Disentanglement and Local Directions of Variance

Alexander Rakowski[1][✉] and Christoph Lippert[1,2]

[1] Hasso Plattner Institute for Digital Engineering, University of Potsdam, Germany
[2] Hasso Plattner Institute for Digital Health at Mount Sinai, New York, United States
{alexander.rakowski,christoph.lippert}@hpi.de

Abstract. Previous line of research on learning disentangled representations in an unsupervised setting focused on enforcing an uncorrelated posterior. These approaches have been shown both empirically and theoretically to be insufficient for guaranteeing disentangled representations. Recent works postulate that an implicit PCA-like behavior might explain why these models still tend to disentangle, exploiting the structure of variance in the datasets. Here we aim to further verify those hypotheses by conducting multiple analyses on existing benchmark datasets and models, focusing on the relation between the structure of variance induced by the ground-truth factors and properties of the learned representations. We quantify the effects of global and local directions of variance in the data on disentanglement performance using proposed measures and seem to find empirical evidence of a negative effect of local variance directions on disentanglement. We also invalidate the robustness of models with a global ordering of latent dimensions against the local vs. global discrepancies in the data.

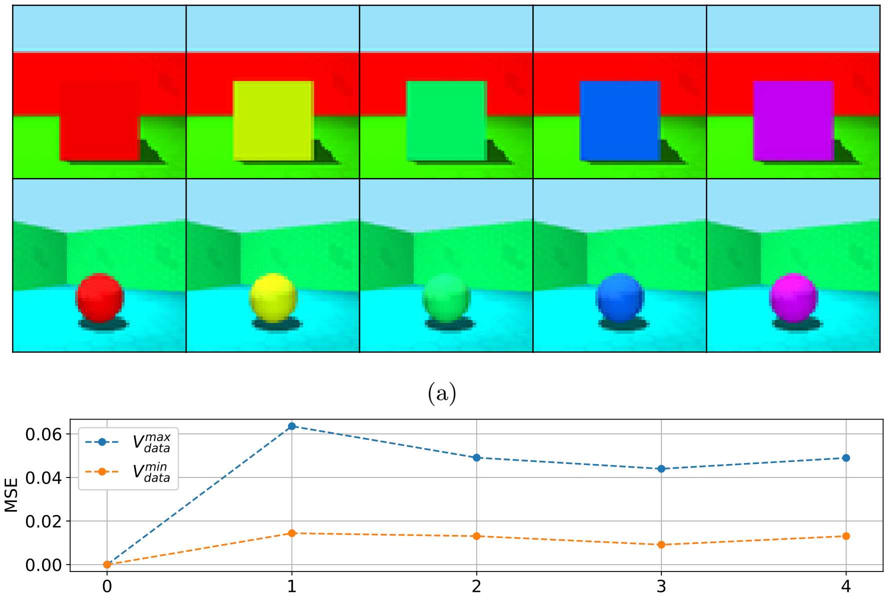
Keywords: Disentanglement · Variational Autoencoders · PCA

1 Introduction

Given the growing sizes of modern datasets and the costs of manual labeling, it is desirable to provide techniques for learning useful low-dimensional representations of data in an unsupervised setting. Disentanglement has been postulated as an important property of the learned representations [2, 28, 12]. Current state-of-the-art approaches utilize Variational Autoencoders (VAEs) [18], a powerful framework both for variational inference and generative modeling.

The primary line of work introduced models that were claimed to disentangle either by controlling the bottleneck capacity [13, 5] or explicitly imposing a factorising prior [19, 17, 6]. However, the validity of these claims has been challenged with the *Impossibility Result* [22, 23], showing the unidentifiability of disentangled representations in a purely unsupervised setting. An exhaustive empirical study from the same work further revealed that these models indeed fail to provide a consistent performance regardless of any tested hyperparameter setting. Instead, the scores exhibit a relatively high variance - disentanglement seems to happen "at random" - or at least cannot be attributed to the proposed techniques.

Meanwhile another line of research emerged, pointing to similarities between VAEs and Principal Component Analysis (PCA) [27] as a possible explanation to why disentanglement can happen in VAE-learned representations [33]. It was further postulated that the **global** structure of variance in the data is an inductive bias that is being exploited, while local discrepancies in the directions of variance in the data should have an adverse effect [41]. An example of different amounts of variance caused by the same underlying factor is shown in Figure 1.



(a)

(b)

Fig. 1: Example of different local impacts of a ground-truth factor in the Shapes3D dataset. **a)**: Top and bottom rows show images generated by interpolating over a single underlying factor - the color of the foreground object - for points belonging to hyperplanes in the latent space where this factor induces the most and the least variance respectively. **b)**: Mean Squared Error between consecutive pairs of these images - blue values correspond to the region with the most induced variance, the orange ones to the one with least variance.

In this work we aim to gain more insight and possibly further validate these claims by performing empirical analyses of properties of both the datasets and trained models. Our contributions are as follows: a) We formulate questions regarding the relations between the structure of variance in data, learned encodings and disentanglement. b) We define measures to quantify these properties and use them to perform statistical analyses to answer these questions c) We design synthetic datasets with hand-controlled specific structures of variance and employ models with more explicit connections to PCA. d) We identify strong connections between the defined measures and model properties and performance - in particular, we seem to find evidence for the negative effect of local variance directions on disentanglement. e) Contrary to the hypothesis from [41] we do not observe benefits of employing models with a global PCA-like behavior in the presence of such local discrepancies.

2 Related Work

Arguably one of the most important works on unsupervised disentanglement is the study of [22], which showed the limitations of current approaches, necessitating a shift towards either weak supervision [3, 36, 14, 24, 16] or investigating (potentially implicit) inductive biases [11, 31, 41].

