Diffusion t-SNE for multiscale data visualization

Lan Huong Nguyen[*]
ICME
Stanford University
Stanford, CA 94305
lanhuong@stanford.edu

Susan Holmes
Department of Statistics
Stanford University
Stanford, CA 94305
susan@stat.stanford.edu

1 Introduction

Neighbor embedding (NE) methods generate reduced data representations, where points originally close in the input are also close in the output, but where points originally distant are not necessarily distant in the output embedding. This property is desirable when one is interested only in recovering the inherent grouping membership and obtaining a good cluster separation [1, 6]. t-SNE by Van der Maaten and Hinton [11] is a very popular dimensionality reduction method for visualizing high-dimensional data. The algorithm has been successfully used on many real datasets for which it has exposed the underlying data groupings; it performs particularly well when applied to datasets composed of well-separated clusters. In particular, the method is now often used on biological datasets, e.g. on genomics or mass cytometry data to produce visualizations of groups of cell populations and sub-populations [4, 7, 10].

Despite its popularity, t-SNE is unable to correctly represent the local shifts in data density or variance. It has been recognized that cluster sizes and large-scale distances are not interpretable in t-SNE output embeddings [12]. While this property might be harmless when the input data is inherently categorical and one is only interested in discriminating distinct classes of observations (e.g. separating images of cats and dogs), the inability to make comparisons between different regions of the data is a major limitation when visualizing datasets governed by unknown latent continuous factors. By design, t-SNE stretches and shrinks different fragments of the data at different rates. Data distortions are most striking when the sampling density is highly variable across regions. Since many datasets, especially the biological ones, are characterized not by discrete but continuous latent variables, often associated with an unknown continuous process, it is important to use data visualization tools that can accurately capture both the local neighborhoods and the long-range interactions.

Here, we characterize the part of the t-SNE algorithm that induces its undesirable properties. Then, we (1) propose a scaling procedure to recover the relative differences in regional variances, and (2) develop a new algorithm, *Diffusion t-SNE*, for generating multiscale data visualizations. More specifically, we show that the pairwise similarity computations adopted by t-SNE and other NE methods are responsible for both the data contraction and expansion at varying degrees across different neighborhoods as well as for the loss of information of the long-range interactions.

2 Results

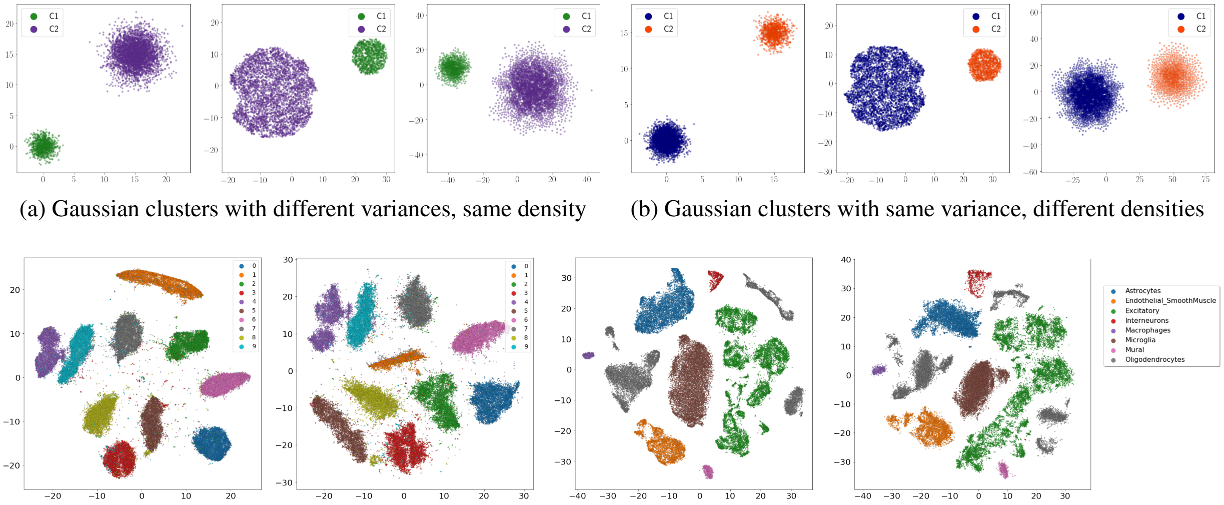
The t-SNE algorithm computes the output embedding coordinates by minimizing the Kullback-Leibler (KL) divergence, $C = KL(P\|Q) = \sum_{i,j} p_{ij} \log\left(p_{ij}/q_{ij}\right)$, between the input ($p_{ij}$) and output ($q_{ij}$) inter-point affinities (see [11] for details). The input affinities, $p_{ij} = (p_{j|i} + p_{i|j})/2N$, are defined as symmetrized *conditional probabilities* of observing a data point in a neighborhood of another, $p_{j|i}$, computed as follows:

