Demystifying Inductive Biases for (Beta-)VAE Based Architectures

Dominik Zietlow [1] **Michal Rolínek** [1] **Georg Martius** [1]

Abstract

The performance of β-Variational-Autoencoders (β-VAEs) and their variants on learning semantically meaningful, disentangled representations is unparalleled. On the other hand, there are theoretical arguments suggesting the impossibility of unsupervised disentanglement. In this work, we shed light on the inductive bias responsible for the success of VAE-based architectures. We show that in classical datasets the structure of variance induced by the generating factors is conveniently aligned with the latent directions fostered by the VAE objective. This builds the pivotal bias on which the disentangling abilities of VAEs rely. By small, elaborate perturbations of existing datasets, we hide the convenient correlation structure that is easily exploited by a variety of architectures. To demonstrate this, we construct modified versions of standard datasets in which (i) the generative factors are perfectly preserved; (ii) each image undergoes a mild transformation causing a small change of variance; (iii) the leading **VAE-based disentanglement architectures fail to produce disentangled representations while the performance of a non-variational method remains unchanged**.

1. Introduction

The task of unsupervised learning of *interpretable* data representations has a long history. From classical approaches using linear algebra e.g. via Principal Component Analysis (PCA) (Pearson, 1901) or statistical methods such as Independent Component Analysis (ICA) (Comon, 1994) all the way to more recent approaches that rely on deep learning architectures.

The cornerstone architecture is the Variational Autoencoder (Kingma & Welling, 2014) (VAE) which clearly demonstrates both high semantic quality as well as good performance in terms of *disentanglement*. Until today, derivates of VAEs (Higgins et al., 2017; Kim & Mnih, 2018a; Chen et al., 2018; Kumar et al., 2017; Klindt et al., 2021) excel over other architectures in terms of disentanglement metrics. The extent of the VAE's success even prompted recent deeper analyses of its inner workings (Rolinek et al., 2019; Burgess et al., 2018; Chen et al., 2018; Mathieu et al., 2018).

If we treat the overloaded term disentanglement to the highest of its aspirations, as the ability to recover the *true generating factors* of data, fundamental problems arise. As explained by Locatello et al. (2019), already the concept of generative factors is compromised from a statistical perspective: two (or in fact infinitely many) sets of generative factors can generate statistically indistinguishable datasets. Yet, the scores on the disentanglement benchmarks are high and continue to rise. This apparent contradiction stems from biases present in used datasets, metrics, and architectures. It was concluded in Locatello et al. (2020) that

> *[...] future work on disentanglement learning should be explicit about the role of inductive biases and (implicit) supervision [...].*

which did not happen for the majority of existing unsupervised approaches. We close this gap for VAE-based architectures on the two most common datasets, namely dSprites (Matthey et al., 2017) and Shapes3d (Burgess & Kim, 2018).

The main hypothesis of this work is that all unsupervised, VAE-based disentanglement architectures are successful because they exploit the same structural bias in the data. The ground truth generating factors are well aligned with the nonlinear principal components that VAEs strive for. This bias can be reduced by introducing a **small change of the local correlation structure** of the input data, which, however, **perfectly preserves the set of generative factors**. We evaluate a set of approaches on slightly modified versions of the two leading datasets in which each image undergoes a modification inducing little variance. We report drastic drops of disentanglement performance on the altered datasets.

On a technical level, we build on the findings by Rolinek et al. (2019) who argued that VAEs recover the *nonlinear principal components* of the data. In other words, they recover a set of scalars that embody the sources of variance

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany. Correspondence to: Dominik Zietlow <dzietlow@tue.mpg.de>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

through a nonlinear mapping, similarly to PCA in the linear setting. We extend their argument by an additional finding that further strengthens this connection. The small modifications of the datasets we propose aim to change the leading principal components by adding modest variance to a set of alternative candidates. The "to-be" leading principal components are specific to each dataset, but they are automatically determined in a consistent fashion.

2. Related work

The related work can be categorized into three research questions: i) defining disentanglement and metrics capturing the quality of latent representations; ii) architecture development for unsupervised learning of disentangled representations; and iii) understanding the inner workings of existing architectures, as for example of β-VAEs. This paper is built upon results from all three lines of work.

Defining disentanglement. Defining the term *disentangled representation* is an open question (Higgins et al., 2018). The presence of learned representations in machine learning downstream tasks, such as object recognition, natural language processing, and others, created the need to *"disentangle the factors of variation"* (Bengio et al., 2013) early on. This vague interpretation of disentanglement is inspired by the existence of a low-dimensional manifold that captures the variance of higher dimensional data. As such, finding a factorized, statistically independent representation became a core ingredient of disentangled representation learning and dates back to classical ICA models (Comon, 1994; Bell & Sejnowski, 1995).

For some tasks, the desired feature of a disentangled representation is that it is *semantically meaningful*. Prominent examples can be found in computer vision (Shu et al., 2017; Liao et al., 2020) and in research addressing the interpretability of machine learning models (Adel et al., 2018; Kim, 2019).

Based on group theory and symmetry transformations, Higgins et al. (2018) provides the *"first principled definition of a disentangled representation"*. Closely related to this concept is also the field of causality in machine learning (Schölkopf, 2019; Suter et al., 2019), more specifically the search for causal generative models (Besserve et al., 2018; 2020). In terms of implementable metrics, a variety of quantities have been introduced, such as the β-VAE score (Higgins et al., 2017), SAP score (Kumar et al., 2017), DCI scores (Eastwood & Williams, 2018) and the Mutual Information Gap (MIG, Chen et al. (2018)).

