Supplementary Material to: Variational Autoencoders Pursue PCA Directions (by Accident)

The supplementary information is structured as follows. We start with a remark on Table 1 and then provide the proofs in Section A.1. Section B reports the details of the experiments followed by additional experiments in Section C.

Remark on Table 1

Some dataset-architecture combinations listed in Table 2 are omitted for the following reasons.

On the one hand, calculating the Disentanglement Score for MNIST and fMNIST does not make sense, as the generating factors are not given (the one categorical label cannot serve as replacement). Consequently, as the values of β are chosen according to this score, we do not report β-VAE numbers for these datasets. On the other hand, for either synthetic task, the regular VAE vastly overprunes, see Figure S1, and the values become meaningless.

A. Proofs

A.1. Proof of Theorem 2

Proof strategy: For part (b), we aim to derive a lower bound on the objective (18), that is independent from the optimization variables $\sigma_j^2(\mathbf{x}^i)$ and V_i. Moreover, we show that this lower bound is tight for some specific choices of $\sigma_j^2(\mathbf{x}^i)$ and V_i, i.e. the global optima. For these choices, all J_i will have orthogonal columns.

The strategy for part (a) is to show that whenever $\sigma_j^2(\mathbf{x}^i)$ and V_i do not induce a global optimum, we can find a small perturbation that decreases the objective function. Thereby showing that local minima do not exist.

Technical lemmas: We begin with introducing a few useful statements. First is the inequality between arithmetic and geometric mean; a consequence of Jensen's inequality.

Lemma S1 (AM-GM inequality)**.** *Let $a_1, \ldots, a_N$ be non-negative real numbers. Then*

$$\frac{1}{N}\sum_{i=1}^{N} a_i \geq \left(\prod_{i=1}^{N} a_i\right)^{1/N} \tag{S1}$$

with equality occuring if and only if $a_1 = a_2 = \cdots = a_n$.

The second bound to be used is the classical Hadamard's inequality.

Lemma S2 (Hadamard's inequality [4])**.** *Let $M \in \mathbb{R}^{k\times k}$ be non-singular matrix with column vectors $c_1, \ldots, c_k$. Then*

$$\prod_{i=1}^{k} \|c_i\| \geq |\det M| \tag{S2}$$

with equality if and only if the vectors $c_1, \ldots, c_k$ are pairwise orthogonal.

And finally a simple lemma for characterizing matrices with orthogonal columns.

Lemma S3 (Column orthogonality)**.** *Let $M \in \mathbb{R}^{n\times d}$ be a matrix and let $M = U\Sigma V^\top$ be its singular value decomposition. Then the following statements are equivalent:*

(a) The columns of M are (pairwise) orthogonal.

(b) The matrix $M^\top M$ is diagonal.

(c) The columns of $\Sigma V^\top$ are (pairwise) orthogonal.

Proof. The equivalence of (a) and (b) is immediate. For equivalence of (a) and (c) it suffices to notice that if we set $M' = \Sigma V^\top$, then

$$M'^\top M' = V\Sigma^\top \Sigma V^\top = M^\top M. \tag{S3}$$

The equivalence of (a) and (b) now implies that M has orthogonal columns if and only if M' does. □

Initial considerations: First, without loss of generality, we will ignore all passive latent variables (in the sense of Definition 1). Formally speaking, we will restrict to the case when the local decoder mappings J_i are non-degenerate (i.e. have non-zero singular values). Now d denotes the dimensionality of the latent space with $d = |V_a|$.

Next, we simplify the loss $L_{\approx\mathrm{KL}}$, Equation 10. Up to additive and multiplicative constants, this loss can be, for a fixed sample $\mathbf{x}^i \in X$, written as

$$\|\mu(\mathbf{x}^i)\|^2 + \sum_{j=1}^{d} -\log(\sigma_j^2(\mathbf{x}^i)). \tag{S4}$$

In the optimization problem (18, 19) the values $\mu(\mathbf{x}^i)$ can only be affected via applying an orthogonal transformation V_i. But such transformation are norm-preserving (isometric) and hence the values $\|\mu(\mathbf{x}^i)\|^2$ do not change in the optimization. As a result, we can restate the constraint (19) as

$$\sum_{\mathbf{x}^i \in X} \sum_{j=1}^{d} -\log(\sigma_j^2(\mathbf{x}^i)) = C_1 \tag{S5}$$

for some constant C_1.

Proof of Theorem 2(b): Here, we explain how Theorem 2(b) follows from the following two propositions.

Proposition S1. *For a fixed sample $\mathbf{x}^i \in X$ let us denote by $c_1, \ldots, c_d$ the column vectors of J_i. Then*

$$\mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 \geq d \left(\prod_{j=1}^{d} \|c_j\|^2 \sigma_j^2(\mathbf{x}^i) \right)^{1/d} \tag{S6}$$

with equality if and only if $\|c_j\|^2 \sigma_j^2(\mathbf{x}^i) = \|c_k\|^2 \sigma_k^2(\mathbf{x}^i)$ for every $j, k \in \{1, \ldots, d\}$.

Proposition S2. *Let $M \in \mathbb{R}^{n \times d}$, where $d < n$, be a matrix with column vectors $c_1, \ldots, c_d$ and nonzero singular values $s_1, \ldots, s_d$. Then*

$$\prod_{j=1}^{d} \|c_j\| \geq \det{}^{\dagger}(M), \tag{S7}$$

where by $\det^{\dagger}(M)$ we denote the product of the singular values of M. Equality occurs if and only if $c_1, \ldots, c_d$ are pairwise orthogonal.