$$p_{j|i} = \frac{\exp(-d_{ij}^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-d_{ik}^2/2\sigma_i^2)}, \ \forall j \neq i \ \text{ and } \ p_{i|i} = 0. \tag{2.1}$$

where d_{ij} denotes the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$. The σ_i^2's are data point specific, and their values are set to maintain a fixed entropy level (determined by a user-specified perplexity parameter, η) for all data points. More specifically, for all $i = 1, \ldots, N$:

$$\log(\eta) = -\sum_{j\neq i} p_{j|i}\log(p_{j|i}) = \sum_{j\neq i}\frac{d_{ij}^2}{2\sigma_i^2}p_{j|i} + \log\sum_{j\neq i}\exp(-d_{ij}^2/2\sigma_i^2). \tag{2.2}$$

[*] Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.



(a) Gaussian clusters with different variances, same density

(b) Gaussian clusters with same variance, different densities

(c) MNIST Dataset

(d) Hrvatin et al. Dataset

Figure 1: Comparison between standard and scaled t-SNE embeddings. From left to right we have panels showing the original data (a, b only), followed by the 2D output embeddings with standard, and then scaled t-SNE (a-d).

There is no closed-form solution to the above expression, and the σ_i's are usually computed through a binary search. The t-SNE's entropy equalization procedure was originally introduced to accommodate non-uniform data density, as variable bandwidth Gaussian kernels can resolve the interactions between nearby neighbors to the same degree either when applied to dense or sparse regions. Unfortunately, equalizing the kernel resolution across all regions also leads to undesirable distortions and loss of information in the output embedding. We show that density and variance are unidentifiable in the t-SNE embedding.

Theorem 2.1. *Two clusters with different volumes (or within-cluster variance) and densities may be indistinguishable in a t-SNE output embedding. In particular, two neighborhoods with variances ν^2, $\nu^2 r^2$, and densities ρ, ρr^p respectively appear to have the same sizes in the t-SNE output embedding.*

The proof of the above theorem can be found in the appendix. We further show that the conditional probabilities can be reformulated as functions of shifted and scaled dissimilarities, $p_{j|i} = \frac{1}{\eta}\exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2)$, where $\bar{d}_i^2 = \sum_{j\neq i} d_{ij}^2 p_{j|i}$.

Since shifting and scaling of dissimilarities occur at nonuniform rates for different i's, the information on local differences in variance and density is lost. Consequently, the cluster sizes are uninformative in t-SNE maps. Intuitively, t-SNE treats the k-nearest-neighbors in a very sparse region the same as the k-nearest neighbors in a dense region, even though the distances between neighbors in these two regions can vary significantly. In order to account for the differences in variance, the conditional probabilities must capture the differences in density along the dataset.

2.1 Preserving local differences in data variance

To alleviate this loss of information, we propose a scaling scheme where the rows of the similarity matrix $(P)_{ij} = p_{j|i}$ are multiplied by factors inversely proportional to the selected bandwidth value. In other words, we use the following formula for adjusted affinities:

$$\tilde{p}_{j|i} = \alpha_i \frac{\exp(-\beta_i d_{ij})}{\sum_{k\neq i}\exp(-\beta_i d_{ik})}, \quad \text{where} \quad \alpha_i = N\frac{\beta_i}{\sum_k \beta_k}, \quad \text{and} \quad \beta_i = \frac{1}{2\sigma_i^2}. \tag{2.3}$$

Scaling $p_{\cdot|i}$ terms by an α_i factor recovers relative differences in the sizes of the neighborhoods considered for each data point i. The resulting affinity between i and its neighbors, $\mathcal{N}_i$, are larger if i is in a dense region, and smaller if it is in a sparse region. Intuitively, larger similarities should be assigned to data points with smaller mean distance to the k-nearest neighbors. Using α_i factors and β_i bandwidths is simply a smooth alternative to using k-nearest neighbors to evaluate inter-point similarities.