Architecture development. The leading architectures for disentangled representation learning are based on VAEs (Kingma & Welling, 2014). Despite originally developed as a generative modeling architecture, its variants have proven to excel at representation learning tasks. In particular, the β-VAE performs remarkably well. It exposes the trade-off between reconstruction and regularization via an additional hyperparameter. Other architectures have been proposed that additionally encourage statistical independence in the latent space, e.g. FactorVAE (Kim & Mnih, 2018b) and β-TC-VAE (Chen et al., 2018). The DIP-VAE (Kumar et al., 2017) suggests using moment-matching to close the distribution gap introduced in the original VAE paper. Using data with auxiliary labels, e.g. time indices of time series data, for which the conditional prior latent distribution is factorized, allowed Khemakhem et al. (2020) to circumvent the unidentifiability of previous models. Similarly, Klindt et al. (2021) used a sparse temporal prior to develop an identifiable model that also performs well on natural data. In this work, we also compare to representations learned by Permutation Contrastive Learning (PCL) (Hyvarinen & Morioka, 2017). This non-variational method conducts nonlinear ICA also assuming temporal dependencies between the sources of variance. The PCL objective is based on logistic regression. Another approach utilizes weak supervision on GANs to achieve disentangled representations in their underlying latent space (Shu et al., 2019).

Understanding inner workings. With the rising success and development of VAE based architectures, the question of understanding their inner working principles became dominant in the community. One line of work tries to answer the question why these models disentangle at all (Burgess et al., 2018). Another closely related line of work showed the tight connection between the vanilla (β-)VAE objective and (probabilistic) PCA (Tipping & Bishop, 1999) (Rolinek et al., 2019; Lucas et al., 2019). The role of the regularization in β-VAEs was explicitly investigated in (Kumar & Poole, 2020). Building on these findings, novel approaches for model selection were proposed (Duan et al., 2020), emphasizing the value of thoroughly understanding these methods. On a less technical side, Locatello et al. (2019) conducted a broad set of experiments, questioning the relevance of the specific model architecture compared to the choice of hyperparameters and the variance over restarts. They also formalized the necessity of inductive biases as a strict requirement for unsupervised learning of disentangled representations. Our experiments are built on their codebase.

3. Background

3.1. Quantifying Disentanglement

Among the different viewpoints on disentanglement, we follow the recent literature and focus on the connection between the discovered data representation and a set of *generative factors*.

Multiple metrics have been proposed to quantify this con-

nection. Most of them are based on the understanding that, ideally, each generative factor is encoded in precisely one latent variable. This was captured concisely by Chen et al. (2018), who proposed the Mutual Information Gap (MIG) – the mean difference (over the N_w generative factors) of the two highest mutual information between a latent coordinate and the single generating factor, normalized by its entropy. For the entropy $H(w_i)$ of a generating factor and the mutual information $I(w_i; z_k)$ between a generating factor and a latent coordinate, the MIG is defined as

$$\frac{1}{N_w} \sum_{i=1}^{N_w} \frac{1}{H(w_i)} \left(\max_k I\left(w_i; z_k\right) - \max_{k \neq k'} I\left(w_i; z_k\right) \right), \tag{1}$$

where $k' = \arg\max_\kappa I\left(w_i, z_\kappa\right)$. More details about MIG, its implementation, and an extension to discrete variables can be found in (Chen et al., 2018; Rolinek et al., 2019). Multiple other metrics were proposed such as SAP score (Kumar et al., 2017), FactorVAE score (Kim & Mnih, 2018a) and DCI score (Eastwood & Williams, 2018) (see the supplementary material of Klindt et al. (2021) for extensive descriptions).

3.2. Variational Autoencoders and the Mystery of a Specific Alignment

Variational autoencoders hide many intricacies and attempting to compress their exposition would not do them justice. For this reason, we limit ourselves to what is crucial for understanding this work: the objective function. For a well-presented description of VAEs, we refer the reader to (Doersch, 2016).

As is common in generative models, VAEs aim to maximize the log-likelihood objective

$$\sum_{i=1}^{N} \log p\left(\mathbf{x}^{(i)}\right), \tag{2}$$

in which $\{\mathbf{x}^{(i)}\}_{i=1}^N = \mathcal{X}$ is a dataset consisting of N i.i.d. samples $\mathbf{x}^{(i)}$ of a multivariate random variable $\mathbf{X}$ that follows the true data distribution. The quantity $p(\mathbf{x}^{(i)})$ captures the probability density of generating the training data point $\mathbf{x}^{(i)}$ under the current parameters of the model. This objective is, however, intractable in its general form. For this reason, Kingma & Welling (2014) follow the standard technique of variational inference and introduce a tractable Evidence Lower Bound (ELBO):

$$\mathbb{E}_{q(\mathbf{z}|\mathbf{x}^{(i)})} \log p\left(\mathbf{x}^{(i)} \mid \mathbf{z}\right) + D_{\mathrm{KL}}\Big(q(\mathbf{z} \mid \mathbf{x}^{(i)}) \parallel p(\mathbf{z})\Big). \tag{3}$$

Here, $\mathbf{z}$ are the latent variables used to generate samples from $\mathbf{X}$ via a parameterized stochastic decoder $p(\mathbf{x}^{(i)} \mid \mathbf{z})$.

The fundamental question of *"How do these objectives promote disentanglement?"* was first asked by Burgess et al. (2018). This is indeed *far from obvious*; in disentanglement the aim is to encode a fixed generative factor in *precisely* one latent variable. From a geometric viewpoint, this requires the latent representation to be **axis-aligned** (one axis corresponding to one generative factor). This question becomes yet more intriguing after noticing (and formally proving) that both objective functions (2) and (3) are *invariant under rotations* for rotationally symmetric latent space priors, as the ubiquitous $p(\mathbf{z}) = \mathcal{N}(0, \mathbb{1})$ (Burgess et al., 2018; Rolinek et al., 2019). In other words, any rotation of a fixed latent representation results in the same value of the objective function and yet β-VAEs consistently produce representations that are axis-aligned and in effect are isolating the generative factor into individual latent variables.

