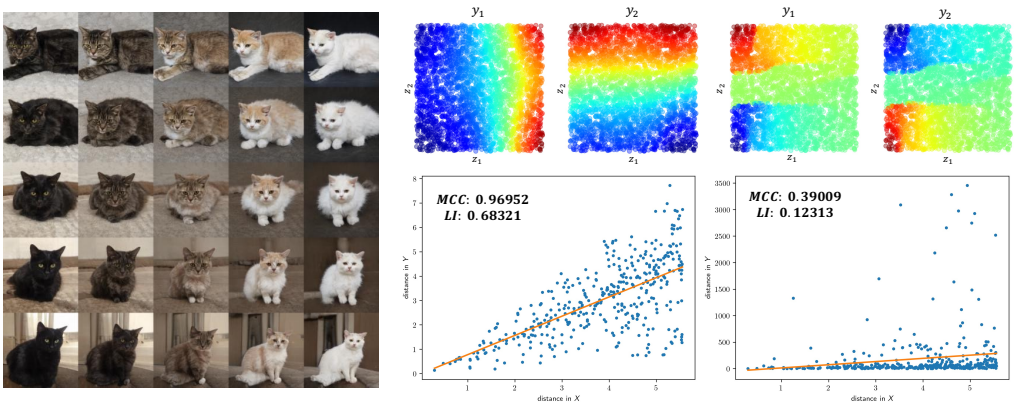
Is Disentanglement Possible?

Anonymous Author(s)
Affiliation
Address
email

Abstract

A common assumption in many domains is that high dimensional data are a smooth nonlinear function of a small number of independent factors. Is it possible to recover the factors from data? In the context of deep models this problem is called "disentanglement" and was recently shown to be impossible without additional strong assumptions [17, 19]. In this paper, we show that the assumption of local isometry together with non-Gaussianity of the factors, is sufficient to provably recover disentangled representations from data. We give examples of cases for which these conditions provably hold, and leverage recent advances in deep generative models to construct manifolds of highly realistic images for which the ground truth latent representation is known, on which we test whether modern and classical methods succeed in recovering the latent factors. For many different manifolds, we find that a spectral method that explicitly optimizes local isometry and non-Gaussianity, consistently finds the correct latent factors, while baseline deep autoencoders do not. We show how to encourage deep autoencoders to find encodings that satisfy local isometry and show that this helps them discover disentangled representations. Overall, our results suggest that in some realistic settings, disentanglement is provably possible, without any domain-specific assumptions.



(a) cat1 (b) HLLE+ICA (c) IRMAE

Figure 1: (a) manifold of highly realistic images with two independent latent factors . Given unlabled images, is it possible to recover the true factors? HLLE+ICA (b) succeeds while a baseline autoencoder (c) learns a highly entangled embedding. In this paper we show that the assumptions of local isometry and non-Gaussianity are sufficient to provably recover disentangled representations.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

1 Introduction

With the recent success of deep learning approaches, there has been a resurgence of interest in the problem of recovering the independent factors that generate high dimensional data. As a motivating example, consider the `cat1` manifold shown in figure 1a. Each image is of size 256×256, but is generated as a smooth, nonlinear function of just two latent independent factors (the fur color and the pose of the cat). Is it possible to recover these latent factors, given only samples of images?

This problem is closely related, yet different from the problem of learning a deep generative model from data. Given a training set $\{x_i\}$, deep generative models find parameters θ such that $x_i = g_\theta(z_i)$ where each latent variable z_i is sampled from a base distribution, and the different components of z are independent. Tremendous progress in deep generative models has been achieved by methods including Generative Adversarial Networks (GAN) [10, 1, 16], Variational Auto-Encoders (VAE) and their variants [18] and optimization-based approaches [5].

While deep generative models have achieved remarkable success in generating realistic images given a random input, the latent variables that they find are often very different from the desired "independent factors" [19]. Returning to our motivating example in figure 1a, a deep generative model may generate perfectly realistic images given random latent vectors z, where some of the coordinates of z have only a weak relationship to the true latent factors, and no single coordinate encodes a latent factor perfectly, as desired.

Figure 1 illustrates this issue. In the top of column (c) we show the representation learned by a particular autoencoder [15] for the data on the left (see figure 5 for additional baseline autoencoders). The embedding is shown as a function of the ground truth factors (z_1: fur color, z_2: cat pose). The autoencoder learned two latent dimensions, y_1 and y_2, and we plot $y_1(z_1, z_2)$, $y_2(z_1, z_2)$ using the "jet" colormap that goes from blue to red. If each learned dimension encodes only one factor, then these plots should have contours that are parallel to the axes. For this particular autoencoder, however, this is not the case. Even though it gives very low reconstruction loss (see supplementary), it can be seen that neither fur color nor pose are directly encoded by either of the learned factors. Such a representation is called "entangled" [4], and there has been much recent work attempting to quantify the extent to which a learned representation is "disentangled" [9, 12, 6].