First, Proposition S2 allows making further estimates in the inequality from Proposition S1. Indeed, we get

$$\mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 \geq d \left(\left(\det{}^{\dagger}(J_i)\right)^2 \prod_{j=1}^{d} \sigma_j^2(\mathbf{x}^i) \right)^{1/d} \tag{S8}$$

and after applying the (monotonous) log function we are left with

$$\log \mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 \geq \tag{S9}$$

$$\log(d) + \frac{2}{d} \log(\det{}^{\dagger}(J_i)) + \frac{1}{d} \sum_{j=1}^{d} \log(\sigma_j^2(\mathbf{x}^i)). \tag{S10}$$

Finally, we sum over the samples $\mathbf{x}^i \in X$ and simplify via (S5) as

$$\sum_{\mathbf{x}^i \in X} \log \mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 \geq$$

$$N \log(d) - \frac{C_1}{d} + \frac{2}{d} \sum_{\mathbf{x}^i \in X} \log(\det{}^{\dagger}(J_i)). \tag{S11}$$

The right-hand side of this inequality is independent from the values of $\sigma_j^2(\mathbf{x}^i)$, as well as from the orthogonal matrices V_i, since these do not influence the singular values of any J_i.

Moreover, it is possible to make inequality (S11) tight (i.e. reach the global minimum), by setting $\sigma_j^2(\mathbf{x}^i)$ as hinted by Proposition S1 and by choosing the matrices V_i such that every J_i has orthogonal columns (this is clearly possible as seen in Proposition 1).

This yields the desired description of the global minima of (18). □

Proof of Proposition S1: We further denote by $r_1, \ldots, r_n$ the row vectors of J_i, and by $a_{r,c}$ the element of J_i at r-th row and c-th column. With sampling $\boldsymbol{\varepsilon}(\mathbf{x}^i)$ according to

$$\varepsilon(\mathbf{x}^i) \sim \mathcal{N}\left(0, \operatorname{diag} \sigma^2(\mathbf{x}^i)\right), \tag{S12}$$

we begin simplifying the objective (18) with

$$\mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 = \mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \sum_{k=1}^{n} \|r_k^{\top} \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 \tag{S13}$$

$$= \sum_{k=1}^{n} \mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|r_k^{\top} \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2. \tag{S14}$$

Now, as the samples $\boldsymbol{\varepsilon}(\mathbf{x}^i)$ are zero mean, we can further write

$$\sum_{k=1}^{n} \mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|r_k^{\top} \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 = \sum_{k=1}^{n} \operatorname{var}(r_k^{\top} \boldsymbol{\varepsilon}(\mathbf{x}^i)). \tag{S15}$$

Now we use the fact that for uncorrelated random variables A and B we have $\operatorname{var}(A + cB) = \operatorname{var} A + c^2 \operatorname{var} B$. This allows to expand the variance of the inner product as

$$\operatorname{var}(r_k^{\top} \boldsymbol{\varepsilon}(\mathbf{x}^i)) = \operatorname{var}\left(\sum_{j=1}^{d} a_{k,j} \boldsymbol{\varepsilon}_j(\mathbf{x}^i) \right) \tag{S16}$$

$$= \sum_{j=1}^{d} a_{k,j}^2 \operatorname{var} \boldsymbol{\varepsilon}_j(\mathbf{x}^i) = \sum_{j=1}^{d} a_{k,j}^2 \sigma_j^2(\mathbf{x}^i).$$

Now, we can regroup the terms via

$$\sum_{k=1}^{n} \operatorname{var}(r_k^\top \boldsymbol{\varepsilon}(\mathbf{x}^i)) = \sum_{k=1}^{n}\sum_{j=1}^{d} a_{k,j}^2 \sigma_j^2(\mathbf{x}^i)$$
$$= \sum_{j=1}^{d}\sum_{k=1}^{n} a_{k,j}^2 \sigma_j^2(\mathbf{x}^i)$$
$$= \sum_{j=1}^{d} \|c_j\|^2 \sigma_j^2(\mathbf{x}^i). \tag{S17}$$

All in all, we obtain

$$\mathop{\mathbb{E}}_{\varepsilon(\mathbf{x}^i)} \|J_i \boldsymbol{\varepsilon}(\mathbf{x}^i)\|^2 = \sum_{j=1}^{d} \|c_j\|^2 \sigma_j^2(\mathbf{x}^i). \tag{S18}$$

from which the desired inequality follows via setting $a_j = \|c_j\|^2 \sigma_j^2(\mathbf{x}^i)$ for $j = 1, \ldots, d$ in Lemma S1. Indeed, then we have

$$\sum_{j=1}^{d} \|c_j\|^2 \sigma_j^2(\mathbf{x}^i) \geq d \left(\prod_{j=1}^{d} \|c_j\|^2 \sigma_j^2(\mathbf{x}^i) \right)^{1/d} \tag{S19}$$

as required. □

Proof of Proposition S2: As the first step, we show that both sides of the desired inequality are invariant to multiplying the matrix M from the left with an orthogonal matrix $U \in \mathbb{R}^{n \times n}$.

For the right-hand side, this is clear as the singular values of UM are identical to those of M. As for the left-hand side, we first need to realize that the vectors c_j are the images of the canonical basis vectors e_j, i.e. $c_j = Me_j$ for $j = 1, \ldots, d$. But since U is an isometry, we have $\|UMe_j\| = \|Me_j\| = \|c_j\|$ for every j, and hence also the column norms are intact by prepending U to M.

