Inferring local structure from pairwise correlations

Mahajabin Rahman [*]
Department of Physics, Emory University, Atlanta, Georgia 30322, USA

Ilya Nemenman [†]
Department of Physics, Department of Biology, and Initiative in Theory and Modeling of Living Systems, Emory University, Atlanta, Georgia 30322, USA

(Received 8 May 2023; accepted 11 September 2023; published 29 September 2023)

To construct models of large, multivariate complex systems, such as those in biology, one needs to constrain which variables are allowed to interact. This can be viewed as detecting "local" structures among the variables. In the context of a simple toy model of two-dimensional natural and synthetic images, we show that pairwise correlations between the variables—even when severely undersampled—provide enough information to recover local relations, including the dimensionality of the data, and to reconstruct arrangement of pixels in fully scrambled images. This proves to be successful even though higher order interaction structures are present in our data. We build intuition behind the success, which we hope might contribute to modeling complex, multivariate systems and to explaining the success of modern attention-based machine learning approaches.

DOI: 10.1103/PhysRevE.108.034410

Introduction. The problem of data-driven inference of laws governing a complex system has become a staple of what theorists do [1]. This is particularly true in fields like biological physics, where high-throughput experiments generate large datasets, consisting of $T \gg 1$ samples of $N \gg 1$ measured variables, such as the activities of neurons or the presence or absence of mutations [2,3]. A probabilistic model $P(\mathbf{x})$ of such a system would include exponentially many coupling constants, which is unrealistic for experiments with $N \sim 10^2$–10^3 and $T \sim N$–$10N$ at best. In traditional physical systems, the combinatorial problem does not exist because variables can only interact over short distances, so the total number of interaction coefficients is $O(N)$. The interaction structure may also be sparse even for complex biological systems (e.g., a neuron may project into many, but not all other neurons). However, using sparsity is difficult until one knows which specific variables can interact [4–7]. Can we infer this effective local structure for complex systems from data alone? (Note that somewhat similar questions are also asked in the field of detecting community structure in complex networks [8–10].)

Physical locality and the constraints it imposes on the interactions are strongly dependent on the physical dimensionality of a problem. The effective dimensionality is usually either unknown or undefined for many complex biological systems. However, if any local, sparse interaction structure exists, it can be specified by a list of interaction partners (effective neighbors) of each variable under consideration, requiring only $O(N)$ numbers. For comparison, the covariance matrix between the measured variables has $O(N^2)$ independent numbers, though many may be noisy. Therefore, it is plausible that the locality can be determined from the pairwise covariance matrix alone, even for systems where interactions can be of higher order or long range (though presumably decaying with the effective distance between the interacting variables). This counting argument suggests that such locality reconstruction might be possible even when $T \sim N$, so that most empirical correlations are dominated by statistical noise [11]. Explicitly showing that this can be done in a specific problem is the goal of this article.

We show that the structure of the pairwise covariance matrix is sufficient to reconstruct effective local relations, even when interactions are manifestly dense and of higher order. Specifically, we analyze a dataset of black and white images, where the notion of locality is clear, the dimensionality of the data is known (images are planar), correlations are critical and hence strong, and higher order interactions among pixels abound [12]. We randomly permute the geometric location of the pixels and show that, based only on pairwise correlations within these randomized data, we can recover the dimensionality of the problem and restore the geometric arrangement of the pixels to a high accuracy, reconstructing the original images with relatively small computational costs. This result is encouraging for using pairwise correlations to determine local interaction neighborhoods in more complex datasets where the true geometric structure is unknown, and it opens doors to investigating if modern neural networks implicitly perform similar computations.