Connections between autoencoders and PCA have been investigated, among others, in [1, 29, 33, 41] and between PCA modifications (e.g., pPCA, RPCA) in [7, 25, 20]. [37] note the bias of VAEs towards PCA solutions and illustrate it with examples of toy data where this bias is beneficial and ones where it is not. Furthermore, the framework of Deep Restricted Kernel Machines [38] provides a Kernel PCA [34] interpretation of autoencoders. Such models with imposed orthogonality constraints on the latent space encodings are shown to be a competitive alternative for VAEs in terms of disentanglement [40].

We build largely upon the findings of the recent work of [41], who postulate a global structure of variance in the datasets as an existing inductive bias. While providing an insightful perspective on the problem, they aim to validate this claim somewhat indirectly, generating artificial data from trained neural networks. The use of such complex non-linear models can obfuscate and change the actual problem being analyzed. Instead, we work directly on the original data, analyzing the local and global variance structures with quantities computed either analytically or estimated using Monte Carlo procedures. [8] is another work connected to ours, as they compute similarities between learned encodings of multiple trained models. These quantities are then used to arguably successfully select in an unsupervised manner models that disentangle, based on the assumption that "disentangled representations are all alike". We are not aware, to the best of our knowledge, of other works that conduct an empirical analysis of properties of both data and models, in the context of disentanglement.

3 Disentanglement, PCA and VAEs

3.1 Preliminaries

Variational Autoencoders Introduced in [18], VAEs are a framework for performing variational inference using latent variable models. The objective is to find parameters $\boldsymbol{\theta}$ that maximize the Evidence Lower Bound (ELBO), which we will denote as $\mathcal{L}(\boldsymbol{\theta}, \mathbf{x})$, which lower-bounds the otherwise intractable marginal log-likelihood of the N observed data points $\mathbf{x}^{(i)}$:

$$\sum_{i}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \sum_{i}^{N} \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})] - D_{KL}(q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})||p(\mathbf{z})) \tag{1}$$

In practice the prior distribution over the latent variables $p(\mathbf{z})$ is taken to be the PDF of a standard Normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, while $q_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}}$ are modeled with a neural network model each (encoder and decoder), optimized with a gradient ascent algorithm of choice using the objective function:

$$-\mathcal{L}(\boldsymbol{\theta}, \mathbf{x}) = \mathcal{L}_{Rec}(\mathbf{x}', \mathbf{x}) + \beta \mathcal{L}_{KLD}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) \tag{2}$$

where $\boldsymbol{\mu}, \boldsymbol{\sigma} = Enc_{\theta}(\mathbf{x})$ is the local parameterization of $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$, $\mathbf{x}' = Dec_{\theta}(\mathbf{z})$ is the reconstructed input and $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ is its sampled latent representation. $\mathcal{L}_{Rec}$ is usually based on Gaussian or Bernoulli priors, yielding the mean squared error (MSE) or binary cross-entropy (BCE) losses. The second term, $\mathcal{L}_{KLD}(\cdot, \cdot)$, is the Kullback-Leibler divergence between the (approximate) posterior and prior distributions. For the standard Normal prior it can be computed in closed form:

$$\mathcal{L}_{KLD}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \sum_{j=1}^{d} [\boldsymbol{\mu}_j^2 + \boldsymbol{\sigma}_j^2 - \log \boldsymbol{\sigma}_j^2 - 1] \tag{3}$$

This can also be understood as a regularization term, forcing the distribution $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ of latent codes $\mathbf{z}$ around each data point $\mathbf{x}$ to match the (uninformative) prior. The β term in Eq. 2 acts as a hyperparameter controlling the tradeoff between reconstruction quality and regularization, introduced with the β-VAE model [13]. It is worth highlighting, that even though both Enc_{θ} and Dec_{θ} are deterministic functions, $\mathcal{L}_{Rec}(\mathbf{x}', \mathbf{x})$ contains hidden stochasticity induced by sampling $\mathbf{z} \sim Enc_{\theta}(\mathbf{x})$ which in turn generates $\mathbf{x}' = Dec_{\theta}(\mathbf{z})$.

Disentanglement This property is usually defined with respect to the *ground-truth factors of variation* - a set of d random variables $\mathbf{G} = (G_1, \ldots, G_d)$ which generate the (usually higher-dimensional) data $\mathbf{X}$ via an unknown function $f : \mathbb{R}^d \mapsto \mathbb{R}^k$. It is commonly assumed that the factors are mutually independent, i.e., coming from a factorized distribution, e.g., $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The task is then to learn a representation $r : \mathbb{R}^k \mapsto \mathbb{R}^l$, where each dimension is dependent on at most one factor G_i. There exist several metrics for measuring disentanglement of representations [13, 17, 6, 32, 9, 19], however the majority of them was found to be strongly correlated [22].

Thus for brevity we report only the Mutual Information Gap (MIG) [6] throughout this work, as it has the advantage of being deterministic and having relatively few hyperparameters. Informally, it first computes the mutual information between each factor and latent dimension of the representation, and stores these values in a $d \times l$ matrix. For each G_i, its corresponding MIG is then taken to be the difference (gap) between the two highest normalized entries of the i-th row. Intuitively, if a factor is encoded with a single latent, this value will be close to 1, and close to 0 otherwise (for an entangled representation).