This allows us to restrict to matrices M for which the SVD has a simplified form $M = \Sigma V^\top$. Next, let us denote by $\Sigma_{d\times d}$ the $d \times d$ top-left submatrix of Σ. Note that $\Sigma_{d\times d}$ contains all nonzero elements of Σ. As a result, the matrix $M' = \Sigma_{d\times d} V^\top$ contains precisely the nonzero rows of the matrix M. This implies

$$M^\top M = M'^\top M'. \tag{S20}$$

In particular, the column vectors c'_j of M' have the same norms as those of M. Now we can write

$$\prod_{j=1}^{d} \|c_j\| = \prod_{j=1}^{d} \|c'_j\| \geq |\det(M')| = \det{}^\dagger(M), \tag{S21}$$

where the inequality follows from Lemma S2 applied to nonsingular matrix M'. Equality in Lemma S2 occurs precisely if the columns of M' are orthogonal. However, according to Lemma S3 and (S20), it also follows that the columns of M' are orthogonal if and only if the columns of M are. Note that Lemma S3(c) is needed for covering the reduction performed in the first two paragraphs. □

Proof of Theorem 2(a): We show the nonexistence of local minima as follows. For any values of $\sigma_j^2(\mathbf{x}^i)$ and V_i that do not minimize the objective function (18), we find a small perturbation that improves this objective.

All estimates involved in establishing inequality (S11) rely on either Lemma S1 or Lemma S2, where in both cases, the right-hand side was kept fixed. We show that both of these inequalities can be tightened in such fashion by small perturbations in their parameters.

Lemma S4 (Locally improving AM-GM)**.** *For any nonnegative values $a_1, \ldots, a_N$ for which*

$$\frac{1}{N}\sum_{i=1}^{N} a_i > \left(\prod_{i=1}^{N} a_i \right)^{1/N} \tag{S22}$$

there exists a small perturbation a'_i of a_i for $i = 1, \ldots, N$ such that

$$\frac{1}{N}\sum_{i=1}^{N} a_i > \frac{1}{N}\sum_{i=1}^{N} a'_i \geq \tag{S23}$$
$$\left(\prod_{i=1}^{N} a'_i \right)^{1/N} = \left(\prod_{i=1}^{N} a_i \right)^{1/N} \tag{S24}$$

Proof. Since (S22) is a sharp inequality, we have $a_i > a_j$ for some $i \neq j$. Then setting $a'_i = a_i/(1+\delta)$, $a'_j = a_j(1+\delta)$, and $a'_k = a_k$ otherwise, will do the trick. Indeed, we have $a_i a_j = a'_i a'_j$ as well as $a_i + a_j > a'_i + a'_j$ for small enough δ. This ensures both S23 and S24. □

An analogous statement for Lemma S2 has the following form.

Lemma S5 (Locally improving Hadamard's inequality)**.** *Let $M \in \mathbb{R}^{k\times k}$ be a non-singular matrix with SVD $M = U\Sigma V^\top$, and column vectors $c_1, \ldots, c_k$, for which*

$$\prod_{i=1}^{k} \|c_i\| > |\det M|. \tag{S25}$$

Then there exists an orthogonal matrix V', a small perturbation of V, such that if we denote by $c'_1, \ldots, c'_k$ the column vectors of $M' = U\Sigma V'^\top$, we have

$$\prod_{i=1}^{k} \|c_i\| > \prod_{i=1}^{k} \|c'_i\|. \tag{S26}$$

Proof. We proceed by induction on k. For $k = 2$, it can be verified directly that for some small δ (in absolute value) setting $V' = VR_\delta$, where R_δ is a 2D rotation matrix by angle δ, achieves what is required.

For the general case, the sharp inequality (S25) implies that $c_i^\top c_j \neq 0$ for some pair of $i \neq j$. Without loss of generality, let $i = 1$, $j = 2$. In such case, we consider $V' = VR_\delta^{2D}$, where

$$R_\delta^{2D} = \begin{pmatrix} R_\delta & \\ & \mathcal{I}_{k-2} \end{pmatrix} \tag{S27}$$

is a block diagonal matrix, in which R_δ is again a 2×2 rotation matrix. By design, we have $c_i = c'_i$ for $i > 2$. This, along with the fact that U can be set to $\mathcal{I}_k$ (isometry does not influence either side of (S25)), allows for a full reduction to the discussed two-dimensional case. □

It is easy to see that the performed perturbations continuously translate into perturbations of the parameters $\sigma_j^2(\mathbf{x}^i)$ and V_i in estimates (S19) and (S21). Consequently, any non-optimal values of $\sigma_j^2(\mathbf{x}^i)$ and V_i can be locally improved. This concludes the proof.