Problem setup. In order to investigate if locality manifests itself in correlations, we require large, uniform datasets, and hence a generative model of images with features and correlations of different scales. Additionally, we need the dependence of pixel-pixel correlation on pixel separation to be generally similar to that of natural images. To accomplish this, we generate $T = 40\,000$ synthetic images as our training set,

[*] mahajabin.rahman@emory.edu
[†] ilya.nemenman@emory.edu

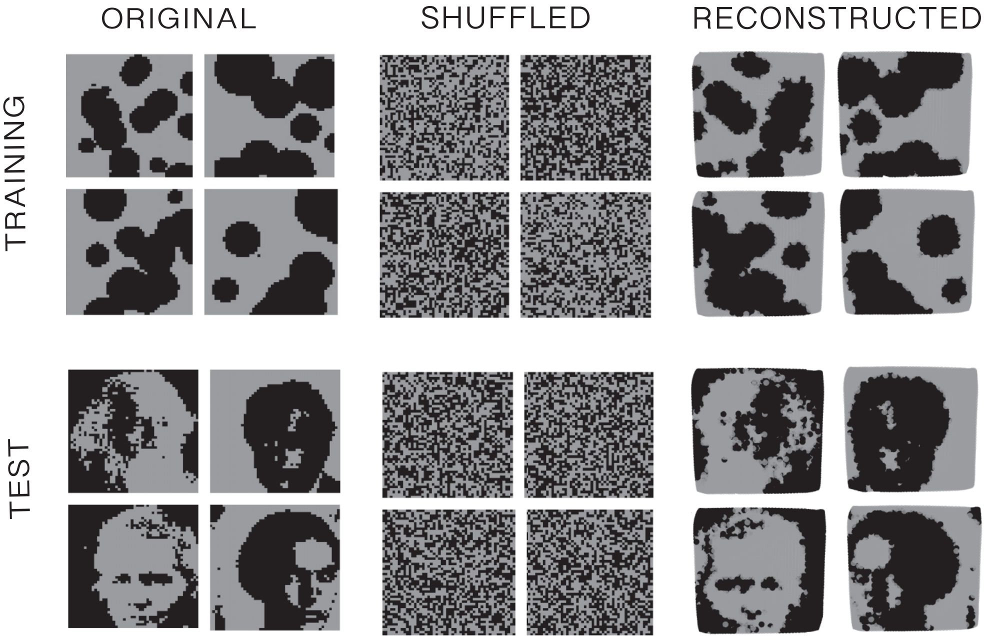

FIG. 1. Original, shuffled, and reconstructed images. The top half of the figure shows four examples from the *dead leaves* set of images (see text), which we use for training and detection of local pixel neighborhoods. The bottom half shows four examples of natural images, the test set, which we analyze based on local relations inferred for the training data. Pixels from original images (left) are shuffled according to one fixed random permutation (center). We then determine the optimal embedding of the shuffled pixels using t-SNE, such that pixels highly correlated over the training set are embedded closer to each other (see text). The shuffled images are then reconstructed (right) by mapping their pixels into planar coordinates according to this optimal embedding, and then performing a single global rotation and rescaling of all images to closer match the originals. Qualitatively (see text for quantitative metrics), training and test reconstructed images are very close to the original ones, indicating that correlations can be used for defining local neighborhoods in, at least, image data.

$I_i(\mathbf{x})$, $i = 1, \ldots, T$, $\mathbf{x} = (x, y)$, and $0 \leqslant (x, y) \leqslant 50$, using the dead leaves model [13–15]. This model simulates occlusion in natural images by piling opaque round objects onto an empty 50×50 canvas. We choose the maximum size of opaque circles, $r_{i,\max}$, at random from a uniform distribution between 1 and 19 pixels, and determine the number of circles of radius $r_{i,\max}$ required to make the image 50% opaque, denoted as n_i. We then uniformly sample the centers of the n_i circles and choose their radii uniformly between 0 and $r_{i,\max}$. The circles are placed on the canvas, and every pixel covered by at least one circle is marked as opaque, so that typically less than 50% of all pixels in each image end up being opaque, as shown in Fig. 1. To demonstrate the weak sensitivity of our conclusions to these specific parameter choices, we also assemble a second dataset consisting of 5000 50×50 pixel natural images of landscapes, faces, and animals, as shown in Fig. 1, which is used as our test data, on which the predictions are also applied.