We demonstrate the effectiveness of our *scaled t-SNE* on a simulated Gaussian clusters dataset, the handwritten digits MNIST dataset, and the single-cell dataset by Hrvatin et al. [4]. Fig. 1 provides a comparison between standard and scaled t-SNE embeddings for each dataset. Standard t-SNE distorts the original data in an uncontrolled way, often expanding the highly dense (e.g. blue cluster in (b)) or low variance (e.g. orange cluster in (c)) areas. This leads to unidentifiability between fluctuations in variances and the ones in densities. For example, the standard t-SNE embeddings (middle panels) appear the same for two distinct cases: a pair of Gaussian clusters with the same density but different variances (a), and a pair of Gaussian clusters with different densities and the same variance (b). In contrast, our scaling procedure generates output embeddings (right-most panels) where cluster sizes are consistent with the underlying truth despite the cluster imbalance in the number of member observations. The scaled t-SNE embeddings correctly exhibit differences between the two Gaussian clusters in (a) and (b). Unlike standard t-SNE, the scaled t-SNE outputs embeddings where, as expected, in (c) the orange cluster corresponding to the least variable digit "1" appears the smallest. For the Hrvatin dataset (d), the standard t-SNE plot (left) shows all clusters to have the same densities; consequently, their sizes reflect only the number of samples included. Scaled t-SNE shows differences in densities between different clusters, with the brown cluster of microglia now much denser but of the same size as the orange cluster of endothelial, smooth muscle cells; this implies the two clusters exhibits roughly similar levels of inter-cell variability. Scaled t-SNE allows one to discriminate between the local differences in variances and and those in data sampling densities, as unlike the standard t-SNE, it does not even out the cluster densities in the output embedding.

2.2 Recovering multiscale structures

Learning the global geometry is most challenging when data lies approximately on a non-linear manifold. When non-linear structures are present, one must "unfold" the latent manifold to learn the global data geometry; it is necessary to accurately learn the long-range interactions between data points. While t-SNE maps preserve the local neighborhood identities, they usually provide a poor representation of the large-scale structures. Since t-SNE utilizes an exponentially decaying Gaussian kernel to transform dissimilarities into inter-point affinities, the medium and long range interactions are all converted to similarly negligible values. Consequently, the information on the relationship between distant data points is effectively lost and the global landscape of the input data cannot be retrieved.

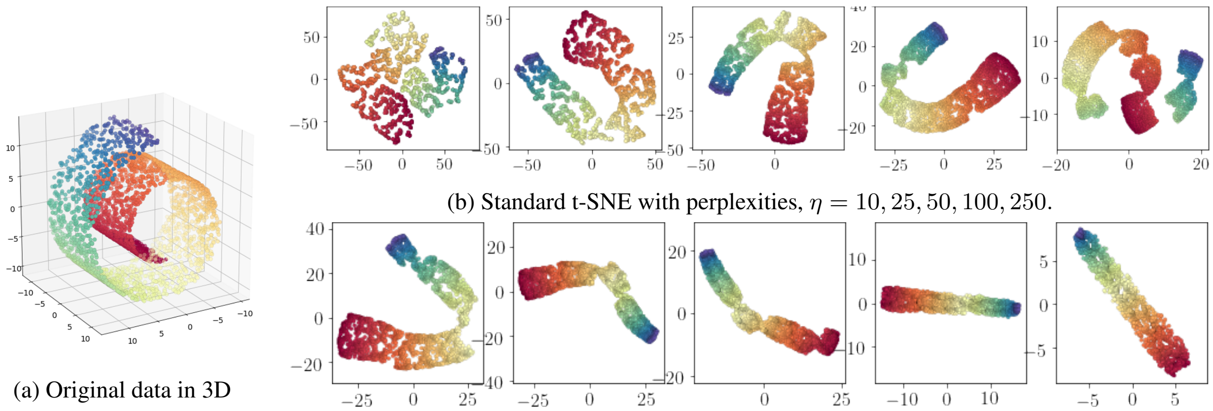
To counter this weakness, we leverage the fact that t-SNE's conditional probabilities are conceptually similar to transition probabilities used in Diffusion Maps (DM) [2] to characterize the directions of fast and slow propagation of a random walk defined on the dataset. Running a Markov chain forward or taking powers of the transition probability matrix, is equivalent to propagating and integrating the local information to obtain a global profile of the entire dataset. Thus, we propose a new *diffusion t-SNE* algorithm, where the t^{th} power of the conditional probability matrix, $P^t_{\text{cond.}}$ (with $P_{\text{cond.}}$ defined in Eq 2.1), is used to incorporate the information on the long-range interactions into the computation of the inter-point affinities. The newly defined "integrated proximity" measures allow the KL divergence optimization problem to find an output embeddings that can accurately reflect the large-scale structures in the input data. Diffusion t-SNE allows the user to generate different views of the data by computing multiple embeddings with different choices of t.