A.2. Rotational invariances

Let us start by fleshing out the common elements of the proofs of Propositions 2 and 3. In both cases, the encoder and decoder mappings $\mathrm{Enc}_{\varphi,U}$, $\mathrm{Dec}_{\theta,U}$ induce joint distributions $p_U(\mathbf{x}, \mathbf{z})$, $q_U(\mathbf{x}, \mathbf{z})$ described as

$$p_U(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x} \mid U^\top \mathbf{z}) \tag{S28}$$

$$q_U(\mathbf{x}, \mathbf{z}) = q(\mathbf{x})q(U^\top \mathbf{z} \mid \mathbf{x}) \tag{S29}$$

Lemma S6. *For every $\mathbf{x}^i \in X$ we have $p(\mathbf{x}^i) = p_U(\mathbf{x}^i)$.*

Proof. We simply compute

$$\begin{aligned} p_U(\mathbf{x}^i) &= \int p_U(\mathbf{x}^i, \mathbf{z})\, \mathrm{d}\mathbf{z} \\ &= \int p(\mathbf{z})p(\mathbf{x}^i \mid U^\top \mathbf{z})\, \mathrm{d}\mathbf{z} \\ &= \int p(U\mathbf{z})p(\mathbf{x}^i \mid \mathbf{z})\, \mathrm{d}\mathbf{z} \\ &= \int p(\mathbf{z})p(\mathbf{x}^i \mid \mathbf{z})\, \mathrm{d}\mathbf{z} = p(\mathbf{x}^i), \end{aligned}$$

where in the third equality we used the Change of Variable Theorem to substitute $U\mathbf{z}$ for $\mathbf{z}$ (keep in mind that $|\det(U)| = 1$ as U is an orthogonal matrix). In the fourth equality, we used the rotational symmetry of the prior $p(\mathbf{z})$. □

Proof of Proposition 2. This immediately follows from Lemma S6. □

Proof of Proposition 3. We utilize the full identity from ELBO derivation. For fixed $\mathbf{x}^i \in X$ we have [2]

$$\mathrm{ELBO} = D_{\mathrm{KL}}(q_U(\mathbf{z} \mid \mathbf{x}^i) \parallel p_U(\mathbf{z} \mid \mathbf{x}^i)) + \log p_U(\mathbf{x}^i) \tag{S30}$$

In order to prove invariance of ELBO to the choice of U, it suffices to prove invariance of the right-hand side of (S30). Due to Proposition (3) we only need to focus on the KL term. Similarly as in the proof of Lemma S6, we calculate

$$\begin{aligned} &D_{\mathrm{KL}}(q_U(\mathbf{z} \mid \mathbf{x}^i) \parallel p_U(\mathbf{z} \mid \mathbf{x}^i)) \\ &= \int q_U(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q_U(\mathbf{z} \mid \mathbf{x}^i)}{p_U(\mathbf{z} \mid \mathbf{x}^i)}\, \mathrm{d}\mathbf{z} \\ &= \int q_U(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q_U(\mathbf{z} \mid \mathbf{x}^i) \cdot p_U(\mathbf{x}^i)}{p_U(\mathbf{z}) \cdot p_U(\mathbf{x}^i \mid \mathbf{z})}\, \mathrm{d}\mathbf{z} \\ &\overset{(3)}{=} \int q(U^\top \mathbf{z} \mid \mathbf{x}^i) \log \frac{q(U^\top \mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{z}) \cdot p(\mathbf{x}^i \mid U^\top \mathbf{z})}\, \mathrm{d}\mathbf{z} \\ &\overset{(4)}{=} \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{U}\mathbf{z}) \cdot p(\mathbf{x}^i \mid \mathbf{z})}\, \mathrm{d}\mathbf{z} \\ &\overset{(5)}{=} \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i) \cdot p(\mathbf{x}^i)}{p(\mathbf{z}) \cdot p(\mathbf{x}^i \mid \mathbf{z})}\, \mathrm{d}\mathbf{z} \\ &= \int q(\mathbf{z} \mid \mathbf{x}^i) \log \frac{q(\mathbf{z} \mid \mathbf{x}^i)}{p(\mathbf{z} \mid \mathbf{x}^i)}\, \mathrm{d}\mathbf{z} \\ &= D_{\mathrm{KL}}(q(\mathbf{z} \mid \mathbf{x}^i) \parallel p(\mathbf{z} \mid \mathbf{x}^i)), \end{aligned}$$

where we again used the Change of Variable Theorem in equality (4), rotational symmetry of $p(\mathbf{z})$ in equality (5), and Lemma S6 in equality (3). □

A.3. Other proofs

Proof of Proposition 1. Recall from Lemma S3 that column orthogonality of M is equivalent to $M^\top M$ being a diagonal matrix.

(b) ⇒ (a): Let $M = U\Sigma V^\top$ where $|V|$ is a permutation matrix. Then

$$M^\top M = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma' V^\top \tag{S31}$$

where $\Sigma' = \Sigma^\top \Sigma$ is a diagonal matrix. But then $V\Sigma' V^\top$ only permutes the diagonal entries of Σ' (and possibly flips their signs). In particular, $V\Sigma' V^\top$ is also diagonal.

(a) ⇒ (b): Let again $M = U\Sigma V^\top$ be some SVD of M and assume $M^\top M = D$ for some diagonal matrix D. Since M has d distinct nonzero singular values, $M^\top M$ has d distinct nonzero eigenvalues (diagonal elements). Moreover, these eigenvalues are precisely the squares of the singular values captured by Σ. Next, if we denote by P the permutation matrix for which PDP^{-1} has decreasing diagonal elements, we can write

$$PDP^{-1} = \Sigma^\top \Sigma \tag{S32}$$

Then using (S32) and the SVD of M similarly as in (S31), we obtain

$$D = M^\top M = V\Sigma^\top \Sigma V^\top = VPDP^{-1}V^\top. \tag{S33}$$

Further, the resulting identity $(VP)D = D(VP)$ implies that columns of VP are eigenvectors of D, i.e. the canonical basis vectors. Since VP is additionally orthogonal, these eigenvectors are normalized. It follows that $|VP|$ is a permutation matrix and the conclusion follows.