An intriguing result from the recent literature on disentanglement is that the problem is impossible without additional assumptions, even with infinite data [19, 17]. For example, [17] argues that "models with any form of unconditional prior $p(z)$ are unidentifiable", and advocates using conditional priors (where z is conditioned on a third variable such as time). This non-identifiability is easiest to see in the case of a base distribution that is standard normal: we can always rotate z by an arbitrary rotation matrix Q to obtain another random variable $\tilde{z} = Qz$, that is also standard normal. If $g(z) = x$ then we can also define $\tilde{g} = gQ^T$ and clearly $\tilde{g}(\tilde{z}) = x$. This issue is well-understood in the context of linear functions of latent factors, i.e. Independent Component Analysis [13]. In the linear ICA setting, it is known that while the problem is non-identifiable when the latent factors are Gaussian, it can be solved if the latent factors have a sufficiently non-Gaussian distribution. The recent work of [19, 17], points out that in the nonlinear case, the identifiability problem is much more severe: even if the independent factors are non-Gaussian, there can be many non-linear transformations $\tilde{z} = q(z)$ that preserve the distribution of z. Again, by defining $\tilde{f}(\tilde{z}) = f(q^{-1}(\tilde{z}))$, for $f(z) = x$, we can perfectly generate the data from latent variables that have the desired distribution yet are highly entangled. This suggests that distenanglement is impossible without additional strong assumptions on the architecture and the data even with infinite data, and indeed Locatello et al. demonstrated empirically that many deep autoencoder methods fail to learn a disentangled representation.

In this paper we show the power of one very general assumption that can enable the learning of disentangled representations. The assumption of *local isometry*, i.e. that an ϵ change in the latent representation corresponds to an ϵ change in the observed variables. In the bottom of figure 1b,c we show a scatter plot of the distances $d(x_i, x_j)$ and $d(y_i, y_j)$ for randomly chosen pairs of nearby points. If the embedding satisfies local isometry, these points should lie on a straight line. As can be seen, the highly entangled autoencoder embedding does not satisfy local isometry: there is almost no correlation between the magnitude of change in the latent space Y and the magnitude of change in the original input space X.

Specifically, we prove that as the number of datapoints goes to infinity, a combination of two classical methods: Hessian Eigenmaps [8] and fastICA [13], is guaranteed to find a disentangled

representation when the manifold parameterization satisfies local isometry and the base distribution is sufficiently non-Gaussian. Figure 1b shows the result of our proposed method on the same data that the autoencoder was trained on. The contours in the plots are now axis-aligned, indicating that a perfectly disentangled representation was learned. As shown in the bottom plot, there is a strong correlation between the pairwise distances of points in the input space X and their corresponding representations in the latent space Y.

In the remaining sections, we first derive analytical conditions on manifold parameterizations that guarantee that disentanglement will be possible and give examples of realistic cases for which these conditions hold. We leverage recent advances in deep generative models to construct manifolds of highly realistic images for which the ground truth latent representation is known, and test whether different methods succeed in recovering the latent factors (up to permutation). For many different manifolds, we find that methods that explicitly optimize local isometry and non-Gaussianity consistently find the correct latent factors, while baseline deep autoencoders do not. Finally, we discuss ways to encourage deep autoencoders to find encodings that satisfy local isometry, and show that this helps them discover disentangled representations. Overall, our results suggest that in some realistic settings, disentanglement is possible without any domain-specific assumptions.

2 Local Isometry + non-Gaussianity make disentanglement possible

Our main theoretical result shows that two classical algorithms are guaranteed to find disentangled representations under certain assumptions. We first define the assumptions and the algorithms and then state the result.

A mapping g from z to x satisfies *local isometry* if for all z, the Jacobian $J(z) = \frac{\partial g}{\partial z}$ satisfies:

$$J(z)^T J(z) = I$$

Note that local isometry is not only a property of the manifold, but also a property of the latent representation. Thus a semi-circle with the parameterization $(x, \sqrt{1-x^2})$ does not satisfy local isometry but the same semi-circle with the parameterization $(\cos\theta, \sin\theta)$ does satisfy the constraint. Indeed, assuming that the dimension of z is much smaller than the dimension of x, then any manifold can be approximated by an isometric mapping. [2].