Impossibility Result In a fully unsupervised setting we never observe the true values of $\mathbf{g}$ and only have access to the corresponding points in the data space $\mathbf{x} = f(\mathbf{g})$ generated from them. Variational autoencoders are arguably the state-of-the-art method in this setting, where the representation r is defined by the encoder network, with its corresponding posterior distribution $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$. However, the main difficulty lies assessing whether a representation is disentangled, without having access to $\mathbf{g}$. There exists a potentially infinite number of solutions yielding the same marginal $p(\mathbf{x})$ and $p(\mathbf{z})$, which VAEs are optimized to match, but different conditional $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ and $q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ [22]. This is due to the rotational invariance of the prior distribution - applying a rotation to $\mathbf{z}$ in the encoder does not change $p(\mathbf{z})$. By undoing the rotation in the decoder, the marginal $p(\mathbf{x})$ remains unchanged, keeping the value of the ELBO objective (Eq. 1) intact.

3.2 Disentanglement in a PCA Setting

Unsupervised disentanglement is thus fundamentally impossible in a general setting. It is believed however, that certain inductive biases might help to overcome this difficulty. In particular, a PCA-like behavior of VAEs has been postulated as one [33, 41]. Recall that the PCA objective can be formulated as minimizing the squared reconstruction error between the original and reconstructed data:

$$\min_{\mathbf{W}} \|\mathbf{X} - \mathbf{W}\mathbf{W}^\top \mathbf{X}\|_2^2, \quad s.t.\ \forall i : \|\mathbf{W}_i\| = 1 \tag{4}$$

where $\mathbf{W}$ is usually taken to be of a smaller rank than $\mathbf{X}$. We can draw the analogy between disentanglement and PCA by treating $\mathbf{W}^\top$ and $\mathbf{W}$ as (linear) encoder and decoder networks respectively, and $\mathbf{z} = \mathbf{W}^\top \mathbf{x}$ as the low-dimensional representations of the data. A PCA model could thus be able to disentangle, given that the underlying factors G that generated the data coincide with the principal components [12]. An important benefit of this approach is the potential to escape the problem of non-identifiability of the solution highlighted in Section 3.1, due to the imposed ordering of principal components. Adopting this view allows us to reason about the conditions regarding the data for identifiability of a potentially disentangled representation. For the solution of PCA to be unique, the data covariance matrix $\mathbf{X}^\top\mathbf{X}$ must have non-degenerate eigenvalues. This translates to the assumption, that each G_i should have a different impact on the data. While it is unlikely that these values will be exactly the same in real-life scenarios, one can still recover differing solutions due to noise in the data or stochasticity of optimization algorithms. If the alignment with PCA directions were in fact an inductive bias beneficial for disentangling, we could ask the following:

Question 1 *Do models disentangle better on datasets where the ground-truth factors of variation contribute to the observed data with different magnitudes?*

3.3 PCA Behavior in Variational Autoencoders

The relation between autoencoders and PCA has been studied from several different angles, under the assumption of linear networks. Just the L_2 reconstruction objective is already enough for a linear autoencoder to learn the subspace spanned by the PCA solution [4, 1], while the actual principal components can be identified by performing Singular Value Decomposition of the model's weights [29]. The variational mechanisms of VAEs provide further connections to PCA. Specifically, the variance components $\boldsymbol{\sigma}^2$ promote local orthogonality of the decoder [33] and even alignment with local the principal components [41] around a point $\mathbf{x}$. While insightful, these findings hold under several simplifying assumptions, such as linearity of the decoder and varying $\boldsymbol{\sigma}_i^2$, which must not necessarily be true in practice. The optimal solution also depends on the value β, via its effect on $\mathcal{L}_{KLD}$. However indirect, these analogies to PCA are argued to be crucial for disentanglement [33, 41, 40, 26], leading us to ask:

Question 2 *Given the assumption from Question 1, will models with a more explicit PCA behavior achieve better disentanglement scores (e.g., MIG)?*

Another caveat of the above is that it is is a **local** (wrt. the data space) effect, as in the standard VAE framework values of $(\boldsymbol{\sigma}^{(i)})^2$ depend on $\mathbf{x}^{(i)}$. In case of non-linearly generated data these locally uncovered directions must not match the global ones. It is postulated in [41] that a consistent structure of variance in the dataset is an inductive bias allowing VAEs to disentangle.

Question 3 *Are local vs. global discrepancies present in the benchmark datasets? Do they have an influence on disentanglement scores obtained by the models?*

If the assumption about the negative effect of non-global PCA behavior in VAEs were indeed true, a natural solution would be to employ models which order the latent dimensions globally [35, 25].

Question 4 *Are models explicitly imposing a global ordering of latents robust to the inconsistency of variance in the data?*

We describe two such models employed in our study in Section 5.2.

4 Measuring Induced Variance and Consistency

4.1 Ground-truth Factor Induced Variance

Since the impact of the j-th ground-truth factor on the generated data cannot be computed analytically in the benchmark datasets, we define it instead as the average per-pixel variance induced by interventions on that factor:

$$V^i_{data}(\mathbf{g}) = \frac{1}{k}\sum_{j}^{k} \operatorname*{Var}_{g_i \sim p(g_i)}(p(\mathbf{x}_j|\mathbf{g}_{(\mathbf{g}_i=g_i)})), \tag{5}$$

which is computed locally on realizations of $\mathbf{g}$. By marginalizing over $\mathbf{z}$ we can obtain a global measure of influence: $\overline{V}^i_{data} = \int V^i_{data}(\mathbf{g})p(\mathbf{g})d\mathbf{g}$. We can estimate the values of $\overline{V}^i_{data}$ via a Monte Carlo procedure: we sample N points in the ground-truth space, and for each of them generate additional points by iterating over all possible values of $\mathbf{g}_i$ while keeping all the other factors $\mathbf{g}_{j\neq i}$ fixed. Using these values we define the "spread" of per-factor contributions as:

$$S_{data} = \underset{i\in[d]}{\mathrm{Var}}(\overline{V}^i_{data}) \tag{6}$$

over the (finite) set of all per-factor values of $\overline{V}^i_{data}$ for a given dataset.