□

Proof of Proposition 4. First, note that for any random variable $\mathbf{X} \in \mathbb{R}^k$ with $\mathbb{E}\mathbf{X} = \mu$ and a constant $\mathbf{b} \in \mathbb{R}^k$, the following identity holds

$$\mathbb{E}\,\|\mathbf{X} - \mathbf{b}\|^2 = \mathbb{E}\,\|\mathbf{X} - \mu\|^2 + \|\mu - \mathbf{b}\|^2. \tag{S34}$$

In our case, we set $\mathbf{X} = \mathrm{Dec}_\theta(\mathrm{Enc}_\varphi(\mathbf{x}^i))$, the unbiasedness assumption translates to $\mathbb{E}\mathbf{X} = \mathrm{Dec}_\theta(\mu(\mathbf{x}^i))$, and finally we set $\mathbf{b} = \mathbf{x}^i$.

The identity we obtain, is exactly what was required to prove. □

B. Experimental details

B.1. Disentanglement Score

As introduced in the paper, for disentangled representations, single latent variables should be sensitive to individual generating factors and insensitive to all others. To quantify this behavior, for each generating factor w_i, all latent variables are evaluated for their sensitivity to w_i. The sensitivity difference between the two most responsive variables then reflects both desired properties; the sensitivity of the associated best matching latent variable and also the insensitivity of all others. A set of quantities capturing disentanglement can therefore be described as

$$\text{Disent.} = \frac{1}{N_{\text{labels}}} \sum_{i=1}^{N} \left(\frac{A_{i,m(i)} - A_{i,s(i)}}{M_i} \right) \tag{S35}$$

$$\text{for} \quad m(i) = \arg\max_l \left(A_{i,l}\right) \tag{S36}$$

$$\text{for} \quad s(i) = \arg\max_{k \neq m(i)} \left(A_{i,k}\right), \tag{S37}$$

where $A_{i,j}$ is some sort of sensitivity measure of latent variable z_j with respect to the generating factor w_i and M_i is a normalization constant, ensuring the summands fall into the interval $(0, 1)$.

The recently proposed Mutual Information Gap (MIG) [1] uses the Mutual Information as a measure of how the latent variables depend on the generating factors. For the normalisation, the entropy of the generating factor is used.

$$A_{i,j} = \mathrm{MI}(w_i, z_j) \tag{S38}$$

$$M_i = \mathrm{H}(w_i) \tag{S39}$$

For discrete generating factors $\{w_i\}$, the normalization with the entropy $H(w_i)$, binds the MIG to the $(0, 1)$ interval, as expected. For continuous generating factors on the other side, this does not hold. In fact, differential entropy can be zero or even negative and no good normalization is possible.

To treat this shortcoming, we introduce the slightly modified *Disentanglement score* such that it comprises continuous and discrete variables alike. Rather than using mutual information measurements, we employ powerful nonlinear regressors and classifiers for the two different classes of latent variables. The predictability of a generating factor from a given latent coordinate indirectly reflects how much information the two share.

Accordingly, we define the Disentanglement score as in Equation S35 by defining $A_{i,j}$ as the prediction performance of the regressor/classifier for predicting generating factor w_i from the latent coordinate z_j. The normalization factor is then the performance of the best constant classifier/regressor. In case of regression with mean square error, this is simply the standard deviation of the generative factor.

More precisely,

$$A_{i,j} = \begin{cases} \sqrt{\mathrm{var}(w_i)} - \sqrt{\mathrm{mse}_{z_j \to w_i}}, & \text{for regression} \\ \mathrm{accuracy}_{z_j \to w_i}, & \text{for classification} \end{cases} \tag{S40}$$

and

$$M_i = \begin{cases} \sqrt{\mathrm{var}(w_i)}, & \text{for regression.} \\ \mathrm{accuracy}^{\text{const}}_{z_j \to w_i}, & \text{for classification.} \end{cases}$$

We used the SciPy implementation of a k-nearest-neighbors classifier and regressor with default settings (e.g. $k = 5$) to measure the Disentanglement Score. The regressor/classifier was trained on 80% of the test data and evaluated on the remaining 20%.

B.2. DtO via Integer Programming

The *Distance to Orthogonality* (DtO) describes the Frobenius norm of the difference between a matrix V and its closest signed permutation matrix $P(V)$. Using mixed-integer linear programming (MILP) formulation, we find the closest permutation matrix as the optimum P^* of the

Table S1. Overview over the used datasets and network architectures. The nonlinearities are only applied in the hidden layers. Biases are used for all datasets.