2.2.1 Swiss roll

Manifold learning techniques are often tested on their ability to unfold a Swiss roll, Fig. 2(a). We use this example to highlight the t-SNE's inability to accurately represent the global structure of non-Euclidean datasets. Fig. 2 (b) shows the output embeddings of a Swiss roll generated with standard t-SNE using different perplexity settings. For all values of the parameter, t-SNE fails to unfold the flat surface curved in 3D. Even for large perplexity, the standard algorithm returns disconnected patches of the continuous latent sheet.

This simple example challenges the popular belief that increasing perplexity values improves the global structure recovery [5]. In practice, raising perplexity simply increases the kernel bandwidths used in computations of the inter-point affinities, and effectively expands the areas considered the local neighborhoods. This implicitly assumes that moderate distances measured on the input data can accurately capture the relationships between moderately distant data points. Unfortunately, the moderate and large distances measured in the ambient space cannot approximate the intrinsic distances well if the data manifold is highly curved. Thus, in general increasing the perplexity could not improve the recovery of the large-scale structures of non-linear datasets.

In contrast, diffusion t-SNE utilizes powers of the transition matrix and is able to reveal the structure of a Swiss roll at varying scales. Keeping a small perplexity value, $\eta = 25$, that corresponds to small kernel bandwidths, and increasing the value of the time step parameter, $t = 5, 10, 20, 50, 100$, allows the methods to successfully unfold the curved manifold, Fig. 2 (c).



(a) Original data in 3D

(b) Standard t-SNE with perplexities, $\eta = 10, 25, 50, 100, 250$.

(c) Diffusion t-SNE with perplexity, $\eta = 25$, and time step, $t = 5, 10, 20, 50, 100$.

Figure 2: Swiss roll (a) embedding. Comparing effect of varying perplexity value in standard t-SNE (b) and varying time step parameter in Diffusion t-SNE (c).

2.2.2 Embryoid Body differentiation

We also test our method on a single cell expression dataset from a 27-day time course study of the embryoid body (EB) differentiation generated by Moon et al. [8]. The dataset [2] contains scRNA-seq samples from human embryonic stem cells (hESCs) differentiating to embryoid bodies (EB), collected in 3-day intervals. We use the same data pre-processing steps as the ones performed by the authors of the original paper. [3] We then compute the first 50 principal components on the processed single cell data and apply standard and diffusion t-SNE to generate 2D embeddings.

Since the dataset is governed by a differentiation process, the latent structure should involve gradual, continuous changes rather than distinct clusters, as in cases of categorical image classes. It is thus important to recover accurate structures present in the input data at different scales, i.e. it is of value to show both the relationship between cells collected on the same days (small-scale), and the differentiation patterns exhibited by cells across different stages (large-scale).

Standard t-SNE generates embeddings with disconnected patches even at a $\eta = 1000$; a red data patch is incorrectly disconnected from the remaining stem cells visible in the bottom part of the right-most panel in Fig. 3 (b). This behavior is similar to the one observed for the Swiss-roll in Fig. 2 (b), where t-SNE with $\eta = 250$ destroys the global data geometry by tearing apart the underlying structure.