3.3. Resolution via Nonlinear Connections to PCA

A mechanistic answer to the question raised in the previous subsection was given by Rolinek et al. (2019). The formal argument showed that under specific conditions which are typical for β-VAEs (called *polarized regime*), the datapoint-wise linearization of the model performs PCA in the sense of aligning the "sources of variance" with the local axes. **The resulting alignment often coincides with finding the components of the datasets ground truth generating factors**. Fig. 1 illustrates the difference between local and global PCA. Note that the principal directions of a non-degenerate uniform distribution are the Cartesian axes. PCA as a linear transformation is aligning the embedding following the overall (global) variance. Nonlinear VAEs are aligning the latent space according to the local structure (the local principal components of the almost uniform clusters). This behavior stems from the convenient but uninformed choice of a *diagonal posterior*, which breaks the symmetry of (2) and (3). This connection with PCA was also reported by Stuehmer et al. (2020), alternatively formalized by Lucas et al. (2019) and converted into performance improvements in an unsupervised setting by Duan et al. (2020). Strictly speaking, the formal statements of Rolinek et al. (2019) are limited and only claim that β-VAEs strive for local orthogonality which, in the linear case, is a strong similarity to PCA.

4. Methods

We first tighten the connection between VAEs and PCA, secondly introduce the general data generation scheme of commonly used disentanglement datasets, and lastly turn this understanding into an experimental setup that allows for empirical confirmation that the success of VAE based architectures mostly relies on the local structure of the data.

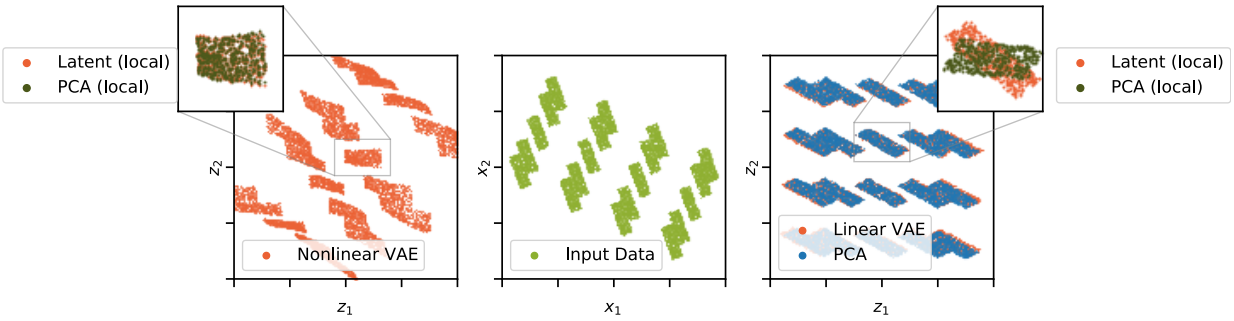

Figure 1: Distribution of latent encodings for an input distributed as depicted in the middle (data dimensionality equals latent dimensionality). The linear VAE's encoding matches the PCA encoding remarkably well (right); both focus on aligning with axes based on the global variance. The nonlinear VAE (left) is, however, more sensitive to local variance. It picks up on the natural axis alignment of the microscopic structure. The insets show the enlarged area and PCA performed only on the local subset of the point cloud. Our argument in this work is that misaligning the microscopic structure with respect to the ground truth generating factors leads to decreased *convenient* bias in the data.

4.1. Theoretical Support of the Connection to PCA

The argument established by Rolinek et al. (2019) is technically incomplete to justify the equivalence of linear VAEs and PCA. Strictly speaking, the core message of that work is that VAE decoders tend to be locally orthogonal. The actual alignment of the latent space is insufficiently described by that finding. However, Lucas et al. (2019) argue for the similarity of linear VAEs to probabilistic PCA. We now show a more technical connection between classical PCA and linear VAEs which allows for easier understanding of the consequent subsections. We try to stay close to the language of Rolinek et al. (2019) and partially reuse their arguments.

The canonical implementation of the β-VAE uses a normal posterior with diagonal covariance matrix and a rotationally symmetric $p(\mathbf{z}) = \mathcal{N}(0, \mathbb{1})$ latent prior. This, together with a Gaussian decoder model, turns the ELBO (3) into the tractable loss function

$$\mathcal{L} = \mathbb{E}_i \left(\mathcal{L}^{(i)}_{\text{rec}} + \beta \mathcal{L}^{(i)}_{\text{KL}} \right) \tag{4}$$

$$\mathcal{L}_{\text{rec}} = \left\| \text{Dec}_\theta(\text{Enc}_\varphi(\mathbf{x}^{(i)})) - \mathbf{x}^{(i)} \right\|^2$$
$$\mathcal{L}_{\text{KL}} = \frac{1}{2} \sum_j \left({\boldsymbol{\mu}^{(i)}}_j^2 + {\boldsymbol{\sigma}^{(i)}}_j^2 - \log({\boldsymbol{\sigma}^{(i)}}_j^2) - 1 \right)$$

for an encoder Enc_φ parameterized by φ, a decoder Dec_θ parameterized by θ, and $\mathbf{z}^{(i)} = \text{Enc}_\varphi(\mathbf{x}^{(i)}) = \boldsymbol{\mu}^{(i)}(\mathbf{x}^{(i)}) + \boldsymbol{\varepsilon}^{(i)}$, $\boldsymbol{\varepsilon}^{(i)} \sim \mathcal{N}(0, {\boldsymbol{\sigma}^{(i)}}^2(\mathbf{x}^{(i)}))$. Since $\mathbf{z}^{(i)}$ is unbiased around $\boldsymbol{\mu}^{(i)}(\mathbf{x}^{(i)})$, we find that