4.2 Local Directions of Variance

Building up on the definition of the per-factor induced variance we can define the consistency of a factor's impact on the data:

$$C^i_{data} = \underset{\mathbf{g}\sim p(\mathbf{g})}{\mathrm{Var}}(V^i_{data}(\mathbf{g})) \tag{7}$$

In contrast to $\overline{V}_{data}$, which measures the average impact over the data, C_{data} quantifies the variability of it. For example, data generated by a (noiseless) linear transformation would be considered perfectly consistent ($\forall i : C^i_{data} = 0$). This must not be true however for more complex data - changes in color of a particular object will induce a smaller total change in regions of the data space where the said object has a smaller size or is occluded - see Figure 1. To estimate these values empirically we use an analogous procedure as in Section 4.1.

4.3 Consistency of Encodings

To quantify which latent dimensions of a model encode a certain ground-truth factor i we measure the amount of variance induced in each dimension j by changes in that factor around a $\mathbf{g}$:

$$V^{i,j}_{enc}(\mathbf{g}) = \underset{g_i\sim p(g_i)}{\mathrm{Var}}(q(\mathbf{z}_j|\mathbf{x})p(\mathbf{x}|\mathbf{g}_{(\mathbf{g}_i=g_i)})) \tag{8}$$

We estimate these values empirically in a similar manner as V_{data} and C_{data} (Sections 4.1, 4.2), with an additional step of passing the obtained points in the data (image) space through a trained encoder network. Due to the non-linear nature of the networks modeling $q(\mathbf{z}|\mathbf{x})$ there is no guarantee that a given factor will be mapped to the same latent dimensions **globally** over the dataset. Variance of V_{enc} over the dataset can then be used as a proxy measure of the consistency of encodings:

$$C^i_{enc} = \frac{1}{l}\sum_{j=1}^{l} C^{i,j}_{enc} = \frac{1}{l}\sum_{j=1}^{l} \underset{\mathbf{g}\sim p(\mathbf{g})}{\mathrm{Var}}(V^{i,j}_{enc}(\mathbf{g})) \tag{9}$$

High values of C^i_{enc} might indicate that the i-th factor is encoded differently over different parts of the data space - perhaps due to differing local vs. global structures of variance. Note that this is irrespective of whether the representation is disentangled or not - instead it can be thought of as consistency in preserving the same rotation of the latent space wrt. the ground-truth factors.

5 Experimental Setup

5.1 Datasets

We constructed two families of synthetic datasets to investigate the effects of the spread of per-factor induced variances S_{data} and consistency of per-factor induced variances C_{data} on disentanglement and properties of VAE models.

Synthetic Data with Varying $\overline{V}_{data}$ In the former case we generate higher-dimensional data $\mathbf{x} \in \mathbb{R}^k$ from low-dimensional ground-truth factors $\mathbf{z} \in \mathbb{R}^d$ via random linear mappings $\mathbf{W} \in \mathbb{R}^{d \times k}$ with unit-norm columns:

$$\mathbf{x} = \mathbf{z}\mathbf{W} = \sum_i \mathbf{z}_i \mathbf{W}_i; \;\; \|\mathbf{W}_i\| = 1; \;\; \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\sigma}^2 \mathbf{I}) \tag{10}$$

This is a non-trivial task in terms of disentangling, since (linear) ICA results require at most one factor to be Gaussian for identifiability [15, 10]. Values of $\overline{V}^i_{data}$ can be computed in closed form, since the data points are defined as linear transformations of Gaussian random variables. We create several different datasets by varying the diversity of entries of $\boldsymbol{\sigma}^2$, ranging from all factors having the same variance ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_d^2$) to all of them being different ($\sigma_1^2 < \sigma_2^2 < \ldots < \sigma_d^2$). Note that for this dataset we have $\forall i : C^i_{data} = 0$ due to the linearity of the data-generating process.

Synthetic Data with Non-global Variance Structure The second variant is constructed to have differing local principal directions. We define three ground-truth factors: two zero-centered Normal random variables z_1, z_2 with $\sigma_1^2 < \sigma_2^2$, and a third variable $z_3 \sim \mathcal{U}\{1, k\}$. We then construct the data vectors $\mathbf{x} \in \mathbb{R}^k$ the following way:

$$\mathbf{x}_i = \begin{cases} z_1, & \text{if } i \leq z_3 \\ z_2, & \text{otherwise} \end{cases} \tag{11}$$

Thus z_3 defines the number of entries in $\mathbf{x}$ equal to z_1, while the rest is filled with z_2. Because of that it also changes the local principal components of $\mathbf{x}$ (and their corresponding eigenvalues). There are regions of high z_3 where it is more profitable for the encoder to transmit z_1 with more precision than z_2 - even though the latter induces more variance in $\mathbf{x}$ globally ($\overline{V}^2_{data} > \overline{V}^1_{data}$). Thus, contrary to the previous dataset, we have $C^1_{data}, C^2_{data} > 0$.

Benchmark Data We also used commonly established "benchmark" datasets, first employed together in the study of [22]: *dSprites*, *noisy-*, *color-* and *scream-dSprites*, *NORB*, *Cars3D* and Shapes3D [13, 22, 21, 30, 17]. They all consist of 64×64 sized images, created in most cases artificially, with known corresponding ground-truth factors used to generate each image.