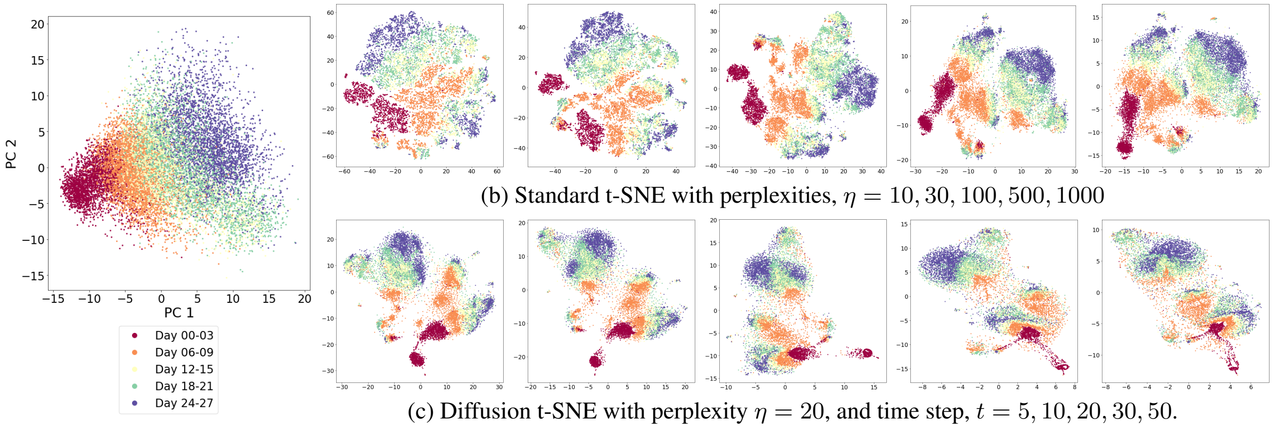
In Fig. 3 (c), we showed that using diffusion t-SNE with increasing t parameters produces multi-scale visualizations of the data, consistently capturing small to large scale structures. Unlike the standard t-SNE, our algorithm always shows all the red stem cells (day $0 - 3$) closely together. Cells at later stages are more spread out across the embedding space, as expected due to the differentiation process. The embedding configuration becomes more connected as the value of t increases, implying a shift of focus from local to global.

3 Discussion and conclusion

In this work, we study the properties of t-SNE, focusing on the limitations due to the method's data transformation procedures. We note many other embedding techniques, including LargeVis [9] or UMAP, exploit similar technique involving a conversion of distances to inter-point affinities. In particular, the varying bandwidth Gaussian kernels are widely implemented to estimate pairwise proximities. As derived in this article, the t-SNE's conditional probabilities can be expressed as inverse exponential on shifted and scaled pairwise distances, where the shifting and scaling is applied using different factors for each data point. This results in distortions in the output embedding with the most noticeable effects occurring when the input data is unevenly sampled across the observed space. In general, variance and density are unidentifiable, while cluster sizes and the distances between cluster centroids are not interpretable in the t-SNE output embedding. We introduce two effective modifications to t-SNE that help alleviate these undesired characteristics: (1) a scaling scheme that allows the inter-point affinites to retain information on the regional fluctuations in variance and (2) a diffusion t-SNE algorithm that, by varying a time step parameter, can recover the data geometry at different scales, including the global structure.

[2] EB data downloaded from: `https://data.mendeley.com/datasets/v6n743h5ng/1`

[3] EB data pre-processing steps: `https://github.com/KrishnaswamyLab/PHATE`



(b) Standard t-SNE with perplexities, $\eta = 10, 30, 100, 500, 1000$

(c) Diffusion t-SNE with perplexity $\eta = 20$, and time step, $t = 5, 10, 20, 30, 50$.

(a) PCA

Figure 3: Embryoid Body dataset visualization with PCA (a), standard t-SNE (b) and diffusion t-SNE (c) for different choices parameters.

References

[1] Sanjeev Arora, Wei Hu, and Pravesh K. Kothari. An analysis of the t-SNE algorithm for data visualization. *CoRR*, abs/1803.01768, 2018.

[2] Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5 – 30, 2006. Special Issue: Diffusion Maps and Wavelets.

[3] Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392), 2018.

[4] Sinisa Hrvatin, Daniel R. Hochbaum, M. Aurel Nagy, Marcelo Cicconet, et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nature Neuroscience*, 21(1):120–129, 2018.

[5] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *bioRxiv*, 2019.

[6] George C. Linderman, Manas Rachh, Jeremy G. Hoskins, Stefan Steinerberger, and Yuval Kluger. Efficient algorithms for t-distributed stochastic neighborhood embedding. *CoRR*, abs/1712.09005, 2017.

[7] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202 – 1214, 2015.

[8] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, et al. Visualizing structure and transitions for biological data exploration. *bioRxiv*, 2019.

[9] Jian Tang, Jingzhou Liu, Ming Zhang, and Qiaozhu Mei. Visualizing large-scale and high-dimensional data. In *Proceedings of the 25th International Conference on World Wide Web*, pages 287–297. International World Wide Web Conferences Steering Committee, 2016.

[10] Bosiljka Tasic, Zizhen Yao, Lucas T. Graybuck, Kimberly A. Smith, et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.

[11] Laurens J. P. van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[12] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-SNE effectively. *Distill*, 2016.