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_i \left(\mathcal{L}^{\mu}_{\text{rec}}(\mathbf{x}^{(i)}) + \mathcal{L}^{\text{stoch}}_{\text{rec}}(\mathbf{x}^{(i)}) \right) \tag{5}$$
$$\mathcal{L}^{\text{stoch}}_{\text{rec}}(\mathbf{x}^{(i)}) = \left\| \text{Dec}_\theta(\text{Enc}_\varphi(\mathbf{x}^{(i)})) - \text{Dec}_\theta(\boldsymbol{\mu}^{(i)}) \right\|^2$$
$$\mathcal{L}^{\mu}_{\text{rec}}(\mathbf{x}^{(i)}) = \left\| \text{Dec}_\theta(\boldsymbol{\mu}^{(i)}) - \mathbf{x}^{(i)} \right\|^2 .$$

We hereby assume linear models $\boldsymbol{\mu}^{(i)} = M_E \mathbf{x}^{(i)}$, $\text{Dec}_\theta(\mathbf{z}^{(i)}) = M_D \mathbf{z}^{(i)}$ and denote the SVD decomposition of M_D as $M_D = U \Sigma V^\top$.

We can now state a constraint optimization problem of a simplified VAE objective as

$$\min_{\Sigma, U, V} \mathbb{E}_i \left(\left\| U \Sigma V^\top \boldsymbol{\varepsilon}^{(i)} \right\|^2 \right) \tag{6}$$
$$\text{s.t.} \quad \mathbb{E}_i \left(\mathcal{L}^{(i)}_{\approx \text{KL}} \right) = c_{\approx \text{KL}}. \tag{7}$$

where only the stochastic part of the reconstruction loss is minimized and $c_{\approx \text{KL}}$ is a constant. The term $\mathcal{L}_{\approx \text{KL}}$ is the KL loss in the polarized regime, where ${\boldsymbol{\sigma}^{(i)}}^2 \ll -\log({\boldsymbol{\sigma}^{(i)}}^2)$ (element-wise):

$$\mathcal{L}_{\approx \text{KL}} = \sum_j \left({\boldsymbol{\mu}^{(i)}}_j^2 - \log({\boldsymbol{\sigma}^{(i)}}_j^2) \right). \tag{8}$$

The 'decoder matrix' of the classical PCA contains the eigenvectors of the covariance matrix C. By SVD decomposing the zero-mean data matrix $X = U_X \Sigma_X V_X^\top$, we find

$$C = X^\top X = V_X \Sigma_X^2 V_X^\top. \tag{9}$$

For encoding data with PCA, the eigenvectors of V_X are typically sorted according to their eigenvalue by a permutation matrix P, which leads to the PCA decoder as

$$M_{\text{PCA}} = V_X^\top \Sigma_X^2 P. \tag{10}$$

To tighten the connection between VAEs and PCA, we compare $M_D = U \Sigma V^\top$ to $M_{\text{PCA}} = V_X^\top \Sigma_X^2 P$.

Theorem 1 (Linear VAEs perform PCA) *In a setting that precisely isolates the freedom in choosing U, Σ, and V, and under mild non-degeneracy assumptions (full description is available in the supplementary material), the*

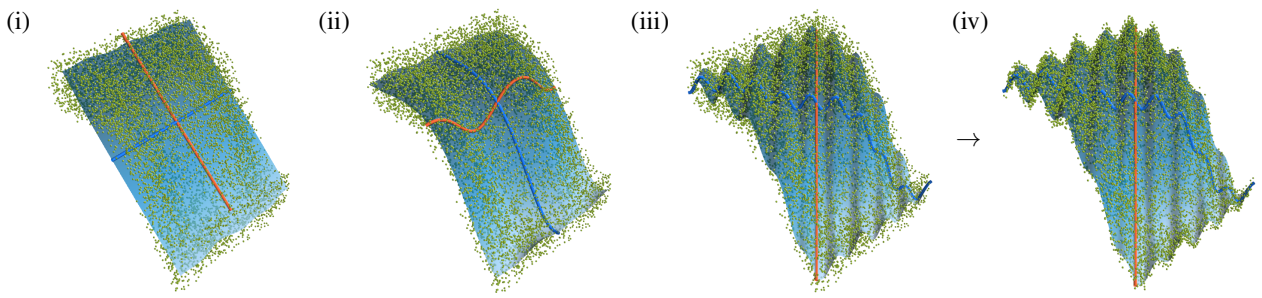

Figure 2: Illustrations for linear and nonlinear embeddings. From left to right: (i) a 3 dimensional point cloud and the corresponding two-dimensional PCA manifold (blue surface) with the canonical principal components (red/blue curves), (ii) a nonlinear two-dimensional manifold with a latent traversal, (iii) a locally perturbed two-dimensional manifold with its principal components which are rotated with respect to (ii), (iv) the goal of our modifications is to move each datapoint closer to this *entangled* manifold.

following holds: For any $X \in \mathbb{R}^{n\times m}$, the solution to (6, 7)

$$\Sigma^\star, U^\star, V^\star = \underset{\Sigma,U,V}{\arg\min}\, \mathbb{E}_i \left(\left\| U\Sigma V^\top \boldsymbol{\varepsilon}^{(i)} \right\|^2 \right), \qquad (11)$$

satisfies (in a "PCA-like" way)

$$V^\star \quad \text{is a signed permutation matrix},$$
$$U^\star = V_X^\top.$$

It was known for long that linear autoencoders, trained on L^2 reconstruction loss, span the same space as PCA (Bourlard & Kamp, 1988; Baldi & Hornik, 1989). The additional similarity that VAEs produce orthogonal mappings, like PCA, was presented by (Rolinek et al., 2019). With the final connection presented here, even the alignment of the embedding is shown to be identical. For the sake of brevity, the proofs of the statements can be found in the supplementary material.