We consider the correlation between pixels at positions $\mathbf{x}$ and $\mathbf{x}'$ in our dataset,

$$c_{\mathbf{x}\mathbf{x}'} = \frac{1}{T}\sum_{i=1}^{T} I_i(\mathbf{x})I_i(\mathbf{x}') - \bar{I}(\mathbf{x})\bar{I}(\mathbf{x}'), \tag{1}$$

$$\bar{I}(\mathbf{x}) = \frac{1}{T}\sum_{i=1}^{T} I_i(\mathbf{x}), \tag{2}$$

where T is the number of images in the dataset, and $\bar{I}(\mathbf{x})$ is the mean value of pixel intensity at position $\mathbf{x}$ over all images. We expect a strong dependence between the geometric distance of pixels $d^2 = (x - x')^2 + (y - y')^2$ and $c_{\mathbf{x}\mathbf{x}'}$, with pixels closer to each other being more likely to have the same color. We explore this dependence for both the synthetic and the natural data in Fig. 2. The general structure of $c_{\mathbf{x}\mathbf{x}'}(d)$ curves is similar for both datasets, with a rapid decay to zero indicating a statistical association between the strength of the correlation and the pixel-to-pixel distance. However, there is a distribution of correlations at the same distance d due to the large number of pixel pairs at each distance in the same image. The standard deviation of the correlation over pairs

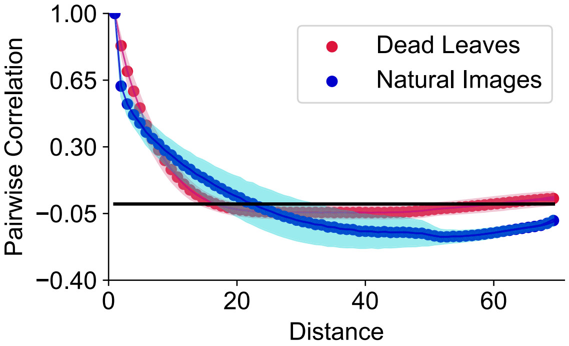

FIG. 2. Pixel-pixel pairwise correlation $c_{\mathbf{x}\mathbf{x}'}$ vs pixel-to-pixel distance d. The distances were grouped in 71 evenly spaced bins, since the range of distances in a 50×50 image are between 0 and 70.7. For each bin, we plot the correlation averaged over all pairs of pixels falling into the bin (dots) and the standard deviation (colored bands). Blue: natural images; red: dead leaves synthetic data.

of pixels the same distance apart (denoted as color bands in Fig. 2) measures the inhomogeneity of the images, and it is much larger for the natural data set.

The wide correlation bands in Fig. 2 indicate that the relationship between correlation and distance is not deterministic, even at very large data set sizes. This raises questions if it would be possible to use correlations to reconstruct the relative geometric position of pixels. To test this, we destroy the local structure by choosing a random permutation π, which reshuffles pixel positions, $\mathbf{x}^* = \pi(\mathbf{x})$. We plot original images $I(\mathbf{x})$ and their shuffled versions $I(\mathbf{x}^*)$ in Fig. 1.

Reconstructing local relations. In order to reconstruct the local relations among the shuffled pixels, we employed t-SNE, an algorithm designed to find an optimal embedding of the pixel coordinates in a metric space [16]. Other low-dimensional embedding methods could have been used [17–19], and we view our choice as a demonstration that the reconstruction of the local structure is possible with some methods, and not an endorsement of a specific method, t-SNE in this case.

t-SNE takes as inputs the matrix of Euclidean distances between the variables (pixels in our case) and the desired dimensionality of the embedding. Its output is a set of coordinates for each variable in the constructed embedding. t-SNE constructs the embedding by minimizing the overall distortion between the geometric arrangement of variables evaluated in the original and in the embedding space, as captured by the Kullback-Leibler divergence, or D_{KL}. The distortion is nonuniformly penalized, so that preserving distances between nearby variables is prioritized over those between faraway variables. It is not guaranteed to find a globally optimal embedding, and its solution are controlled by the initial condition and by two additional input parameters: the *perplexity* and the *early exaggeration* (EE). The former determines the effective number of "neighbors" (variables that are allowed to affect the embedding coordinates of a given variable), while the latter controls the global clumpiness of the embedding. We specifically chose t-SNE over other methods designed to produce geometric embeddings of data because strong correlations (small distances) in our data set are well sampled, whereas small correlations (and hence large distances) suffer from large statistical noise. Thus, preserving them in the embedding is not crucial (cf. Fig. 2).