5.2 Models

Baseline As a baseline for the experiments we took the β-VAE model. For the benchmark datasets we used pretrained models from the study of [22], which are publicly available for download[3] (apart from models trained on the Shapes3D dataset).

Global Variance VAE Introduced in [25], this model uses a global variance vector i.e., $\boldsymbol{\sigma}^2$ is not a function of $\mathbf{x}$. More formally, the posterior $q_{\boldsymbol{\theta}}$ is defined as:

$$q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I}) = \mathcal{N}(Enc_\theta(\mathbf{x}), \boldsymbol{\sigma}^2\mathbf{I}) \tag{12}$$

where $\boldsymbol{\sigma}^2 \perp\!\!\!\perp \mathbf{X}$. Note that the variance components are still learned from the data, just kept as constant parameters after training. For linear networks, the global solution of this model coincides with the solution of pPCA [39].

Hierarchical Non-Linear PCA (h-NLPCA) Introduced in [35], this model imposes an explicit ordering of the latent dimensions in terms of their contribution to reconstruction. This is done by using a modified reconstruction loss:

$$\mathcal{L}_{Rec\_H}(\theta, \mathbf{x}) = \frac{1}{l}\sum_{i}^{l} \mathcal{L}^{i}_{Rec\_H}(\theta, \mathbf{x}) = \frac{1}{l}\sum_{i}^{l} \mathcal{L}_{Rec}(Dec_\theta(Enc_\theta(\mathbf{x}) \odot \boldsymbol{\delta}_i)), \tag{13}$$

where $\boldsymbol{\delta}_i$ denotes a "masking" vector with $l - i$ leading 1's and i trailing 0's. $\mathcal{L}^{i}_{H\_Rec}$ thus measures the reconstruction loss when allowing information to be encoded only in the first i latent dimensions, forcing each $\mathbf{z}_i$ to be more beneficial for reconstruction than $\mathbf{z}_{i+1}$.

6 Results

6.1 The Effect of Different Per-factor Contributions

Synthetic Data We created several variants of the dataset defined in Section 5.1 with increasing values of S_{data}. There is barely any change in in disentanglement scores wrt. S_{data} for β-VAE and Global Variance models. On the other hand, the h-NLPCA models exhibit a much stronger relation, outperforming the baseline the more the higher the dataset's S_{data} (see Figure 2). This might indicate that the explicit ordering of dimensions in h-NLPCA is beneficial for settings with high S_{data}. Even more notable, however, is the fact that even these models underperform against a simple PCA baseline.

[3] `github.com/google-research/disentanglement_lib`

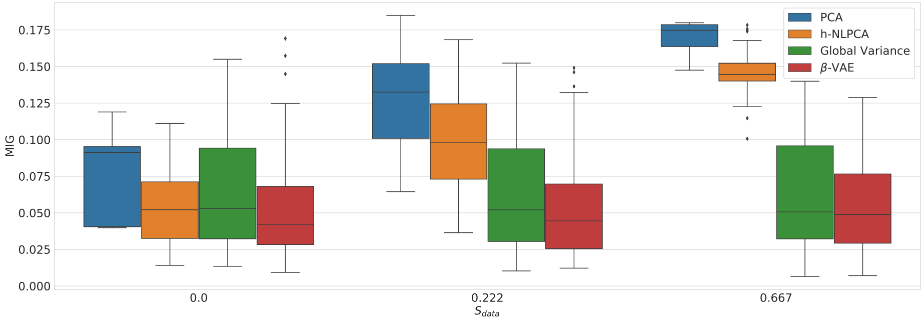


Fig. 2: MIG scores obtained on synthetic data with controlled S_{data}. Models imposing an explicit ordering of latent dimensions wrt. impact on reconstruction (PCA and h-NLPCA) perform best, especially when $S_{data} > 0$.

Benchmark Data This relation is also visible on the benchmark datasets. Figure 3 shows the Pearson correlation of estimated values of S_{data} with obtained MIG scores. Interestingly, while β-VAE models exhibit an arguably strong correlation (> 0.70) for smaller values of β, increasing the regularization strength seems to consistently weaken this relation. This seems contradictory with the regularization loss term being postulated as the mechanism inducing PCA-like behavior in VAEs. The other models exhibit a somewhat reverse relation: while still present, the correlations are weaker, but they grow with increased β-s.

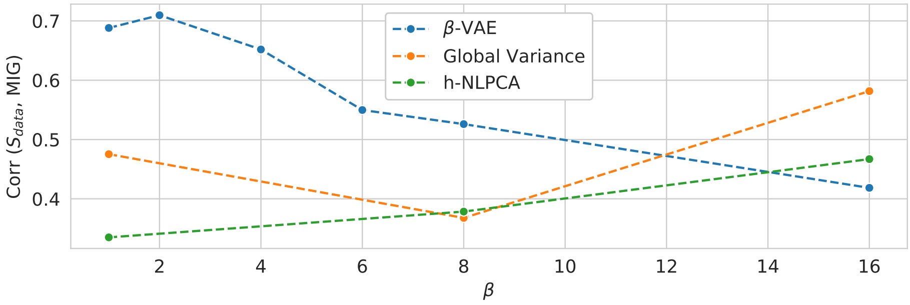


Fig. 3: Pearson correlation coefficient between S_{data} for each benchmark dataset with MIG scores obtained on it (y-axis), for different values of β (x-axis). There is an arguably strong correlation, which decreases with higher β-s for β-VAE while increasing for the other models.