Although this does not directly translate to a universal statement about the linearization of a nonlinear model, it provides an intuition for that case as well. An important observation is that **the alignment of the latent space is mostly driven by the distribution of the latent noise**. When generalizing this statement to the linearization of a nonlinear decoder, the effect of the noise stays local. As a consequence, local changes of the data distribution can potentially lead to a disruptive change in the latent alignments, without inducing large global variance. This idea is depicted in Fig. 2.

4.2. Linear vs. Nonlinear Embeddings

One less obvious observation is that the "isolation" of different sources of variance relies on the non-linearity of the decoder. The region in which the linearization of the decoder around a fixed $\boldsymbol{\mu}^{(i)}(\mathbf{x}^{(i)})$ is a reasonable approximation suggests a certain radius of the relevant local structure. Since in many datasets the local principal components are well aligned with the intuitively chosen generating factors, β-VAEs recover sound global principal components. If, however, the local structure obeys a different "natural" alignment, the VAE could prefer it, and in return not disentangle the ground truth generating factors.

4.3. The Generative Process

The standard datasets for evaluating disentanglement all have an explicit generation procedure. Each data point $\mathbf{x}^{(i)} \in \mathcal{X}$ is an outcome of a generative process g applied to input $\mathbf{w}^{(i)} \in \mathcal{W}$. Imagine that g is a function rendering a simple scene from its specification w containing *as its coordinates* the background color, foreground color, object shape, object size, etc. By design, the individual generative factors are statistically independent in $\mathcal{W}$. All in all, the dataset $\mathcal{X} = \left(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(n)}\right)$ is constructed with $\mathbf{x}^{(i)} = g(\mathbf{w}^{(i)})$, where g is a mapping from the generative factors to the corresponding data points.

In this paper, we design a modification $\widetilde{g}$ of the generative procedure g that changes the local structure of the dataset $\mathcal{X}$, whilst barely distorts each individual data point. In particular, for each $\mathbf{x}^{(i)} \in \mathcal{X}$, we have under some distance measure $d(\cdot,\cdot)$, that

$$d\big(\mathbf{x}^{(i)}, \widetilde{g}(\mathbf{w}^{(i)})\big) \leq \varepsilon. \qquad (12)$$

How to design $\widetilde{g}$ such that despite an ε-small modification, VAE-based architectures will create an entangled representation? Following the intuition from Sec. 3.3, Fig. 1 and Fig. 2, we *misalign* the local variance with respect to the generating factors in order to promote an alternative (entangled) latent embedding. This is precisely the step from (iii) to (iv) in Fig. 2.

To avoid hand-crafting this process, we can exploit the fol-

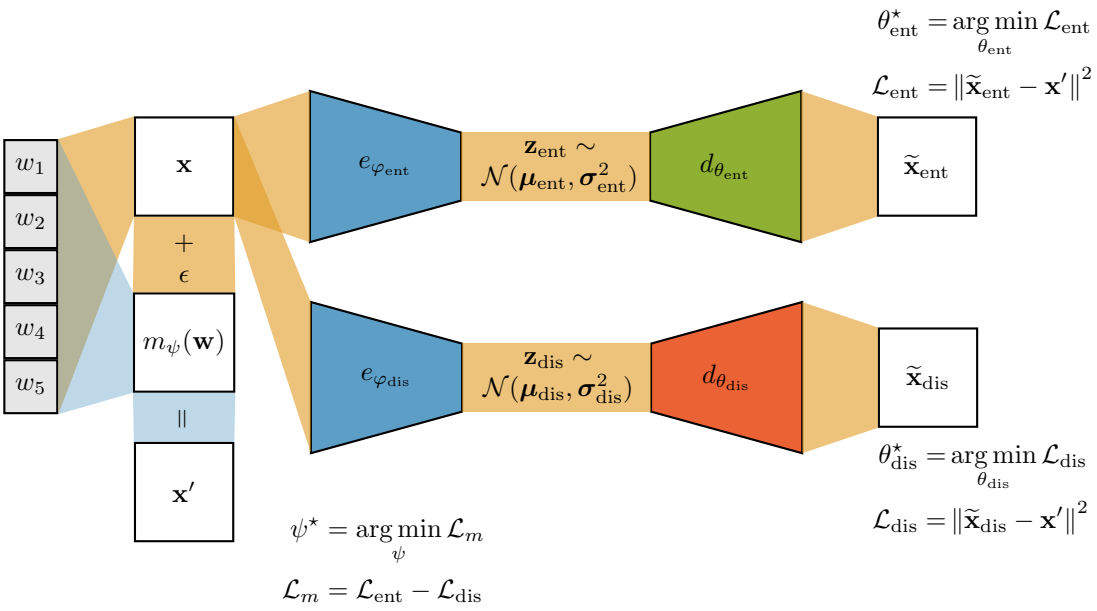

Figure 3: A schematic visualization of the image generation process. Starting from ground truth generating factors $\mathbf{w}$, two β-VAE encoder-decoder pairs are initialized such that one (top) produces entangled and the other (bottom) disentangled representations. Another decoder-like network m is trained to produce additive manipulations to the original images x. The encoders are frozen and fed with the original images. The set of ground truth generating factors $\mathbf{w}$ stays untouched by the modification.

lowing observation. VAE-based architectures suffer from large performance variance over e.g. different random initializations. This hints at an existing ambiguity: two or more candidates for the latent coordinate system are competing minima of the optimization problem. Some of these solutions perform well, others are "bad" in terms of disentanglement – they correspond to (ii) and (iii) in Fig. 2 respectively. Below, we elaborate on how to foster the entangling and diminish the disentangling solutions.