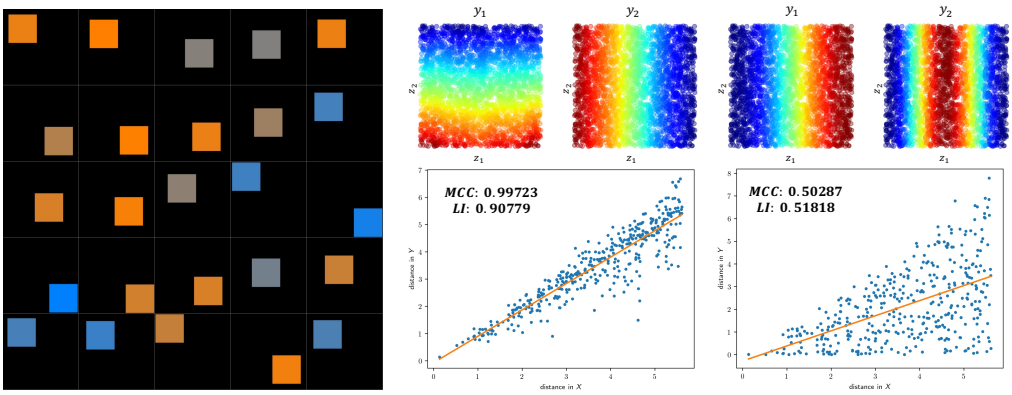
Given a distribution over $p(z)$ we say that z is sufficiently non-Gaussian if linear ICA is guaranteed to recover z (up to permutation) from a linear mixing Az, given infinite training examples. Specifically we use the fastICA algorithm [13] which searches for projections of the data which maximize the deviation between the expectation of a nonlinear function of the projection, $G(u)$, and the same expectation with a Gaussian distribution. It is guaranteed to be consistent "for most reasonable choices of G, and distributions of the s_i. In particular, if $G(u) = u^4$, the condition is fulfilled for any distribution of non-zero kurtosis. In that case, it can also be proven that there are no spurious optima"[14].

The two manifold learning algorithms we discuss in the theorem are Hessian Eigenmaps (HLLE) [8] and Laplacian Eigenmaps (LEM) [3]. At a high level, both algorithms receive as input a set of points $\{x_i\}_{i=1}^M$, a scalar that defines a local neighborhood (typically k, the number of nearest neighbors to include in the neighborhood), and d, the dimension of the manifold. They proceed by constructing an $M \times M$ matrix (L, the Laplacian matrix for LEM, and H, the Hessian matrix, for HLLE) and choosing the $d+1$ eigenvectors of the matrix with smallest eigenvalues. The output of the algorithms is y_i a vector of length d that gives the embedding of x_i, and is computed from the ith component of the d eigenvectors after discarding the trivial constant eigenvector. A full description of both algorithms is given in the supplementary material.

Finally, we refer to HLLE+ICA as an algorithm that first runs HLLE and then runs ICA on the embedding found by HLLE. Similarly LEM+ICA. We are now ready to state our two main results.

Theorem 1: Let $\{x_i\}$ be samples from a manifold $x = g(z)$. If g satisfies local isometry, z is i.i.d. and each coordinate is uniform, then as the number of samples goes to infinity, HLLE+ICA and LEM+ICA will provably find a disentangled representation of the true latent factors z.

Proof: Donoho and Grimes [8] showed that as the number of examples goes to infinity, the H matrix converges to the (continuous) Hessian operator on the manifold whose $d+1$ eigenfunctions with zero



(a) `shapes-xc` (b) HLLE+ICA (c) LEM+ICA

Figure 2: An illustration of the difference between Theorem 1 and Theorem 2. (a) Samples from the `shapes-xc` manifold generated by two independent factors (diagonal position, color). The mapping from these factors to the images does not satisfy local isometry, causing LEM+ICA to fail (c), but a mapping from a monotonic transformation of the two factors to the observations does satisfy local isometry, allowing HLLE+ICA to succeed (b).