6.2 The Effect of Non-global Variance Structure in the Data

Does Inconsistency of Variance Structure in Data Correlate with Inconsistency of Encodings? We observed prevailing correlations between the consistency of per-factor induced variance in the dataset C^i_{data} and the corresponding consistency of learned encodings of that factor $C^{i,\cdot}_{enc}$, with the strongest values obtained by the Global Variance models. Table 1 shows values of the per-dataset Pearson correlation coefficient averaged over all β values and random seeds. It seems that factors whose V_{data} vary across the data tend to also be encoded with different latent dimensions, depending on the location in the data space. Figure 4 shows the mean correlation over all datasets wrt. values of β. While almost constant for Global Variance models, these correlations steadily decrease with higher β-s. One can see that for smaller values (≤ 8) this effect is much stronger - up to a coefficient of over 0.7. This could mean that stronger regularization could be beneficial for alleviating non-global mappings of factor-latent being caused by non-global variance structure in the data.

Table 1: Per-dataset correlation between consistency of per-factor induced variance in data C^i_{data} and variance in dimensions of learned representations $C^{i,\cdot}_{enc}$ for models trained on the benchmark datasets. While differing in strength, it is clearly present in almost all cases, the only exception being Cars3D for the h-NLPCA models.

Model	dSprites	Color-dS.	Noisy-dS.	Scream-dS.	Norb	Cars3D	Shapes3D	Mean
β-VAE	0.302	0.598	0.432	0.501	0.447	0.462	0.314	0.437
Global Var.	0.710	0.862	0.409	0.647	0.287	0.501	0.379	0.542
h-NLPCA	0.307	0.558	0.247	0.407	0.339	-0.124	0.635	0.338

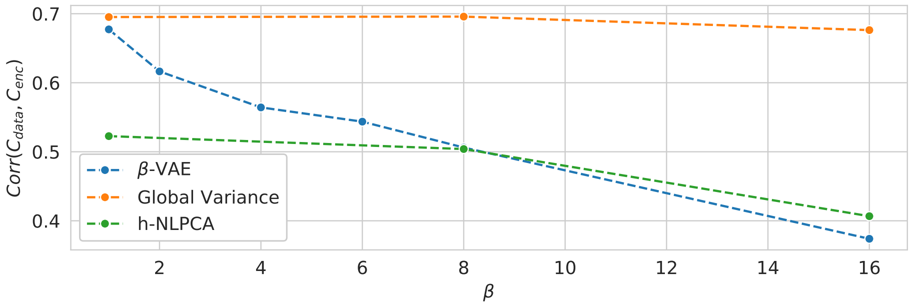


Fig. 4: Correlations between C^i_{data} and $C^{i,\cdot}_{enc}$ (y-axis) averaged over all datasets, for different values of β (x-axis). For models with a non-global variance structure a clear negative relation is visible.

Do These Inconsistencies Correlate with Disentanglement Scores? A perhaps more interesting question is whether the above effects translate to disentanglement of learned representations. In Section 6.1 we analyzed the effects of global, per-dataset, variance structures. Here we are interested in a more fine-grained, per-factor effect caused by the discrepancies between local and global directions of variance.

First we investigate whether disentanglement is correlated with inconsistency of encodings. This could reveal an indirect effect of C_{data} on disentanglement, propagated through its correlation with C_{enc} seen in the previous section. Table 2 shows per-dataset correlations between a model's C_{enc} and its obtained MIG score - we observe negative correlations in most settings, albeit weaker than these from the previous section. While present for all model families, they are strongest for β-VAE.

We also look for a direct relation with C_{data}. To account for the correlation we observed in Section 6.1 (see Figure 3), we first fit an ordinary least squares model on each dataset's values of S_{data} and MIG scores across different β values, and compute the correlations between C_{data} on the residuals. While weaker than with S_{data}, there are still negative correlations present - models tend to perform worse on datasets with higher average C_{data} (see Figure 5). Interestingly, increasing β seems to amplify this effect.

Table 2: Correlation between the MIG scores and C_{enc} for models trained on the benchmark datasets. Negative correlations seem to be most prevalent for the β-VAE models.

Model	dSprites	Color-dS.	Noisy-dS.	Scream-dS.	Norb	Cars3D	Shapes3D	Mean
β-VAE	-0.612	-0.439	-0.320	-0.754	0.058	-0.054	-0.657	-0.397
Global Var.	-0.110	-0.251	-0.194	-0.532	0.069	-0.033	-0.207	-0.180
h-NLPCA	-0.103	-0.068	-0.081	-0.035	-0.105	-0.410	-0.337	-0.163

6.3 The Effect of Non-global Variance Structure in the Models

Synthetic data In Figure 6 we compare performance of models on the synthetic data with non-global structures of variance (see Section 5.1). In this case, as opposed to the results from Section 6.1, the PCA baseline does not outperform the VAE-based models. For lower β values ($\{10^{-1}, 10^{0}\}$) the Global Variance models perform best. However, for higher β-s there is a drastic decrease in performance, where their MIG scores fall below even those of PCA. On the other hand, the β-VAE and h-NLPCA models do not exhibit this sudden drop in performance. This might suggest that when the directions of variance change over the dataset, the ability to adapt the variance structure locally improves robustness against tighter bottlenecks.

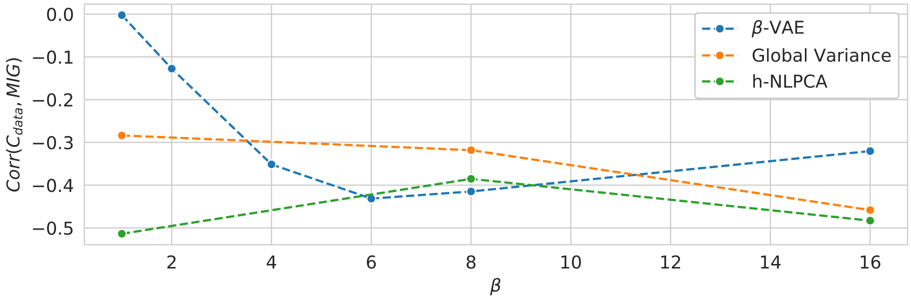


Fig. 5: Correlation between C_{data} of benchmark datasets and obtained MIG scores (y-axis) for different values of β (x-axis). At least a weak negative correlation is present for all tested models, often growing with higher β-s.