Our modifications are not an implementation of (Locatello et al., 2019, Theorem 1). We **do not modify the set of generative factors, but slightly alter the generating process** to target a specific subtlety in the inner working of VAEs.

Given any dataset, our modification process has three steps:

(i) Find the most disentangled and the most entangled latent space alignment that a β-VAE produces over multiple restarts.

(ii) Optimize a generator that manipulates images to foster and diminish their suitability for the entangled and disentangled model respectively.

(iii) Apply the manipulation to the whole dataset and compare the performance of models trained on the original and the modified dataset.

4.4. Choice of Fostered Latent Coordinate System

Over multiple restarts of β-VAE, we pick the model with the lowest MIG score. This gives us an entangled alignment that is expressible by the architecture. Although any choice of metric is valid for this model selection (e.g. UDR (Duan et al., 2020)), we chose MIG for the sake of simplicity. The latent variables of each of the models capture the nonlinear principal components of the data. Similarly to PCA, we can order them according to the variance they induce. The order is inversely reflected by the magnitude of the latent noise values. We find the j'th principal components $s_j^{(i)}$ as

$$s_j^{(i)}\big(\mathbf{x}^{(i)}\big) = \text{enc}\big(\mathbf{x}^{(i)}\big)_{k^{(j)}} \tag{13}$$

$$k^{(j)} = \underset{l \notin \{k^{(0)}, k^{(1)}, \ldots, k^{(j-1)}\}}{\arg\min} \big\langle \boldsymbol{\sigma}_l^2 \big\rangle. \tag{14}$$

This procedure of sorting the most *important* latent coordinates is consistent with (Higgins et al., 2017) and (Rolinek et al., 2019). The analogy to PCA is that the mapping $s^{(j)}(\mathbf{x}^{(i)})$ gives the j'th coordinate of $\mathbf{x}^{(i)}$ in the new (nonlinear) coordinate system.

4.5. Dataset Manipulations

We will now describe the modification procedure assuming the data points are $r \times r$ images. The manipulated data-point $\mathbf{x}'^{(i)}$ is of the form $\mathbf{x}'^{(i)} = \mathbf{x}^{(i)} + \varepsilon m\big(\mathbf{w}^{(i)}\big)$ where the mapping $m\colon \mathbb{R} \to \mathbb{R}^r \times \mathbb{R}^r$ is constrained by $\|m(\mathbf{w}^{(i)})\|_\infty \leq 1$

for every $\mathbf{w}^{(i)}$. Then inequality (12) is naturally satisfied for the maximum norm.

The abstract idea of how to achieve a change of the latent embedding coordinate systems can be visualized using the intuition following from Eq. (14). We can think of two VAE latent spaces where one is considered disentangled ($\{\boldsymbol{\mu}_{\text{dis}}^{(i)}, \boldsymbol{\sigma}_{\text{dis}}^{(i)}\}$) and the other is entangled ($\{\boldsymbol{\mu}_{\text{ent}}^{(i)}, \boldsymbol{\sigma}_{\text{ent}}^{(i)}\}$), as two sets of nonlinear principal directions, and the variance each of the dimensions capture is reflected in the magnitude of $\boldsymbol{\sigma}^{(i)}$. We are aiming to alter the dataset such that its entangled representation is superior over the disentangled representation, in the sense of being *cheaper* to decode with respect to the reconstruction loss. In other words, projecting the dataset to the manifold supported by $\mathbf{z}_{\text{ent}}^{(i)}$ should result in a lower loss in Eq. (5) than projecting it to the manifold supported by $\mathbf{z}_{\text{dis}}^{(i)}$. A naive way of doing so is by moving each image closer to its projections on the first principal components of the entangled representation and further away from those of the disentangled representation. Instead of hand-crafting this operation, we can optimize for it directly.

This idea can be turned into an end-to-end trainable architecture as depicted in Fig. 3. We want to change the dataset such that it is more convenient to encode it in an entangled way. Starting with two pretrained models, we fix their encoders and keep feeding them the original images. This ensures that the latent encoding stays unchanged, as we want to compare their suitability for reconstruction. The decoders are trained to minimize the reconstruction loss given the entangled representation:

$$\theta_{\text{ent}}^{\star} = \arg\min_{\theta_{\text{ent}}} \mathcal{L}_{\text{rec}}^{\text{ent}}\left(\mathbf{x}'^{(i)}, \mathbf{z}^{(i)}\right),$$
$$\theta_{\text{dis}}^{\star} = \arg\min_{\theta_{\text{dis}}} \mathcal{L}_{\text{rec}}^{\text{dis}}\left(\mathbf{x}'^{(i)}, \mathbf{z}^{(i)}\right).$$

We initialize this network with the parameters of the disentangled model $\theta_{\text{dis}}, \varphi_{\text{dis}}$ and the entangled model $\theta_{\text{ent}}, \varphi_{\text{ent}}$ respectively. We introduce a network to learn the additive manipulation, m_ψ with parameters ψ. The parameters are trained to minimize the reconstruction loss of the entangled VAE and to increase the loss of the disentangled VAE via its effect on the dataset:

$$\psi^{\star} = \arg\min_{\psi}\left(\mathcal{L}_{\text{rec}}^{\text{ent}}\left(\mathbf{x}'^{(i)}, \mathbf{z}^{(i)}\right) - \mathcal{L}_{\text{rec}}^{\text{dis}}\left(\mathbf{x}'^{(i)}, \mathbf{z}^{(i)}\right)\right).$$

It is worth noting that both latent spaces were suitable for reconstructing the images of the original dataset. **The major play that the network m_ψ has, is to utilize the different ways the noise was distributed across the latent space.**

5. Experiments

To experimentally validate the soundness of the manipulations, we need to demonstrate the following:

1. **Effectiveness of manipulations.** Disentanglement metrics should drop on the altered datasets across VAE-based architectures. We do not expect to see changes on non variational methods.
2. **Comparison to a trivial modification.** Instead of the proposed method, we modify with uniform noise of the same magnitude. The disentanglement scores for the algorithms on the resulting datasets should not drop significantly, as this change does not alleviate the existing bias.
3. **Robustness.** The new datasets should be hard to disentangle even after retuning hyperparameters of the original architectures.