4 Appendix

4.1 Unidentifiability of density and variance

Theorem 2.1. *Two clusters with different volumes (or within-cluster variance) and densities may be indistinguishable in a t-SNE output embedding. In particular, two neighborhoods with variances ν^2, $r^2\nu^2$, and densities ρ, $r^p\rho$ respectively appear to have the same sizes in the t-SNE output embedding.*

Proof. Consider two well separated clusters $X_k^{(1)} \in \mathcal{C}_1 \subset R^p$ and $X_l^{(2)} \in \mathcal{C}_2 \subset \mathbb{R}^p$. Let $\mathcal{C}_1$ and $\mathcal{C}_2$ be hyper-spheres of radii R_1 and R_2, where $R_2 = rR_1$. Further let $\{X_k^{(1)}\}_{k=1}^N$ and $\{X_l^{(2)}\}_{i=1}^N$ data points be sampled uniformly at random from $\mathcal{C}_1$ and $\mathcal{C}_2$ respectively. The two clusters have unequal densities $\rho_1 = \rho_2/r^p$ and occupy different volumes, $V_1 = r^p V_2$. Suppose $X_i^{(1)}$ lies approximately near the center of $\mathcal{C}_1$ and $X_j^{(2)}$ near the center of $\mathcal{C}_2$, with $\|X_i^{(1)} - X_j^{(2)}\|_2^2 \gg R_2$ (since well separated).

Let $D_{ii'} = \|X_i^{(1)} - X_{i'}^{(1)}\|_2$ be the distance from i to another data point in the same cluster $i' \in \mathcal{C}_1$, and similarly $D_{jj'}$ the distance of j to $j' \in \mathcal{C}_2$. Note that these distances can be expressed as uniform random variables, $D_{ii'} = R_1 U_{i'}$ and $D_{jj'} = R_2 U_{j'}$ with $U_{i'}$, $U_{j'} \sim \text{Uniform}_{[0,1]}$. Let $Z_{ii'} = D_{ii'}^2/2\sigma_1^2 = R_1^2 U_{i'}^2/2\sigma_1^2$, and $Z_{jj'} = D_{jj'}/2\sigma_2^2 = r^2 R_1^2 U_{j'}^2/2\sigma_2^2$. Using the t-SNE conditional probability definition in (2.1), we have:

$$p_{i'|i} \approx \begin{cases} \frac{\exp(-Z_{ii'})}{\sum_{k \neq i, k \in \mathcal{C}_1} \exp(-Z_{ik})} & i' \in \mathcal{C}_1 \\ 0 & i' \in \mathcal{C}_2 \end{cases}$$

The approximation in the denominator removes the negligible proximity terms corresponding to pairs of observations from different clusters. Similarly, we have $p_{j'|j} \approx \frac{\exp(-Z_{jj'})}{\sum_{k \neq j, k \in \mathcal{C}_2} \exp(-Z_{jk})}$ for $j' \in \mathcal{C}_2$. Note that, t-SNE requires that σ_1 and σ_2 be set so that $p_{\cdot|i}$, $p_{\cdot|j}$ have equal Shannon entropy (2.2). Setting $\sigma_2^2 = r^2\sigma_1^2$, would satisfy this constraint, as $Z_{ii'}$ and $Z_{jj'}$ would have the same distribution.

However, with $\sigma_2^2 = r^2\sigma_1^2$, the of $X_i^{(1)}$ and $X_j^{(2)}$ to other data points, $p_{\cdot|i}$ and $p_{\cdot|j}$ would be indistinguishable. That is the affinity of $X_i^{(1)}$ to its k^{th}-nearest-neighbor (k-NN) would be roughly equal to the affinity of $X_j^{(2)}$ to its k-NN, $p_{j_k|j} \approx p_{i_k|i}$, despite the fact that the corresponding distances are roughly r times larger $D_{jj_k} \approx rD_{ii_k}$. For data points, i and j, not in the center of the clusters, the situation is similar and setting the bandwidth's respecting the same ratio, obtains indistinguishable $p_{\cdot|i}$ and $p_{\cdot|j}$ with equal entropy.

With the intra-cluster conditional probabilities indistinguishable from each other, $p_{\cdot|i}$ and $p_{\cdot|j}$, the information on the discrepancy in the within-cluster variance between $\mathcal{C}_1$ and $\mathcal{C}_2$ is lost. As a result, there is no possibility for the t-SNE algorithm to preserve the relative cluster sizes when finding the optimal embedding by minimizing the KL-divergence to the input affinities that do not resolve the differences between the two clusters.