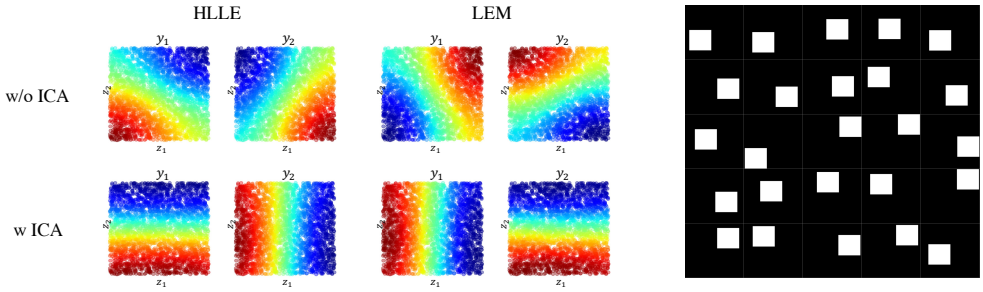
eigenvalue are spanned by the latent factors. They also showed that if the support of $p(z)$ is connected, then HLLE will recover the isometric embedding of the manifold up to a linear transformation. The assumption of non-Gaussianity means that running fastICA on the output of HLLE will provably recover the original independent factors. Similarly, Belkin and Niyogi [3] have shown that when the number of examples goes to infinity, the L matrix converges to the (continuous) manifold Laplacian, and Singer and Coifman [23] have shown that under local isometry, the $d+1$ eigenfunctions with smallest eigenvalues are monotonic functions of the true latent functions. Since the eigenvalues are equal, LEM may recover a linear combination of the true factors and again the assumption of non-Gaussianity means that running fastICA on the output of LEM will provably recover the original independent factors.

The problem with theorem 1 is that it requires the latent factors to satisfy local isometry and the definition of local isometry is very sensitive: even when the two latent factors are pose and color, there are many different ways to parameterize each factor (e.g. pose can be measured in degrees or radians, color can be measured in RGB or YUV, etc.), and different parameterizations will in general not preserve local isometry. The following theorem is much more general.

Theorem 2: Let $\{x_i\}$ be samples from a manifold $x = g(z)$. If there exists an alternative representation of the manifold $x = g_2(s)$ with $s = Af(z)$, for any matrix A and pointwise nonlinearity f, such that g_2 satisfies local isometry and $f(z)$ is sufficiently non-Gaussian, then as the number of samples goes to infinity, HLLE+ICA will provably find a disentangled representation of the true latent factors z.

Proof: As in the proof of theorem 1, the results of Donoho and Grimes guarantee that HLLE will find $Bs = BAf(z)$ for some matrix B. Again, the assumption of non-Gaussianity guarantees that running ICA on the output of HLLE will recover $f(z)$.

Figure 2 illustrates the difference between the two theorems. Images of squares are generated by two latent factors: color and location, which are sampled uniformly and independently. Changing the color of the square by ϵ has a much greater effect on the image compared to changing the location by ϵ, so $f(z)$ does not satisfy local isometry and thus theorem 1 does not hold, causing LEM+ICA to fail. However, since there exists a rescaling of the latent variables that satisfies local isometry, theorem 2 does hold and HLLE+ICA succeeds.



(a) embeddings (b) `shapes-xy`

Figure 3: Results of running HLLE+ICA and LEM+ICA on 3000 samples from a 2D manifold created by translating a white square on a black background. The results of [7] guarantee that this manifold satisfies local isometry and our theorem thus guarantees that spectral methods followed by ICA will recover a perfectly disentangled representation. As can be seen, the algorithms work perfectly even with finite data.

3 Manifolds for which local isometry provably holds

The previous section shows that when the ground truth factors are related to an isometric encoding of the manifold, then disentanglement is possible. But this will not be the case if the "ground truth" factors are a highly entangled function of an isometric encoding. Thus referring again to figure 1, if the "ground truth" factors were the two variables found by the autoencoder rather than "fur color" and "pose", then we cannot expect HLLE+ICA to recover them. This raises the question, for what kind of manifolds and representations will the conditions of theorems 1 and 2 hold? We now give examples of seven cases.

The first five examples were given in [7] ”(1) Translation of simple black objects on a white background; (2) Pivoting certain simple black objects on a white background around a fixed point; (3) Morphing of boundaries of black objects on a white background; (4) Articulation of ‘fingers’ of a digital ‘hand’; (5) Articulation of a cartoon face by arranging its eyebrows, eyelids, and lips.”[7].

Two more examples are (6) color images of aligned faces in which the latent variables control the color of the hair, eyes and mouth; and (7) gene expression data in which the latent variables activate disjoint subsets of the genes. The last two examples can be shown to satisfy the condition that $J^T J$ is diagonal (but not necessarily equal to the identity). However, we can independently reparameterize each latent factor using a monotonic transformation that makes $J^T J = I$. Thus, they satisfy the condition of theorem 2 and not theorem 1.

Figure 3 shows the results of HLLE and LEM on `shapes-xy`, a 2D manifold of squares at different positions. This is equivalent to the first case discussed in [7], and therefore the mapping from latent factors to images will provably satisfy local isometry. Together with theorem 1, disentanglement is guaranteed to be possible for this manifold. Indeed as shown in figure 3, both LEM and HLLE find the correct factors up to a linear transformation, and a perfectly disentangled representation with the addition of ICA.