To use t-SNE for embedding the shuffled pixels, we first transform their correlation matrix ($\mathbf{C}$) into a distance matrix ($\mathbf{D}$) by $\mathbf{D} = \exp(-\mathbf{C})$. We tried other transformations, which gave comparable results. Passing the distance matrix $\mathbf{D}$ to t-SNE, we get the embedding coordinates (the reconstruction) in the space of the requested dimensionality for each pixel $\mathbf{x}_{\mathrm{reconst}} = f(\mathbf{x}^*)$ in the original data set. If the embedding is in two dimensions (2D), in addition to the KL divergence, we can evaluate the quality of the reconstruction by directly comparing the original and the reconstructed image coordinates. To avoid issues related to global rotations and the mirror symmetry, which t-SNE cannot recover, we perform the comparison by calculating the correlation coefficient between distances of every pair of pixels in the original image on the one hand, $d(\mathbf{x}, \mathbf{x}')$, and the distances between their coordinates in the reconstructed embedding on the other, $d(\mathbf{x}_{\mathrm{reconst}}, \mathbf{x}'_{\mathrm{reconst}}) = d(f(\pi(\mathbf{x})), f(\pi(\mathbf{x}')))$.

Results. To select the appropriate perplexity, we note that larger perplexity values increase the number of neighbors that inform the algorithm about a pixel's position, and hence reduce statistical fluctuations that can come from having a small number of neighbors. However, in our data, the pairwise correlation bands start to dip below zero at a radius of $r \approx 12$ pixels; cf. Fig. 2. Thus, there are at most about $n_{\mathrm{c}} = \pi r^2 \approx 450$ pixels that are positively correlated with a pixel far away from the boundary and can be considered its useful neighbors for reconstructing the pixel's position. Pixels at the boundary have even fewer neighbors. Further, most of n_{c} positive correlations are low, making them uninformative, and potentially even detrimental. Therefore, we expect a good reconstruction of local arrangements with perplexity of only a fraction of n_{c}, and a worsening quality if more neighbors are included. To further develop this intuition, we notice that T images made from $N = T/q < T$ independent pixels will result in nonzero correlations purely due to statistical fluctuations when $q \sim 1$. The celebrated Marchenko-Pastur bound [11] suggests that we should not trust eigenvalues of covariance matrices smaller than

$$\lambda_+ = \sigma^2(1 + \sqrt{q})^2, \tag{3}$$

where σ^2 is the variance of an individual pixel. On the other hand, correlations stronger that λ_+ are probably reliably known. In other words, if there are $n(\lambda_+)$ eigenvalues in the pixel-pixel correlation matrix $\mathbf{C}$ that are above the λ_+ cutoff, then $n(\lambda_+)$ linear combinations of pixels can be used reliably to embed a given pixel. Thus $n(\lambda_+)$ provides the lower estimate on the optimal perplexity range, while n_{c} is the upper bound. As Fig. 3 shows, empirically, the optimal perplexity is closer to $n(\lambda_+)$ than to n_{c}, and we will use the optimal perplexity $p_{\mathrm{opt}} = n(\lambda_+)$ in all analyses, unless otherwise specified.

We know of no heuristics to set a good EE value from first principles, as a robust theory of hyperparameter selection in t-SNE based on a given data set does not exist. Rather, hyperparameter selection tailored to data sets previously involved trying combinations of hyperparameters (including learning rate and steps) within large ranges [20,21]. Thus, we explore a range of EE values in the analyses below.

Figure 3(a) shows D_{KL} for the quality of the embedding in 2D. For all but the smallest EE values, D_{KL} decreases until the perplexity reaches λ_+, then stays relatively flat until it starts increasing well below the naive estimate of 450. Similarly, the correlation between the original distances between pairs of pixels in the nonpermuted image and the distances between the same pairs in the embedding space in Fig. 3(b) shows an improving correlation up to the λ_+ bound—to nearly perfect values—and then a dropoff soon after, when too many noisy pixel pairs are used to estimate the local structure. Crucially, these data suggest that small EE is detrimental, but there is little difference in its value past ~10 in the 2D embedding.