4.2 Perplexity and the effective number of neighbors

Definition 4.1. For computed bandwidth parameters σ_i's, and a chosen error level, $0 < \epsilon < 1$, let the set of *effective neighbors* of the ith data point be $\mathcal{N}_i = \{j\colon \exp(-d_{ij}^2/2\sigma_i^2) > \epsilon/N \text{ and } j \neq i\}$, where the σ_i's are selected to meet the constraints specified in (2.2). Further let $\bar{d}_i^2 = \sum_{j \neq i} d_{ij}^2 p_{j|i}$ be the *effective distance* of data point i to its neighbors.

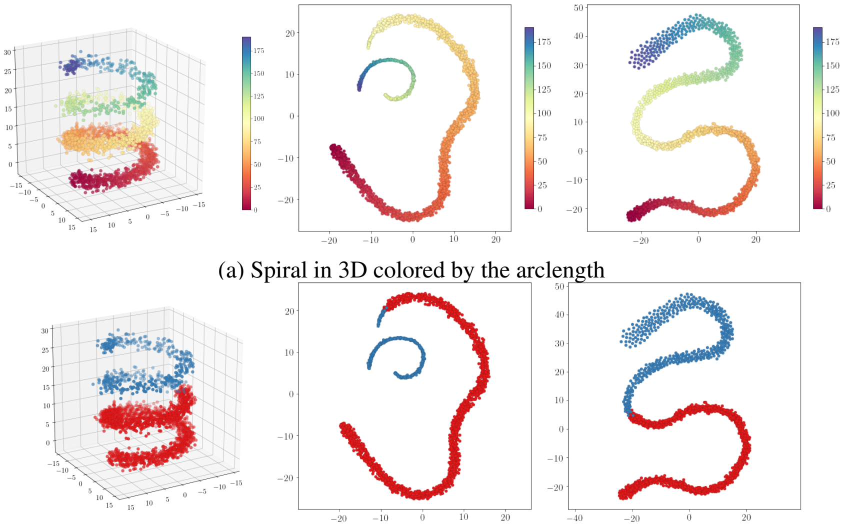
Bellow we show the relationship between the perplexity parameter, η, and the effective number of neighbors. First, we expand the formula for entropy of $p_{\cdot|i}$:

$$\begin{aligned} H_i &= \log\left(\sum_{j \neq i} \exp(-d_{ij}^2/2\sigma_i^2)\right) + \sum_{j \neq i} \frac{d_{ij}^2}{2\sigma_i^2} p_{j|i} \\ &= \log\left(\sum_{j \neq i} \exp(-d_{ij}^2/2\sigma_i^2)\right) + \log\left(\exp\left(\bar{d}_i^2/2\sigma_i^2\right)\right) \\ &= \log\left(\sum_{j \neq i} \exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2)\right) \\ &= \log\left(\sum_{j \in \mathcal{N}_i} \exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2) + \sum_{j \notin \mathcal{N}_i} \exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2)\right) \end{aligned}$$

where $\bar{d}_i^2 = \sum_{j \neq i} d_{ij}^2 p_{j|i}$. Since $H_i = \log(\eta)$ for all$i = 1, \ldots, N$:

$$\eta = \underbrace{\sum_{j \in \mathcal{N}_i} \exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2)}_{A} + \underbrace{\sum_{j \notin \mathcal{N}_i} \exp(-(d_{ij}^2 - \bar{d}_i^2)/2\sigma_i^2)}_{B}$$

Then, we have:



(a) Spiral in 3D colored by the arclength

(b) Spiral in 3D colored membership to density groups

Figure 4: Comparison of standard and scaled t-SNE. Noisy spiral with the second half sample four times denser than the first half. Standard (middle) and scaled (right) t-SNE (perplexity, $\eta = 100$).

$$
\begin{aligned}
|\mathcal{N}_i| \exp((\bar{d}_i^2 - d^2_{i,\max})/2\sigma_i^2) \ &\leq A \ \leq \ |\mathcal{N}_i| \exp((\bar{d}_i^2 - d^2_{i,\min})/2\sigma_i^2) \\
0 \ &\leq B \ \leq \ \epsilon((N - |\mathcal{N}_i|)/N) \exp(\bar{d}_i^2/2\sigma_i^2)
\end{aligned}
$$