Note that this list is by no means exhaustive, but is rather meant to illustrate the generality of the local isometry assumption.

4 Experiments

We conducted experiments on two simple synthetic manifolds and 8 realistic synthetic manifolds. The two simple manifolds were constructed by squares on a black background, where we varied the 2D position in the first manifold (`shapes-xy`), and the diagonal position and the color in the second (`shapes-xc`). Similar manifolds have been used in many recent papers on disentanglement (e.g. [12, 15]), and we ensured that the latent representation has a non-Gaussian distribution by sampling the latent variables from independent uniform distributions.

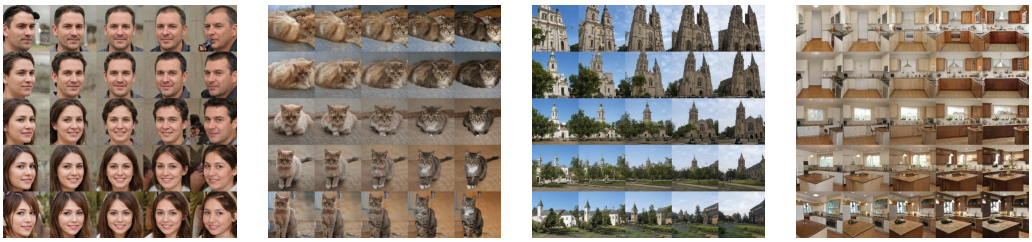


Figure 4: Realistic 2D image manifolds generated with StyleGAN and GANSpace. Each dataset consists of 3000 samples. The rows correspond to the z_1 direction and the columns to the z_2 direction.

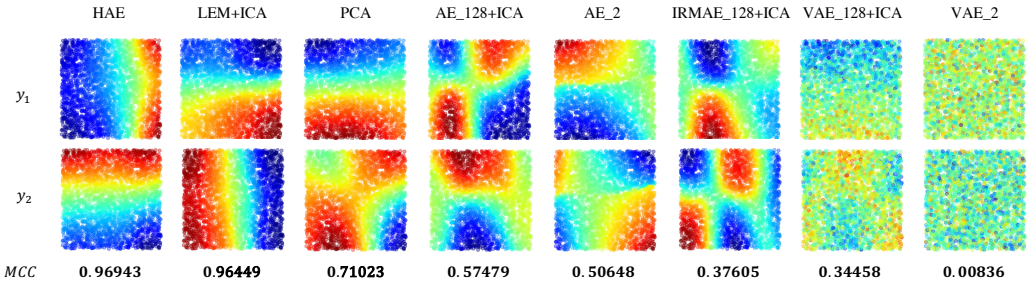


Figure 5: The embeddings found for the `cat1` manifold shown in figure 1 by various methods. None of the baseline methods manage to learn a disentangled representation.

In addition to these simple manifolds, we also leveraged recent advances in deep generative models to create realistic image manifolds where each image is a complex, nonlinear function of a small number of known latent variables.

Specifically we use GANSpace [11] to find interpretable semantic directions in the W space of StyleGAN2 [16], using PCA. To generate a d-dimensional manifold we select d such directions and obtain their corresponding PCA components. We then uniformly sample the coefficients of these d components, while fixing the coefficients of the remaining components to a constant (0). For example, for the `ffhq` manifold, we use the first two PCA components identified by GANSpace, which correspond to head pose and gender, as visible in figure 4. For the `cat2` manifold, the directions we select correspond to pose and fur pattern. The following table lists the 8 realistic manifolds that we generated using GANspace and their independent factors.

	`ffhq`	`cat1`	`cat2`	`horse`	`car`	`kitchen`	`bedroom`	`church`
z_1	gender	pose	pose	pose	viewpoint	structure	viewpoint	zoom level
z_2	pose	color	pattern	color	color	cabinet	illumination	building material

For each dataset we ran both modern deep autoencoder methods as well as the two classical methods described previously: HLLE+ICA (our proposed method) and LEM+ICA. HLLE has a single free parameter k (the number of neighbors) and we find it by choosing the k for which the local isometry score is maximal. The local isometry score is defined as the correlation coefficient between distances in the input space $d(x_i, x_j)$ and distances in the embedding space $d(y_i, y_j)$, when we only consider pairs of *nearby* points x_i, x_j for which $d(x_i, x_j) < \epsilon$, and ϵ was set to include the closest 5% among the pairs of points. Since $d(y_i, y_j)$ is sensitive to a rescaling of the latent representations, we first compute a diagonal rescaling of the representations that minimizes the MSE between the rescaled $d(y_i, y_j)$ and $d(x_i, x_j)$. For some values of k we found that fastICA did not converge (less than 5% of the runs) and we simply ignored the results with that value.