	Optimizer (LR)	**Architecture**	**Latent Dim.**	**Epochs**	β
dSprites	AdaGrad (10^{-2})	**Enc:** $1200 - 1200$ (Relu) **Dec:** $1200 - 1200 - 1200$ (Tanh)	5	50	4
Synth. Lin.	Adam (10^{-3})	**Enc:** No hidden Layers (Lin) **Dec:** No hidden Layers (Lin)	2	600	10^{-4}
Synth. Non-Lin.	Adam (10^{-3})	**Enc:** $60 - 40 - 20$ (Tanh) **Dec:** $60 - 40 - 20$ (Tanh)	2	600	10^{-3}
MNIST	AdaGrad (10^{-2})	**Enc:** 400 (Relu) **Dec:** $500 - 500$ (Tanh)	6	400	1
fMNIST	AdaGrad (10^{-2})	**Enc:** 400 (Relu) **Dec:** $500 - 500$ (Tanh)	6	500	1
CelebA	Adam (10^{-4})	Conv/Deconv: [number of kernels, kernel size, stride] **Enc:** [[32, 4, 2], [32, 4, 2], [64, 4, 2], [64, 4, 2]] (Relu) **Dec:** [[64], [64, 4, 2], [32, 4, 2], [32, 4, 2], [3, 4, 2]] (Relu), first layer is connecting MLP	32	50	4

following optimization problem

$$\begin{aligned} \min_P \sum_{i,j} |V_{i,j} - P_{i,j}| & \qquad \text{(S41)} \\ \text{s.t.} \quad P_{i,j} \in \{-1, 0, 1\} & \qquad \forall\,(i,j) \\ \sum_i |P_{i,j}| = 1 & \qquad \forall\, j \\ \sum_j |P_{i,j}| = 1 & \qquad \forall\, i. \end{aligned}$$

Producing a clean MILP formulation, with purely linear objective and binary integer values, can be achieved with a standard technique; introducing new variables. In particular, we set

$$\begin{aligned} P_{i,j} &= P^+_{i,j} - P^-_{i,j} \qquad \text{(S42)} \\ \text{for} \quad P^+_{i,j}, P^-_{i,j} &\in \{0, 1\} \quad \forall\,(i,j) \end{aligned}$$

and introduce (continuous) variables for the differences $V_{i,j} - P_{i,j}$

$$\begin{aligned} V_{i,j} - P_{i,j} \le D_{i,j} & \qquad \forall\,(i,j) \qquad \text{(S43)} \\ P_{i,j} - V_{i,j} \le D_{i,j} & \qquad \forall\,(i,j). \end{aligned}$$

The final formulation then is

$$\begin{aligned} \min_P \sum_{i,j} D_{i,j} & \qquad \text{(S44)} \\ \text{s.t.} \quad (P^+_{i,j} - P^-_{i,j}) - V_{i,j} \le D_{i,j} & \qquad \forall\,(i,j) \\ V_{i,j} - (P^+_{i,j} - P^-_{i,j}) \le D_{i,j} & \qquad \forall\,(i,j) \\ \sum_i \left(P^+_{i,j} + P^-_{i,j}\right) = 1 & \qquad \forall\, j \\ \sum_j \left(P^+_{i,j} + P^-_{i,j}\right) = 1 & \qquad \forall\, i. \end{aligned}$$

B.3. β-VAE with Full Covariance Matrix

In the derivation of the VAE loss function, the approximate posterior is set to be a multivariate normal distribution with a diagonal covariance matrix. The claim of the paper is that this diagonalization is responsible for the orthogonalization. As one of the control experiments in Section 5 we also implemented VAE with a full covariance matrix.

Two issues now need to be addressed; computing KL divergence in closed form and adapting the reparametrization trick. Regarding the former, the sought identity is

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{N}(\mu, \Sigma) \parallel \mathcal{N}(0, \mathcal{I}_k)) = & \qquad \text{(S45)} \\ \frac{1}{2}\left(\|\mu\|^2 + \operatorname{tr}(\Sigma) - \log\det\Sigma - k\right). & \qquad \text{(S46)} \end{aligned}$$

As for the reparametrization trick, if $\varepsilon \sim \mathcal{N}(0, \mathcal{I}_k)$, it is easy to check that

$$\mu + \Sigma^{1/2}\varepsilon \sim \mathcal{N}(\mu, \Sigma), \qquad \text{(S47)}$$

where $\Sigma = \Sigma^{1/2} \cdot \left(\Sigma^{1/2}\right)^\top$ is the unique Cholesky decomposition of the positive definite matrix Σ.