The weak dependence of neighborhood reconstruction on perplexity suggests that there is enough information about the local structure in just a handful of pairwise distances, so that one can achieve a very good reconstruction even for much smaller sample sizes. Indeed, smaller data sets will have a higher noise in weak correlations between far away pixels. However, a few strong correlations among nearby pixels will

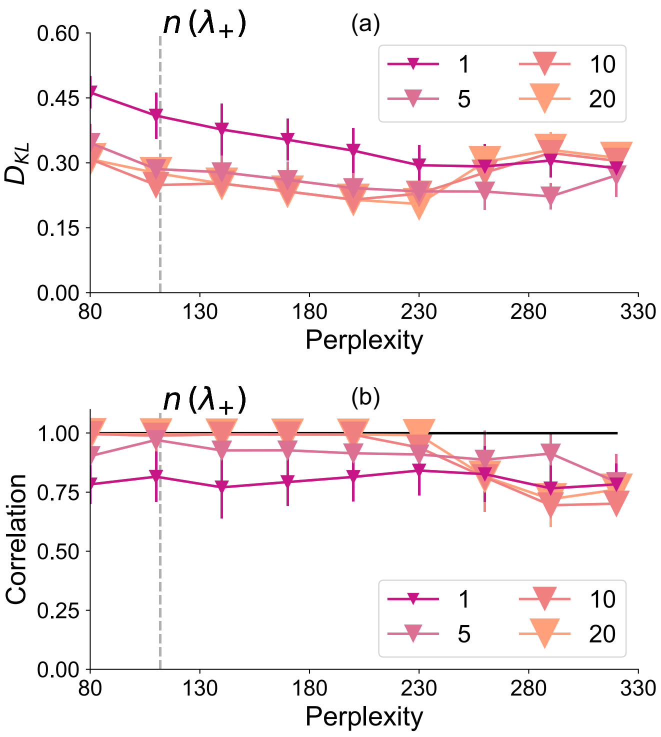

FIG. 3. DKL and correlation between reconstructed and true distances. (a) The quality of the optimal 2D embedding found by t-SNE as a function of the early exaggeration (EE) and the perplexity, averaged over 50 runs, with error bars indicating sample standard deviations. The number of eigenvalues above the statistical significance threshold, $n(\lambda_+)$, is shown. Different colors and marker sizes indicate different EE values. (b) The average correlation between the pixel-to-pixel distances in the nonpermuted data and t-SNE-embedded permuted data. Plotting notations are the same as in (a). Dead leaves (training) data are used in both panels.

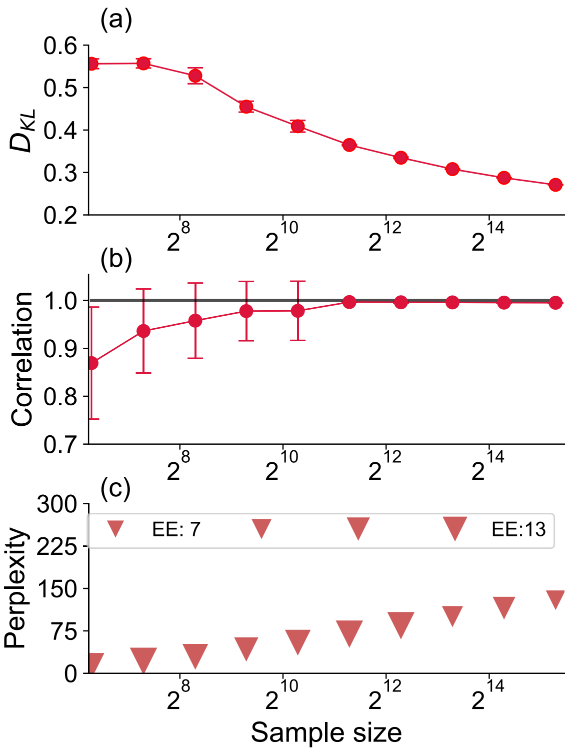

FIG. 4. Effect of the data set size on the detection of local relations. (a) The minimum D_{KL} obtained by t-SNE as a function of the sample size T. (b) The correlation the true distances and t-SNE reconstructed distances. Both panels show averages over ten realizations, and error bars too small to be seen in panel (a) are the standard deviations over the realizations. (c) The optimal perplexity $p_{opt} = n(\lambda_+)$ and the optimal EE value (denoted by the marker size) that achieved the lowest D_{KL} at p_{opt}. The optimal EE barely changes over a 1000-fold change in T.