We also measured the performance of baseline autoencoders on the same datasets. Since these methods have many free parameters, we do not claim that the results shown here are exhaustive. Our main focus is to show that the disentanglement tasks we use here are not trivial: simply using an autoencoder with a small bottleneck along with standard regularization methods is not sufficient to ensure disentanglement. The autoencoders we use were all taken from the code of a recent paper [15]

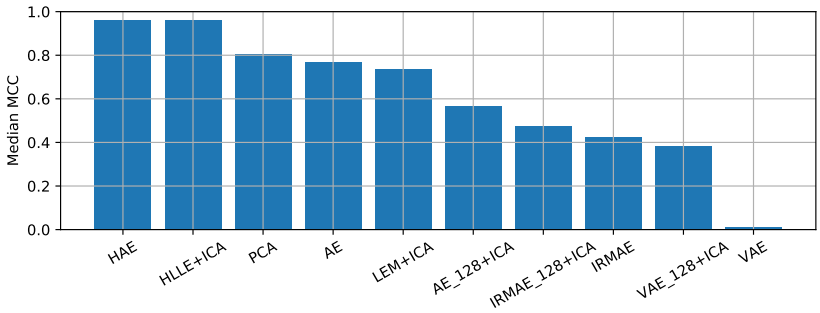


Figure 6: Quantitative results for all methods on the 8 styleGAN manifolds. Our proposed methods succeed in finding a disentangled representation for all 8 manifolds, while the baseline autoencoders are no better than PCA.

and includes an implementation of a "vanilla" autoencoder (AE) with 4 conv+relu layers with stride=2 and one fully-connected layer. The Implicit Rank-Minimizing Autoencoder (IRMAE) includes in addition a number of fully connected linear layers and this has been shown to implicitly regularize the solution towards a low rank embedding. We also tested a basic implementation of a β-VAE with the value of β that was found in [15] to provide the best performance. Since deep autoencoders sometimes fail to learn good representations with very narrow bottlenecks, we trained two versions of all autoencoders: one in which the bottleneck was of width 2 (the true dimensionality of the manifold), and another with a bottleneck of width 128. In the second case, we ran fastICA to find a two dimensional projection of the 128 dimensional latent vector that is as non-Gaussian as possible. We also provide the results of a PCA encoder as a baseline.

Qualitative results are shown in figures 5 and 1 for the `cat1` manifold. None of the baseline autoencoder methods managed to find a disentangled representation, while HLLE+ICA and LEM+ICA found a disentangled one. Evidently the VAE suffered from "posterior collapse", i.e. "when the variational distribution closely matches the uninformative prior" [20].

As a quantitative performance metric we used the Mean Correlation Coefficient (MCC) which is standard in the nonlinear ICA community [17]. To compute it, we first solve the problem of permuting the components of the embedding $y(x)$ so it best matches the latent factors $z(x)$, using the Hungarian algorithm. We then measure the Pearson correlation between the permuted embedding and the latent variables. The results are summarized in figure 6: *our proposed method consistently found a disentangled embedding, while baseline autoencoders did not*.

As expected, run-times of the spectral methods were much faster than the autoencoders. With $k = 30$ neighbors, running HLLE took about 6 seconds per manifold and ICA took less than a second, while training the autoencoder took over an hour.

4.1 Hessian Auto-Encoder (HAE)

In order to encourage autoencoders to learn representations that satisfy local isometry we use a "student-teacher" paradigm. We first run HLLE+ICA and then train the autoencoder to minimize the following loss:

$$L(\theta_1, \theta_2; x, y) = \sum_i \|E(x_i; \theta_1) - y_i\|^2 + \|D(y_i; \theta_2) - x_i\|^2 \tag{1}$$

where θ_1 are the parameters of the encoder $E(x)$, θ_2 are the parameters of the decoder $D(y)$ and y is the embedding calculated by HLLE+ICA. This essentially decouples the learning of the encoder and the decoder. We use exactly the same encoder and decoder architectures as the baseline methods. As can be seen in figure 7 (right), this method, which we call the Hessian Autoencoder (HAE), gives near perfect disentanglement on all manifolds, as opposed to the baseline autoencoders which often find highly entangled embeddings. This entanglement can also be seen in the interpolations. The baseline autoencoders often jump abruptly across the latent space, while HAE transitions smoothly along the manifold, gradually moving in both dimensions, as desired.