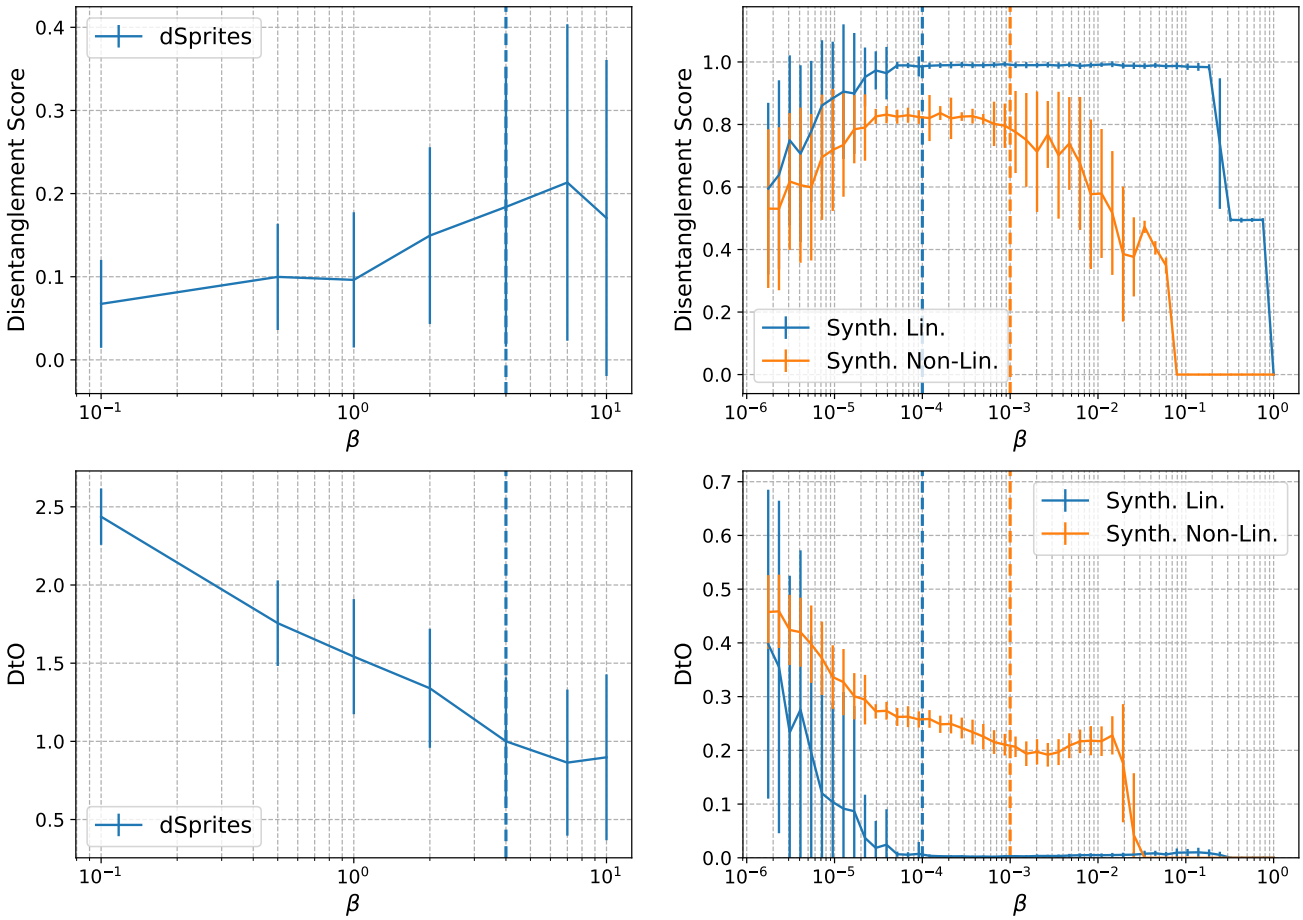

Figure S1. The β hyper-parameter in the β-VAE allows to trade-off reconstruction error and the KL loss such that the desired amount of disentanglement is achieved. The plots show the Disentanglement Score (top) and the DtO (bottom) for dSprites (left) and synthetic datasets (right). The dashed lines indicate the parameter chosen for the experiments.

B.4. Network Details and Training

Table S1 contains the training parameters used for the different architectures. The listed latent dimension is chosen to be the number of independent generating factors, if applicable, and chosen large enough to ensure decent reconstruction loss on all architectures.

All reported numbers are calculated using a previously unseen test dataset. To facilitate this, we split the whole datasets randomly into three parts for training, evaluation and test (containing $80\,\%$, $10\,\%$ and $10\,\%$ of all samples respectively). During development, we use the evaluation dataset, for the final reports we use the test dataset.

B.5. Synthetic Datasets

The linear synthetic dataset is generated with a transformation $f_{\text{lin}}\colon \mathbb{R}^2 \to \mathbb{R}^3$, mapping a unit square $V = [0, 1]^2$ to a 3-dimensional space. The transformation can be decomposed into:

1. stretching along one axis by a fixed factor of 2,
2. trivial embedding into $\mathbb{R}^3$,
3. rotation of $45°$ along the line containing the vector $(1, -1, 1)$.

For the non-linear dataset, the transformation $f_{\text{non-lin}}\colon \mathbb{R}^2 \to \mathbb{R}^6$ is realized by a random initialization of a MLP with one hidden layer (width 10), biases and $\tanh$ nonlinearitites.

Both datasets consist of 50000 samples.

C. Additional Experiments

C.1. Dependence of Disentanglement Score and DtO on β

The choice of β depends on the achievable Disentanglement Score. Figure S1 shows a more thorough analysis of the dependence of both the Disentanglement Score and the

DtO. For too small values of β, the effect of the KL term (and thus the orthogonalization) is negligible. In the other extreme case, too large values values of β result in over-pruning, such that the number of active latent coordinates drops below the number of generating factors.