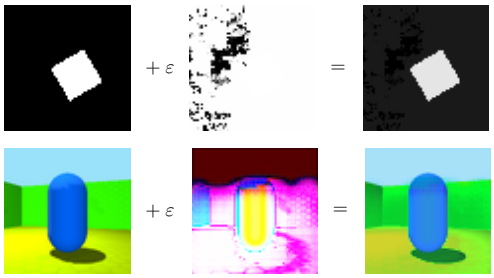

Figure 4: From left to right: Original images, additive manipulations and the altered images. Top row shows an example of dSprites, the bottom for Shapes3D.

5.1. Effectiveness of Manipulations

We deploy the suggested training for the manipulations on two datasets: Shapes3D and dSprites, leading to manipulations as depicted in Fig. 4. In terms of models, we trained four VAE-based architectures (Higgins et al., 2017; Kim & Mnih, 2018a; Chen et al., 2018; Klindt et al., 2021), a regular autoencoder (Hinton & Salakhutdinov, 2006), and (as non-variational methods) PCL (Hyvarinen & Morioka, 2017) as well as the weakly supervised GAN from (Shu et al., 2019) in the full sharing setting. We evaluate on both the original and manipulated datasets. Regularization strengths are used as reported in the literature (or better tuned values), and other hyperparameters are taken from the disentanglement library (Locatello et al., 2019). For the sake of simplicity and clarity, we restricted the latent space dimension to be equal to the number of ground truth generative factors. Most of the architectures have been shown to be capable of pruning the latent space as a consequence of their intrinsic regularization (Stuehmer et al., 2020). While being a perk in real world application scenarios, this behaviour can lead to over- or under-pruning and thereby cloak the

Table 1: MIG scores for unmodified, modified, and noisy datasets. We report the mean and standard deviation over 10 distinct random seeds for each setting. The regular autoencoder serves as a baseline (random alignment). PCL and the weakly supervised GAN from (Shu et al., 2019) are the only disentangling non-variational model. The modification leads to a significant drop in all variational methods.

	dSprites			Shapes3d		
	orig.	**mod.**	**noise**	**orig.**	**mod.**	**noise**
AE	0.09 ± 0.06	0.05 ± 0.02	0.06 ± 0.03	0.06 ± 0.03	0.05 ± 0.03	0.07 ± 0.03
β-VAE	0.23 ± 0.08	0.07 ± 0.09	0.14 ± 0.07	0.60 ± 0.31	0.09 ± 0.14	0.66 ± 0.05
Fac. VAE	0.27 ± 0.11	0.20 ± 0.12	0.16 ± 0.08	0.27 ± 0.18	0.07 ± 0.05	0.33 ± 0.20
TC-β-VAE	0.25 ± 0.08	0.14 ± 0.10	0.20 ± 0.04	0.58 ± 0.20	0.24 ± 0.16	0.60 ± 0.11
Slow-VAE	0.39 ± 0.08	0.27 ± 0.08	0.37 ± 0.09	0.53 ± 0.19	0.13 ± 0.08	0.60 ± 0.10
PCL	0.21 ± 0.03	0.24 ± 0.07	0.24 ± 0.07	0.44 ± 0.06	0.47 ± 0.08	0.40 ± 0.07
Weak sup. GAN	0.45 ± 0.05	0.36 ± 0.02	0.36 ± 0.01	0.69 ± 0.12	0.66 ± 0.12	0.77 ± 0.13

actual difference in the alignment of the latent space. The resulting MIG scores are listed in Tab. 1, other metrics are listed in the supplementary materials. Over all variational models, the disentanglement quality is significantly reduced. Interestingly, even for SlowVAE, an architecture that supposedly circumvents the non-identifiability problem by deploying a sparse temporal prior, the disentanglement reduces. This indicates that the architecture still builds upon the local data structure more than on the weak supervision induced by the temporal sparsity. **PCL and the weakly supervised GAN, as non-variational methods, perform similarly well on the original and the modified architecture**, which is a strong indicator that due to the constraint (12), the main sources of global variance remain unaltered. The modifications indeed only attack the subtle bias VAEs exploit.

5.2. Noisy Datasets

We replace our modification by contaminating each image with uniform pixel-wise noise $[-\varepsilon, \varepsilon]$. The value of ε is fixed to the level of the presented manipulations (0.1 for dSprites and 0.175 for Shapes3D). The results are also listed in Tab. 1. The lack of structure in the contamination does not affect the performance in a guided way and leads to very little effect on Shapes3D. The impact on dSprites is, however, noticeable. Due to the comparatively small variance among dSprites images, the noise conceals the variance from the less important generating factors (such as e.g. orientation).

5.3. Robustness over Hyperparameters

We run a line search over the primary hyperparameter for each architecture. The results are illustrated in Fig. 5. Overall our modifications seem mostly robust for adjusted hyperparameters. Significant increase in the regularization strength allowed for some recovery. More thorough analysis revealed that this effect starts only once the models reach a level of over-pruning, which is a behavior well known to practitioners. We discard the runs that over pruned the latent space (number of active coordinates, i.e. $\mathbb{E}\left(\boldsymbol{\sigma}_i^2\right) < 0.8$, sinks below the dimensionality of the ground truth generating factors). This effect goes along with decreased reconstruction quality and intrinsically prevents the models from recovering all true generating factors and as such renders these cases uninteresting.