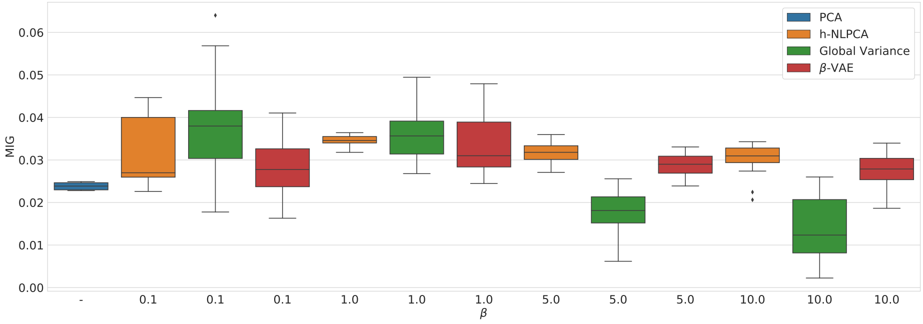


Fig. 6: MIG scores (y-axis) obtained on the dataset with non-global variance structure for different values of β (x-axis). Notable is the sudden drop in performance for the Global Variance VAE for higher β-s.

Benchmark Data We also perform a similar analysis on the benchmark data - the obtained MIG scores for each dataset and β value are reported in Table 3. There seems to be no clear advantage of employing the proposed models over β-VAE. Especially the Global Variance VAE seems to drastically underperform, regardless of the hyperparameter setting. h-NLPCA models achieve results more comparable to the baseline. They seem to have an advantage with weak regularization - perhaps this is due to the PCA-like behavior being induced not only by the KL divergence, but also by the modified reconstruction loss.

7 Conclusions

In this work we approach an existing hypothesis of an inductive bias for disentanglement, from the perspective of analyzing properties of datasets and models directly. Reviewing the existing connections between VAEs and PCA we stated

Table 3: Mean MIG scores obtained on the benchmark datasets for the baseline and proposed models, for different β values (first 9 rows) and averaged over all of them (last 3 rows).

β	Model	Cars3D	Color-dS.	dSprites	Noisy-dS.	Scream-dS.	Shapes3D	Norb
1	β-VAE	0.046	0.067	0.075	0.019	**0.040**	0.076	**0.250**
	Global Var.	0.030	0.027	0.025	0.015	0.031	0.053	0.057
	h-NLPCA	**0.068**	**0.080**	**0.087**	**0.108**	0.022	**0.432**	0.154
8	β-VAE	**0.099**	0.127	**0.124**	0.068	**0.165**	0.291	**0.208**
	Global Var.	0.031	0.048	0.029	0.029	0.049	0.091	0.066
	h-NLPCA	0.092	**0.148**	0.103	**0.087**	0.018	**0.492**	0.130
16	β-VAE	**0.121**	**0.148**	**0.243**	0.080	**0.105**	**0.465**	**0.188**
	Global Var.	0.049	0.030	0.027	0.034	0.042	0.112	0.158
	h-NLPCA	0.101	0.106	0.094	**0.087**	0.001	0.397	0.187
-	β-VAE	**0.089**	**0.114**	**0.147**	0.056	**0.103**	0.277	**0.215**
	Global Var.	0.037	0.035	0.027	0.026	0.041	0.085	0.094
	h-NLPCA	0.087	0.111	0.094	**0.094**	0.014	**0.440**	0.157

questions regarding expected performance of the models wrt. structure of variance in the datasets, and defined quantifiable measures used to answer them.

Models seem to disentangle better on datasets where the per ground-truth factor induced variances vary stronger (Question 1). However there doesn't seem to be a clear benefit of exploiting this relation with models with stronger connections to PCA (Question 2). We also find that local vs. global discrepancies of variance structure are indeed present in the datasets, and are negatively correlated with both consistency of encodings and disentanglement (Question 3). Surprisingly, contrary to the assumption from [41], models with a global ordering of latents seem to be less robust against these discrepancies (Question 4).

We note that albeit seemingly strong in some cases, these correlations should not be taken as exhaustive or causal explanations of the mechanisms governing variational autoencoders and disentanglement. Instead, they are meant to empirically (in-)validate previously stated assumptions and point to new intuitions. Specifically, we see potential in further analyzing the connection to PCA and the role of local directions of variance.

References

1. Baldi, P., Hornik, K.: Neural networks and principal component analysis: Learning from examples without local minima. Neural networks **2**(1), 53–58 (1989)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
3. Bouchacourt, D., Tomioka, R., Nowozin, S.: Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
4. Bourlard, H., Kamp, Y.: Auto-association by multilayer perceptrons and singular value decomposition. Biological cybernetics **59**(4), 291–294 (1988)