still have a small relative error, and this should be sufficient for the reconstruction if a correct perplexity is chosen. We verify this in Fig. 4, where we plot the reconstruction quality as a function of the sample size, T. The quality decreases gracefully as T decreases. The average reconstruction correlation reaches >0.9 at just $T = 157$ images with $n < 30$ neighbors used for the reconstruction. In comparison, there are 2500 pixels and $>3 \times 10^6$ pairwise distances in the images, so that extremely undersampled data sets are still sufficient for establishing locality.

In real data, the correct embedding dimension is often unknown, undefined, or varies across the data set. When possible, it should be inferred from the data directly. Generically, the larger the embedding dimension, the easier it is for t-SNE to produce embeddings with a smaller D_{KL} simply because it is easier to satisfy distance relations implied by **D**. This results in the drop in D_{KL} between the embedding dimensions of 1 and 2 in Fig. 5. However, this freedom comes at a cost that, in higher dimensions, there are many embedding configurations that preserve some relative distances, but distort the global geometry of the image. This makes the reconstruction susceptible to EE, which controls the clustering of the embedding, increasing the initial distance between clusters while clumping the points within the same cluster. As shown in Fig. 5, at high EE and in high dimensions, t-SNE fails to recover from extremely clumpy initializations, resulting in large D_{KL}. Only at or below the true dimensionality of 2 is the algorithm insensitive to EE. This suggests a simple approach to detecting the intrinsic dimensionality of data: look for the

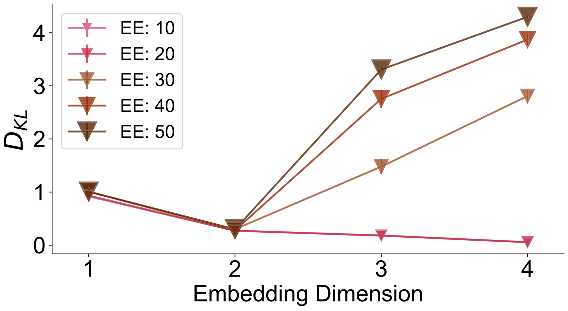

FIG. 5. Embedding quality in different dimensions. D_{KL} for the optimal embedding is plotted for dimensions of 1 to 4 (averaged over ten realizations, with error bars—too small to see—representing the standard deviation of the set). For our dataset of 2D images, D_{KL} drops between 1D and 2D uniformly for all EEs, and then it becomes strongly susceptible to EE, indicating that the intrinsic dimensionality of the data is 2.

largest embedding dimension, at which the reconstruction D_{KL} is the lowest uniformly, independently of the EE. We emphasize this finding: with no assumptions about the structure of the data, t-SNE can recover its intrinsic dimensionality. This bodes well for applications to more interesting data sets. However, Cauchy distributions, which allow t-SNE to preserve local structures while breaking global relations, do not work well in higher dimensions [16]. Thus applicability of this simple approach to large-dimensional data requires additional investigation.

Finally, to illustrate the weak sensitivity of the recovery of the local relations to the nature of the training set, we visualize the reconstructed images. For this, we place a grey or a black dot, corresponding to the color of the shuffled pixel in the original image, at the optimal 2D t-SNE reconstructed coordinates $\mathbf{x}_{\mathrm{recon}}$. We then rescale the reconstructed image to be 50×50 pixels in size (matching the original images), and rotate them to better align to the original images (notice that the rescaling and the rotation are the same for *all* images). This is needed because t-SNE focuses on local geometry only and is unaware of possible global transformations to the data. We show the results for a few select training and test images in the right column of Fig. 1. The quality of the reconstruction for both data sets is visually exceptional, even though the correlation structure is different in both data sets; cf. Fig. 2.