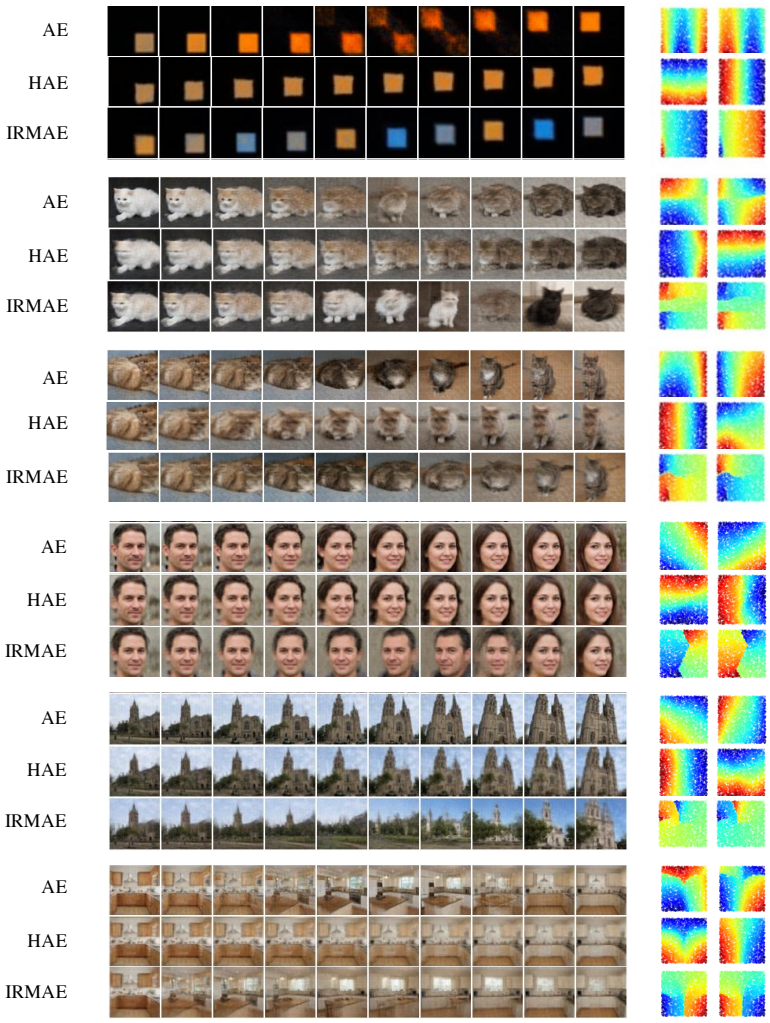


Figure 7: Interpolations in the embedding space of the different autoencoders generated by linearly interpolating between the encodings of the left and right image and then decoding. HAE transitions smoothly along the manifold between the two endpoints, while the baseline autoencoders often fail to do so. For each manifold and model, the corresponding embeddings are visualized to the right of the interpolation sequence. As can be seen, entangled embeddings result in discontinuous interpolations.

5 Related Work

Our work was inspired by the recent work of Locatello et al. [19] who point out that "unsupervised learning of disentangled representations is fundamentally impossible without inductive biases on both the models and the data." They conducted extensive empirical experiments on "more than 12000 models covering most prominent methods and evaluation metrics in a reproducible large-scale experimental study on seven different data sets" and observed "that while the different methods successfully enforce properties encouraged by the corresponding losses, well-disentangled models seemingly cannot be identified without supervision.".

We agree with the fundamental need for inductive biases but our work suggests that a relatively general inductive bias (local isometry) makes disentanglement possible without supervision.

Khemakhem et al. [17] extend the impossibility results and relate them to classical results on identifiability of nonlinear ICA. They state that "models with any form of unconditional prior $p(z)$ are unidentifiable" and argue for using conditional priors (where z is conditioned on a third variable such as time). Our work again acknowledges the non-identifiability issue without inductive bias, but

shows that a relatively general inductive bias allows us to estimate the correct latent factors even with an unconditional prior.

Tenenbaum et al. [24] pointed out the importance of global isometry in enabling manifold learning and Donoho and Grimes [8] pointed out that global isometry may be too strong an assumption and often local isometry is sufficient. Atzmon et al. [2] suggested using local isometry as a regularizer in deep autoencoders and compared this regularizer to other regularizers in the context of visualization of high dimensional data. Our focus, on the other hand, has been to show that local isometry together with non-Gaussianity is sufficient to guarantee disentanglement.

