives poor disentanglement but good reconstructions; (bottom) lowering β over training ($1 \rightarrow 0.001$) gives good disentanglement (see heatmap) and good reconstructions.

K Proof discrepancies in Reizinger et al. (2022)

The following discrepancies concern the displayed proof of Theorem 1 in Reizinger et al. (2022). The two inequalities and omitted remainder below mean that this proof does not establish Theorem 1 as stated; they do not by themselves show that its conclusion is false.

1. p.33, after "triangle inequality": $\left|\mathbb{E}\left[\|a\|^2 - \|b\|^2\right]\right| \leq \mathbb{E}\left[\|a-b\|^2\right]$, where $a = x - f,\ b = -\sum \frac{\partial f}{\partial z_k} ...$
   - (dropping expectations for clarity) this has the form $\left|\|a\|^2 - \|b\|^2\right| \leq \|a-b\|^2$ (*)
   - true triangle inequality: $\left|\|a\| - \|b\|\right| \leq \|a-b\| \implies \|a\| - \|b\|\|^2 \leq \|a-b\|^2$ (by squaring)
     - this differs to (*) since norms are squared inside the absolute operator on the L.H.S.
   - counter-example to (*): $b = x > 0,\ a = x+1 \implies \left|\|a\|^2 - \|b\|^2\right| = |2x+1| > 1 = \|a-b\|^2$
2. next step, p.33: $\mathbb{E}\left[\|(c-e) - (d-e)\|^2\right] \leq \mathbb{E}\left[\|c-e\|^2 + \|d-e\|^2\right]$ where $c = x,\ d = f(z) - \sum \frac{\partial f}{\partial z_k} ...,\ e = f(\mu)$
   - this has the form of the standard triangle inequality $\|a-b\| \leq \|a\| + \|b\|$ except all norms are squared.
   - squaring both sides of the triangle inequality gives an additional cross term on the right that the used inequality omits, without which the inequality does not hold in general.
3. first step, p.34: drops the K term, which bounds the decoder Hessian and higher derivatives (in earlier Taylor expansion)
   - this omission is similar to a step in Kumar & Poole (2020) but is not stated, e.g. in Assumption 1.
   - since K is unbounded, any conclusion omitting it without justification is not valid in general.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our core claims are theoretic and proven in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Our core theorems apply under specific mathematical conditions that are stated. Our theoretic results apply where latent factors are independent, which is clearly stated.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All proofs are given in the appendices and briefly described in the main text where space allows.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Our main contributions are theoretic and we run no novel experiments per se. We include Empirical Support for our theoretical claims that are straightforward to reproduce using available implementations (e.g. VAE).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [N/A]

   Justification: Our main contributions are theoretic and we run no novel experiments per se. We include Empirical Support for our theoretical claims that are straightforward to reproduce using available implementations (e.g. VAE).

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: Our main contributions are theoretic and we run no novel experiments per se. We include Empirical Support for our theoretical claims that are straightforward to reproduce using available implementations (e.g. VAE).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [N/A]

Justification: Our main contributions are theoretic and we run no novel experiments per se. We include Empirical Support for our theoretical claims that are straightforward to reproduce using available implementations (e.g. VAE).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: Our main contributions are theoretic and we run no novel experiments per se. We include Empirical Support for our theoretical claims that are straightforward to reproduce using available implementations (e.g. VAE).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Authors have reviewed the NeurIPS Code of Ethics.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: Our main contributions are theoretic to explain previously observed phenomena. We believe a better understanding of what machine learning models already do should help mitigate the risk of accidental negative societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our main contributions are theoretic and we envisage no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All works relied on are referenced in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our main contributions are theoretic and we release no new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our main contributions are theoretic and we use no human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our main contributions are theoretic and we involve no human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Language models were used to help improve paper drafting and confirm results in the paper, but we do not "involve LLMs as any important, original, or non-standard components"

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.