5. Burgess, C.P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., Lerchner, A.: Understanding disentangling in beta-vae. arXiv preprint arXiv:1804.03599 (2018)
6. Chen, T.Q., Li, X., Grosse, R.B., Duvenaud, D.K.: Isolating sources of disentanglement in variational autoencoders. In: Advances in Neural Information Processing Systems. pp. 2610–2620 (2018)
7. Dai, B., Wang, Y., Aston, J., Hua, G., Wipf, D.: Connections with robust pca and the role of emergent sparsity in variational autoencoder models. The Journal of Machine Learning Research **19**(1), 1573–1614 (2018)
8. Duan, S., Matthey, L., Saraiva, A., Watters, N., Burgess, C., Lerchner, A., Higgins, I.: Unsupervised model selection for variational disentangled representation learning. In: International Conference on Learning Representations (2019)
9. Eastwood, C., Williams, C.K.: A framework for the quantitative evaluation of disentangled representations (2018)
10. Eriksson, J., Koivunen, V.: Identifiability and separability of linear ica models revisited. In: Proc. of ICA. vol. 2003, pp. 23–27 (2003)
11. Gondal, M.W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., Bauer, S.: On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. Advances in Neural Information Processing Systems **32**, 15740–15751 (2019)
12. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016)
13. Higgins, I., Matthey, L., Glorot, X., Pal, A., Uria, B., Blundell, C., Mohamed, S., Lerchner, A.: Early visual concept learning with unsupervised deep learning. arXiv preprint arXiv:1606.05579 (2016)
14. Hosoya, H.: Group-based learning of disentangled representations with generalizability for novel contents. In: IJCAI. pp. 2506–2513 (2019)
15. Hyvärinen, A.: Survey on independent component analysis (1999)
16. Khemakhem, I., Kingma, D., Monti, R., Hyvarinen, A.: Variational autoencoders and nonlinear ica: A unifying framework. In: International Conference on Artificial Intelligence and Statistics. pp. 2207–2217. PMLR (2020)
17. Kim, H., Mnih, A.: Disentangling by factorising. In: International Conference on Machine Learning. pp. 2649–2658. PMLR (2018)
18. Kingma, D.P., Welling, M.: Auto-encoding bariational bayes. arXiv preprint arXiv:1312.6114 (2013)
19. Kumar, A., Sattigeri, P., Balakrishnan, A.: Variational inference of disentangled latent concepts from unlabeled observations. In: International Conference on Learning Representations (2018)
20. Kunin, D., Bloom, J., Goeva, A., Seed, C.: Loss landscapes of regularized linear autoencoders. In: Chaudhuri, K., Salakhutdinov, R. (eds.) Proceedings of the 36th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 97, pp. 3560–3569. PMLR (09–15 Jun 2019)
21. LeCun, Y., Huang, F.J., Bottou, L.: Learning methods for generic object recognition with invariance to pose and lighting. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. vol. 2, pp. II–104. IEEE (2004)
22. Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O.: Challenging common assumptions in the unsupervised learning of disentangled representations. In: international conference on machine learning. pp. 4114–4124. PMLR (2019)
23. Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., Bachem, O.: A commentary on the unsupervised learning of disentangled representations.

In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 13681–13684 (2020)

24. Locatello, F., Poole, B., Rätsch, G., Schölkopf, B., Bachem, O., Tschannen, M.: Weakly-supervised disentanglement without compromises. In: International Conference on Machine Learning. pp. 6348–6359. PMLR (2020)
25. Lucas, J., Tucker, G., Grosse, R.B., Norouzi, M.: Don't blame the elbo! a linear vae perspective on posterior collapse. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
26. Pandey, A., Fanuel, M., Schreurs, J., Suykens, J.A.: Disentangled representation learning and generation with manifold optimization. arXiv preprint arXiv:2006.07046 (2020)
27. 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)
28. Peters, J., Janzing, D., Schölkopf, B.: Elements of causal inference: foundations and learning algorithms. The MIT Press (2017)
29. Plaut, E.: From principal subspaces to principal components with linear autoencoders. arXiv preprint arXiv:1804.10253 (2018)
30. Reed, S.E., Zhang, Y., Zhang, Y., Lee, H.: Deep visual analogy-making. In: Advances in neural information processing systems. pp. 1252–1260 (2015)
31. Ren, X., Yang, T., Wang, Y., Zeng, W.: Rethinking content and style: Exploring bias for unsupervised disentanglement. arXiv preprint arXiv:2102.10544 (2021)
32. Ridgeway, K., Mozer, M.C.: Learning deep disentangled embeddings with the f-statistic loss. In: Advances in Neural Information Processing Systems. pp. 185–194 (2018)
33. Rolinek, M., Zietlow, D., Martius, G.: Variational autoencoders pursue pca directions (by accident). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12406–12415 (2019)
34. Schölkopf, B., Smola, A., Müller, K.R.: Kernel principal component analysis. In: International conference on artificial neural networks. pp. 583–588. Springer (1997)
35. Scholz, M., Vigário, R.: Nonlinear pca: a new hierarchical approach. In: Esann. pp. 439–444 (2002)
36. Shu, R., Chen, Y., Kumar, A., Ermon, S., Poole, B.: Weakly supervised disentanglement with guarantees. In: International Conference on Learning Representations (2019)
37. Stühmer, J., Turner, R., Nowozin, S.: Independent subspace analysis for unsupervised learning of disentangled representations. In: International Conference on Artificial Intelligence and Statistics. pp. 1200–1210. PMLR (2020)
38. Suykens, J.A.: Deep restricted kernel machines using conjugate feature duality. Neural computation **29**(8), 2123–2163 (2017)
39. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61**(3), 611–622 (1999)
40. Tonin, F., Patrinos, P., Suykens, J.A.: Unsupervised learning of disentangled representations in deep restricted kernel machines with orthogonality constraints. arXiv preprint arXiv:2011.12659 (2020)
41. Zietlow, D., Rolinek, M., Martius, G.: Demystifying inductive biases for *beta*-vae based architectures. arXiv preprint arXiv:2102.06822 (2021)