Discussion. Building probabilistic models $P(I(\mathbf{x}))$ of large dimensional biological systems of realistic sizes is a combinatorial challenge, especially as $N > T$. Regularizing the learning problem requires knowing local relations and hence which variables are allowed to interact. Such relations may or may not correspond to neighborhoods in the real space, but some notion of effective locality is crucial for model building. Here we showed that, at least in the context of 2D images, pairwise correlations among observed variables alone are sufficient to recover the locality structure and the dimensionality of the data set using an off-the-shelf visualization algorithm, t-SNE (though we emphasize again that we view t-SNE not as a crucial choice of our approach, but as just one of the methods that we could have used). We provided semiquantitative heuristics based on well-known calculations from the random matrix theory to set parameters of t-SNE to determine the optimal embedding dimension and the optimal reconstruction for this dimension. The resulting inference is so good that, even in the deeply undersampled regime, the permuted images can be fully reconstructed, with a very weak sensitivity to the details of the training set.

The key insight from this analysis is that, even when the sample size is insufficient to estimate all correlations in data, a few strong correlations can still be estimated well. The number of such correlations is then enough to predict the position of a pixel relative to its "neighbors," even if relations to the more distant pixels are unknown. One can then reconstruct the full structure of the data set by traversing the neighborhoods of the pixels individually, which is what t-SNE (and other related low-dimensional visualization algorithms) does implicitly.

Real-world data are usually more complex than 2D images considered here. Dimensionality of such data often depends on the point in the state space and is generally hard to define (see, e.g., Ref. [22] for discussion of this for animal behavior data). Common generative models of real data often are based on low-dimensional or sparse *latent* (rather than visible) structures (see, e.g., Ref. [23] for a relevant discussion in the context of neurobiology). Such latent models also can generate data of varying or undefined global dimensionality in the visible space. Making generalization from a relatively simple problem of 2D images to such complex data is hard. However, we expect that, while direct application of t-SNE or other algorithms that assume a small fixed dimensionality to more complex data may be questionable, our main realization that determining *local* structures only requires a few strong correlations irrespective of the *global* properties of the data is likely to hold for them as well. Thus, we expect that determining local relations and then constraining statistical models to only include interactions within them will decrease the number of parameters that must be inferred to build a model with N variables from $O(2^N)$ to $O(N)$, making such models experimentally tractable. Applications to real data in diverse domains can be plentiful, as possibilities include identifying functional 3D protein structure from coevolution based amino acids [3], chromosomal structure based on base-pair contacts [24], neural activity from correlated firing patterns [2], or "favored" interactions in genome space [25]. All of these biological processes are similar in that the number of possible combinatorial interactions greatly exceeds possible sizes of experimental datasets. Unknown physical constraints create an effective locality, which is distinct from the geometric locality (e.g., allostery allows for long-range interactions in proteins), and it is precisely this effective locality—*a priori* unknown—that offers a possibility for model building. Understanding for which of these domains, if any, our approach works or not and why is likely to be a fruitful direction for future research.

Success of modern machine learning in image analysis has largely been driven by convolutional neural networks, which decrease the dimensionality of the learned statistical model by imposing locality and translational invariance on images [26]. Our analysis suggests that the local structure can be inferred from just a handful of images. Thus, there should exist algorithms for training general fully connected neural networks so that, compared to the convolutional networks, they are only minimally handicapped by not knowing the structure *a priori*.

We point out an intriguing speculative connection to the transformer neural architecture [27] and, more generally, other machine learning models with the so called *attention* mechanism [28]. Transformers are behind the recent success of large language models, such as GPT-3 [29]. They build on an established idea that the identity of a word that should occur at a certain place in text can be determined by the context; that is, by the conditional distribution of a word on its surroundings [30]. However, contexts, specified by identities of words and their relative position in text, must be very large to be practically useful. Hence they are severely undersampled even for exceptionally large training sets. Transformers solve this problem by using the attention mechanism to only focus on the parts of the context, whose correlations with the target are well sampled. A potential future research direction involves exploring if such attention is similar to how we were able to detect an attribute (position) of a pixel by only focusing on highly correlated pixels and ignoring noisy, undersampled relations. If a rigorous connection can be identified, one may be able to use random matrix theory-based estimates, similar to our current analysis, to understand the data set sizes needed for attention-based models to work.