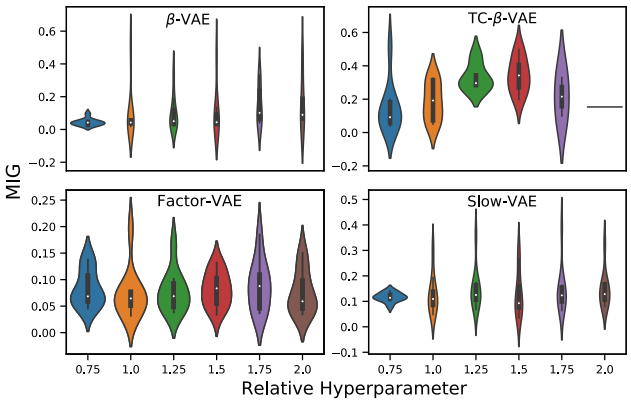

Figure 5: MIG scores for scaled literature hyperparameters over 10 restarts for Shapes3D. Overpruning runs with fewer active units than generating factors were discarded

6. Conclusion

We have shown that the success of β-VAE based architectures is mostly based on the structured nature of the datasets they are being evaluated on. Small perturbations of the dataset can alleviate this structure and decrease the bias that such architectures exploit. Interestingly, even architectures that are proven to be identifiable, like the Slow-VAE, still owe their success to the same bias. PCL and the weakly supervised GAN, however, as non-variational methods, were unaffected by the small perturbation.

It remains an open question whether the same local structure can reliably be found in real world data on which such archi-

tectures could be deployed. If so, fostering the sensitivity of future architectures towards the natural alignment of data could result in a transparent advance of unsupervised representation learning. It would be interesting to investigate and compare the different nonlinear embeddings VAE based architectures find. There are hints of clearly distinct local minima of the optimization problem; their suitability for downstream applications remains unexplored.

Acknowledgements

We thank Maximilian Seitzer, Stefan Bauer and Lukas Schott for the fruitful and invaluable discussions. Georg Martius is a member of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. We acknowledge the support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039B). The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Dominik Zietlow.

References

Adel, T., Ghahramani, Z., and Weller, A. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning*, pp. 50–59, 2018.

Baldi, P. and Hornik, K. Neural networks and principal component analysis: Learning from examples without local minima. *Neural networks*, 2(1):53–58, 1989.

Bell, A. J. and Sejnowski, T. J. An information-maximization approach to blind separation and blind deconvolution. *Neural computation*, 7(6):1129–1159, 1995.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Besserve, M., Shajarisales, N., Schölkopf, B., and Janzing, D. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pp. 557–565. PMLR, 2018.

Besserve, M., Sun, R., Janzing, D., and Schölkopf, B. A theory of independent mechanisms for extrapolation in generative models. *arXiv preprint arXiv:2004.00184*, 2020.

Bourlard, H. and Kamp, Y. Auto-association by multilayer perceptrons and singular value decomposition. *Biological cybernetics*, 59(4):291–294, 1988.

Burgess, C. and Kim, H. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *ArXiv e-prints*, abs/1804.03599, 2018. URL http://arxiv.org/abs/1804.03599.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Comon, P. Independent component analysis, a new concept? *Signal Processing*, 36(3):287 – 314, 1994. ISSN 0165-1684. doi: 10.1016/0165-1684(94)90029-9. Higher Order Statistics.

Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C., Lerchner, A., and Higgins, I. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SyxL2TNtvr.

Eastwood, C. and Williams, C. K. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

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 *International Conference on Learning Representations*, 2017.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

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

Hyvarinen, A. and Morioka, H. Nonlinear ica of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pp. 460–469. PMLR, 2017.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational autoencoders and nonlinear ica: A unifying framework. In Chiappa, S. and Calandra, R. (eds.), *Proceedings of the Twenty Third International Conference on*

Artificial Intelligence and Statistics, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 26–28 Aug 2020.

Kim, H. *Interpretable Models in Probabilistic Machine Learning*. PhD thesis, University of Oxford, 2019.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proc. ICML*, volume 80, pp. 2649–2658. PMLR, 2018a. URL http://proceedings.mlr.press/v80/kim18b.html.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2649–2658. PMLR, 10–15 Jul 2018b. URL http://proceedings.mlr.press/v80/kim18b.html.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Klindt, D. A., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., and Paiton, D. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=EbIDjBynYJ8.

Kumar, A. and Poole, B. On implicit regularization in β-vaes. In *International Conference on Machine Learning*, pp. 5480–5490. PMLR, 2020.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017.

Liao, Y., Schwarz, K., Mescheder, L., and Geiger, A. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5871–5880, 2020.

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 *international conference on machine learning*, pp. 4114–4124, 2019.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21(209):1–62, 2020. URL http://jmlr.org/papers/v21/19-976.html.

Lucas, J., Tucker, G., Grosse, R. B., and Norouzi, M. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, pp. 9408–9418, 2019.

Mathieu, E., Rainforth, T., Siddharth, N., and Whye Teh, Y. Disentangling disentanglement in variational autoencoders. *ArXiv e-prints*, abs/1812.02833, 2018.

Matthey, L., Higgins, I., Hassabis, D., and Lerchner, A. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Pearson, K. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

Rolinek, M., Zietlow, D., and Martius, G. Variational autoencoders pursue PCA directions (by accident). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2019.

Schölkopf, B. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., and Samaras, D. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5541–5550, 2017.

Stuehmer, J., Turner, R., and Nowozin, S. Independent subspace analysis for unsupervised learning of disentangled representations. In Chiappa, S. and Calandra, R. (eds.), *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1200–1210. PMLR, 2020. URL http://proceedings.mlr.press/v108/stuehmer20a.html.

Suter, R., Miladinovic, D., Schölkopf, B., and Bauer, S. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065, 2019.

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.