There has been much work on performance evaluation methods for disentanglement [12, 9]. Here we used the simple MCC performance measure, and verified using visual inspection that it captures the success or failure of disentanglement. A second major difference between our work and most existing papers on evaluating disentanglement is the use of a realistic, complex dataset that we define using StyleGAN. While we have also shown results on simple 2D shapes on constant background, we believe the StyleGAN datasets are much more realistic and capture a much wider range of latent factors that affect images.

We have focused in this paper mainly on classic spectral methods, but the development of spectral methods for manifold learning continues to be an active area of research. Most recently, the work of Pfau et al [22] presents a novel spectral method called GeoMancer and proves that it can correctly find disentangled representations when the true metric is known and the manifold has a nontrivial holonomy. They showed results on synthetic 3D objects on simple background and concluded that their method only works if the true distance (in latent space) was given to the algorithm ("when applied directly to the true latent state vectors, GeoMancer performs exceptionally well, even if the state is rotated randomly. When individual dimensions are multiplied by a random scale factor, the performance degrades, and if a random linear transformation is applied to the state, performance is no better than chance." Our work uses existing spectral methods and we have shown that even on highly realistic images, these methods can disentangle factors such as pose and illumination using only the Euclidean distance in pixel space.

The fact that LEM fails to recover the correct latent factors when local isometry does not hold was previously pointed out by Singer and Coifman [23] and by [21]. Both of them suggested methods for improving LEM: either by pre-computing a local covariance at each point and using it do define affinity or by post-processing the LEM output. On our datasets, HLLE consistently outperformed LEM but we believe improvements to LEM can be of great interest given our theoretical results.

6 Conclusions and Limitations

The main question addressed in our work is whether disentanglement is possible. Our theoretical and empirical results show that when we make two modest assumptions: (1) local isometry and (2) non-Gaussianity, the answer is "yes"; and a combination of two classical methods (HLLE+fastICA) can successfully learn disentangled representations in less than a minute. Our work does not contradict the recent impossiblity theorems on disentanglement which argue that the task is impossible without additional assumptions, but it shows that two very general assumptions suffice with no need for domain-specific assumptions. Moreover, in a set of experiments with highly complex realistic image manifolds, the assumptions appear to hold and disentanglement is possible with a simple, well-understood method.

While our work has shown that disentanglement is possible, we do not mean to suggest that the problem has been solved by running HLLE followed by fastICA. A major limitation of our theoretical result is that we focus on the infinite data limit and similarly the datasets we constructed correspond to a very dense sampling of the manifold (3000 images for a 2D manifold). When the sampling is sparser, it is known that spectral methods can be fooled by sampling noise. Furthermore, running spectral methods on large datasets (e.g. millions of images) is challenging and may require special-purpose approximations. We hope that our "possibility results" will encourage more research that will allow similarly successful disentanglement in a broader range of datasets.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[2] Matan Atzmon, Amos Gropp, and Yaron Lipman. Isometric autoencoders. *CoRR*, abs/2006.09289, 2020.

[3] Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.*, 15(6):1373–1396, 2003.

[4] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[5] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *CoRR*, abs/1707.05776, 2017.

[6] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[7] David Donoho and Carrie Grimes. Image manifolds which are isometric to euclidean space. *Journal of Mathematical Imaging and Vision*, page 5–24, 2005.

[8] David L. Donoho and Carrie Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. *Proceedings of the National Academy of Sciences*, 100(10):5591–5596, 2003.

[9] Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[11] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable GAN controls. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[12] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[13] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[14] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE TRANS. NEURAL NETW*, 10(3):626–634, 1999.

[15] Li Jing, Jure Zbontar, and Yann LeCun. Implicit rank-minimizing autoencoder. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[16] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020.

[17] Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In Silvia Chiappa and Roberto Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

[18] Diederik P. Kingma and Max Welling. An introduction to variational autoencoders. *Found. Trends Mach. Learn.*, 12(4):307–392, 2019.

[19] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *CoRR*, abs/1811.12359, 2018.

[20] James Lucas, George Tucker, Roger B. Grosse, and Mohammad Norouzi. Understanding posterior collapse in generative latent variable models. In *Deep Generative Models for Highly Structured Data, ICLR 2019 Workshop, New Orleans, Louisiana, United States, May 6, 2019*. OpenReview.net, 2019.

[21] David Pfau and Christopher P. Burgess. Minimally redundant laplacian eigenmaps. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Workshop Track Proceedings*. OpenReview.net, 2018.

[22] David Pfau, Irina Higgins, Aleksandar Botev, and Sébastien Racanière. Disentangling by subspace diffusion. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[23] Amit Singer and Ronald R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226 – 239, 2008.

[24] JB Tenenbaum, V De Silva, and JC Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, pages 2319–2323, 2000.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]