Acknowledgments. We are grateful to Sean Ridout for insightful comments about the manuscript. I.N. was supported, in part, by the Simons Investigator Award. M.R. was supported, in part, by Emory University internal funds.

[1] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Machine learning and the physical sciences, Rev. Mod. Phys. **91**, 045002 (2019).

[2] C. Savin and G. Tkačik, Maximum entropy models as a tool for building precise neural controls, Curr. Opin. Neurobiol. **46**, 120 (2017).

[3] D. S. Marks, L. J. Colwell, T. A. Hopf, A. Pagnani, R. Zecchina, and C. Sander, Protein 3D structure computed from evolutionary sequence variation, PLOS ONE **6**, e28766 (2011).

[4] D. L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory **52**, 1289 (2006).

[5] M. Bailly-Bechet, A. Braunstein, A. Pagnani, M. Weigt, and R. Zecchina, Inference of sparse combinatorial-control networks from gene-expression data: A message passing approach, BMC Bioinf. **11**, 355 (2010).

[6] S. Ganguli and H. Sompolinsky, Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis, Annu. Rev. Neurosci. **35**, 485 (2012).

[7] N. Bulso, M. Marsili, and Y. Roudi, Sparse model selection in the highly under-sampled regime, J. Stat. Mech.: Theory Exp. (2016) 093404.

[8] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, Proc. Natl. Acad. Sci. USA **99**, 7821 (2002).

[9] A. Clauset, Finding local community structure in networks, Phys. Rev. E **72**, 026132 (2005).

[10] E. Ravasz and A. Barabási, Hierarchical organization in complex networks, Phys. Rev. E **67**, 026112 (2003).

[11] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists* (Cambridge University Press, Cambridge, 2020).

[12] D. L. Ruderman and W. Bialek, Statistics of Natural Images: Scaling in the Woods, Phys. Rev. Lett. **73**, 814 (1994).

[13] G. Matheron, *Random Sets and Integral Geometry* (John Wiley and Sons, New York, 1974).

[14] A. B. Lee, D. Mumford, and J. Huang, Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model, Int. J. Comput. Vis. **41**, 35 (2001).

[15] X. Pitkow, Exact feature probabilities in images with occlusion, J. Vis. **10**, 42 (2010).

[16] L. van der Maaten and G. Hinton, Visualizing data using t-SNE, J. Machine Learn. Res. **9**, 2579 (2008).

[17] J. Tenenbaum, V. de Silva, and J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science **290**, 2319 (2000).

[18] M. Belkin and P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. **15**, 1373 (2003).

[19] L. McInnes, J. Healy, N. Saul, and L. Großberger, UMAP: Uniform manifold approximation and projection, J. Open Source Softw. **3**, 861 (2018).

[20] R. Gove, L. Cadalzo, N. Leiby, J. M. Singer, and A. Zaitzeff, New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation, Visual Inf. **6**, 87 (2022).

[21] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets, Nat. Commun. **10**, 5415 (2019).

[22] G. Berman, Measuring behavior across scales, BMC Biol. **16**, 23 (2018).

[23] S. Vyas, M. Golub, D. Sussillo, and K. Shenoy, Computation through neural population dynamics, Annu. Rev. Neurosci. **43**, 249 (2020).

[24] E. Lieberman-Aiden, N. L. Van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner *et al.*, Comprehensive mapping of long-range interactions reveals folding principles of the human genome, Science **326**, 289 (2009).

[25] T. Mora, A. Walczak, W. Bialek, and C. Callan, Maximum entropy models for antibody diversity, Proc. Natl. Acad. Sci. USA **107**, 5405 (2010).

[26] W. Rawat and Z. Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. **29**, 2352 (2017).

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. **30**, 6000 (2017).

[28] D. Bahdanau, K. Cho, and Y. Bengio, Neural machine translation by jointly learning to align and translate, *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*, edited by Y. Bengio and Y. LeCun (2015).

[29] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, Language models are few-shot learners, Adv. Neural Inf. Process. Syst. **33**, 1877 (2020).

[30] F. Pereira, N. Tishby, and L. Lee, Distributional clustering of English words, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Columbus, Ohio, USA, 1993), pp. 183–190.