C.2. Degenerate case

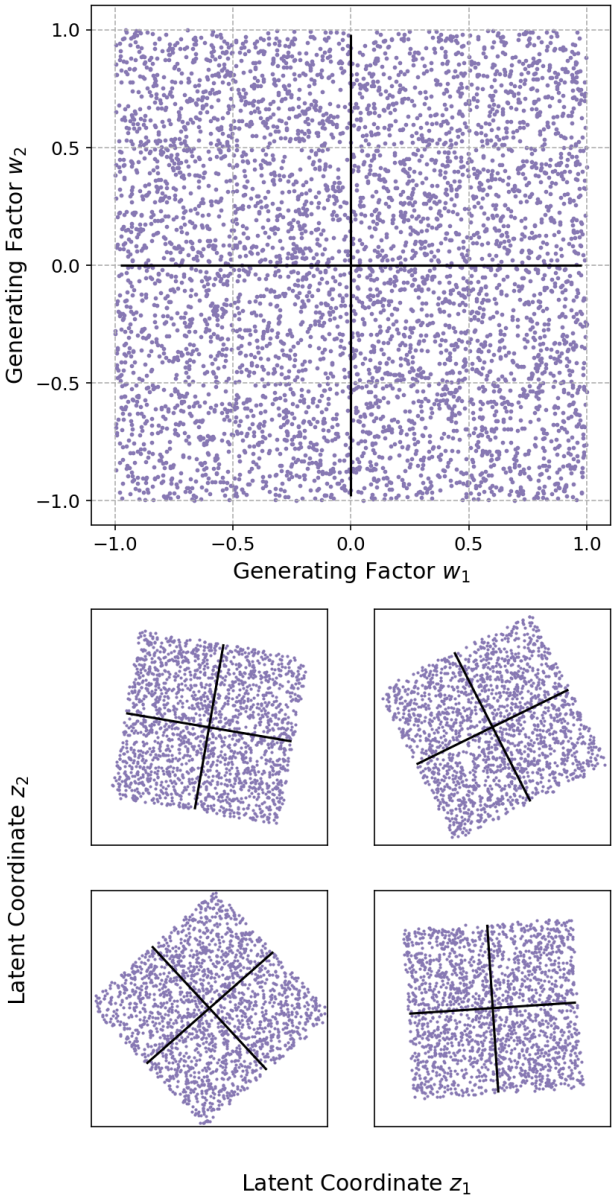

Figure S2. For strong degeneracy, e.g. in the synthetic dataset with the two generating factors w_1 and w_2 on equal, uniform scale (top), the linear β-VAE generates arbitrarily rotated latent representations (bottom) here for the linear synthetic dataset.

Proposition 1 insists that the locally linearized decoder have distinct singular values, otherwise orthogonality of the column vectors does not translate into preserving axes. Here, we design an experiment showing, that this condition is also relevant in practice.

Table S2. Overview of Disentanglement Score and DtO for different ratios of importance between the generating factors for the Synth. Lin. task. A ratio of 1.2 means one generating factor is scaled by 1.2.

Ratio	1.0	1.2	1.5
Disent.	0.51 ± 0.28	0.76 ± 0.25	0.98 ± 0.06
DtO	0.49 ± 0.32	0.20 ± 0.24	0.01 ± 0.06

The dataset in question will be a version of the linear synthetic task where the generating factors have the same scaling, as visualized in the upper plot of Figure S2. Note that any linear encoder applying a simple rotation has both orthogonal columns and equal singular values. But it does not respect the alignment of the original square, as it does not meet the assumptions of Proposition 1.

Behavior of the β-VAE with a linear encoder/decoder network is consistent with this. The bottom part of Figure S2 shows β-VAE latent representations of four random restarts; they expose random alignments. The same effect results in high variances for both the Disentanglement Score and the DtO, as shown in Table S2.

This degeneracy also occurs for PCA. It is easy to check that *any* projection of a unit square on a line *has equal variance*. Hence the greedy PCA algorithm has no preference over which alignment to choose, and the practical choice of alignment is implementation dependent.

This insight reinforces our point that β-VAE (just like PCA) looks for sources of variance rather than for statistical independence.

We can also see in Table S2, that the degeneracy disappears even for small rescaling of the ground truth factors. Since β-VAE promotes normalized latent representations (zero mean, unit variance), the singular values will no longer be equal and the right alignment is found. The same is true for PCA.

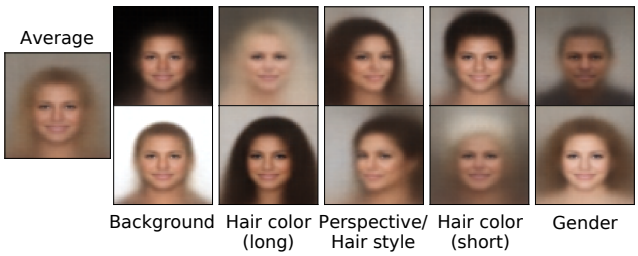

Figure S3. For strong degeneracy, e.g. in the synthetic dataset with the two generating factors w_1 and w_2 on equal, uniform scale (top), the linear β-VAE generates arbitrarily rotated latent representations (bottom) here for the linear synthetic dataset.

C.3. Non-Linear VAE Eigenfaces

In order to highlight the connection with PCA, we use β-VAE to produce a non-linear version of the classical eigenfaces [5] on the CelebA dataset [3]. Fig S3 shows a discrete latent traversal.

Starting from the latent representation z_{mean} of the mean face (over 300 randomly selected datapoints) we feed $\{\mathbf{z}_{\text{mean}} \pm \alpha \mathbf{e}_i\}$ through the decoder, where $\mathbf{e}_i$ are the canonical base vectors. Particularly, we chose i covering the first 5 latent coordinates, sorted by the mean σ_j. The parameter $\alpha = 2.5$ was empirically chosen to be on near the tails of the distribution over $\mathbf{z}^k$.

We can see that unlike classical eigenfaces that mostly reflect photometric properties, the 'nonlinear eigenfaces" capture also semantic features of the data. Note also that the ordering of the 'principal components' by the mean values of σ_j is naturally justified by our work. As was illustrated in Sec. 4.2 of the paper, the first β-VAE 'principle components' also focus on characteristics with high impact on the reconstruction loss (i.e. capture the most variance),

Details about the architecture used are listed in Tab. S1.

References

[1] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *ArXiv e-prints*, abs/1802.04942, 2018. 5

[2] D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *ICLR*, 2014. 4

[3] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 9

[4] L. N. Trefethen and D. Bau III. *Numerical linear algebra*, volume 50. Siam, 1997. 1

[5] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991. 9