where $d^2_{i,\min} = \min_{j \in \mathcal{N}_i} d^2_{ij}$ and $d^2_{i,\max} = \max_{j \in \mathcal{N}_i} d^2_{ij}$, the minimum and maximum distance of i to its effective neighbors. Note that $(\bar{d}_i^2 - d^2_{i,\min})/2\sigma_i^2 \ll 1$, since $\bar{d}_i^2$ are highly skewed towards small squared distances d^2_{ij}s. Therefore, the term $\exp((\bar{d}_i^2 - d^2_{i,\min})/2\sigma_i^2) \approx 1$. Let $\Delta_i = d^2_{i,\max} - d^2_{i,\min}$ be the spread of distances in the neighborhood of i. Then we have:

$$
|\mathcal{N}_i| \exp(-\Delta_i^2/2\sigma_i^2) \ \leq \eta \ \leq \ |\mathcal{N}_i| \exp(\Delta_i^2/2\sigma_i^2) + \mathcal{O}(\epsilon)
$$

when distances to nearby neighbors are similar, $\Delta_i^2 \ll 1$ is small, we have $\eta \approx |\mathcal{N}_i|$.

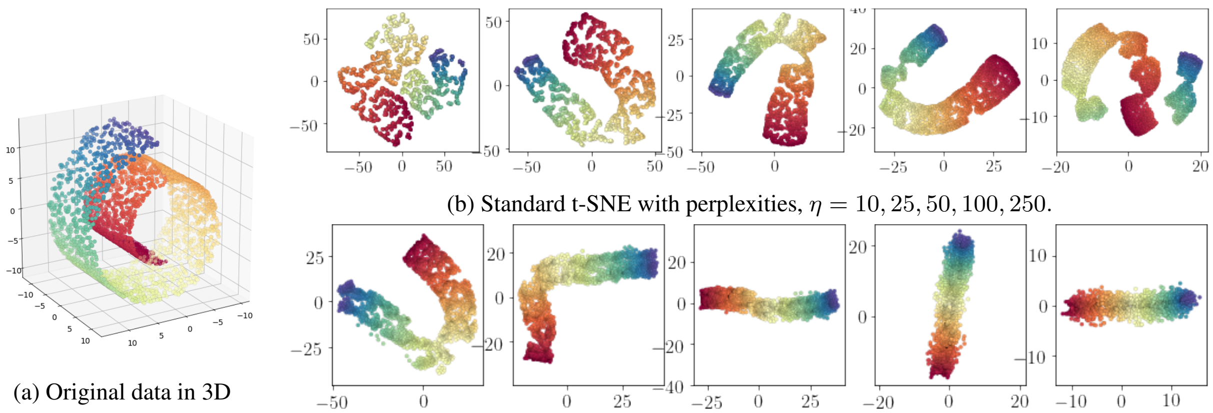
4.3 More examples with scaled and diffusion t-SNE

Spiral in 3D. Below, we show an example where the standard t-SNE also generates misleading representations of datasets governed by continuous gradients, when data sampling is not uniform across regions. In particular, here we simulate data points from a spiral in 3D and add a noise term. Then, we subsample the first half of the spiral, taking every fourth observation. The generated non-uniformly sampled spiral is shown in Fig. 4 (a, d).

Varying density across neighborhoods produces data distortions in the t-SNE output embeddings, just like in previous examples with Gaussian clusters. Despite a large value of the chosen perplexity parameter, t-SNE plots in Fig. 4(b, e) represent the original spiral as two disconnected curves. In addition, the sparser half of the spiral is highly contracted, making it appear much shorter than the second half. The density-dependent distortions can be highly misleading when one is interested in capturing the dynamics or the magnitudes of fluctuations occurring over the course of a continuous process. For example, for a researcher studying cell differentiation process, effective data visualizations should accurately portray changes in the gene expression variability along the developmental trajectory. The t-SNE embeddings stretch and shrink the original data depending on the regional sampling density. Since the data collection procedures are usually independent of the local shifts in variability and do not cover the unknown manifold where the observations reside uniformly, t-SNE maps cannot faithfully depict the original data and the underlying processes.

Introducing **the scaling factor**, our algorithm is able to recover the differences in local variances across neighborhoods. As show in Fig. 4 (c, f), scaled t-SNE returns a spiral represented as a single connected curve in 2D. The embedding does not distort the data and preserves the relative sizes. Even though, the "blue part" of the spiral contains four times fewer observations, in the scaled t-SNE embedding the two halves appear (correctly) to be of the equal lengths. Note that the algorithm also accurately "spreads out" the data points in the blue region allowing the user to detect a lower sampling density in that neighborhood.

Swiss Roll. Here add the results of the scaled diffusion t-SNE applied to Swiss Roll dataset.



(a) Original data in 3D

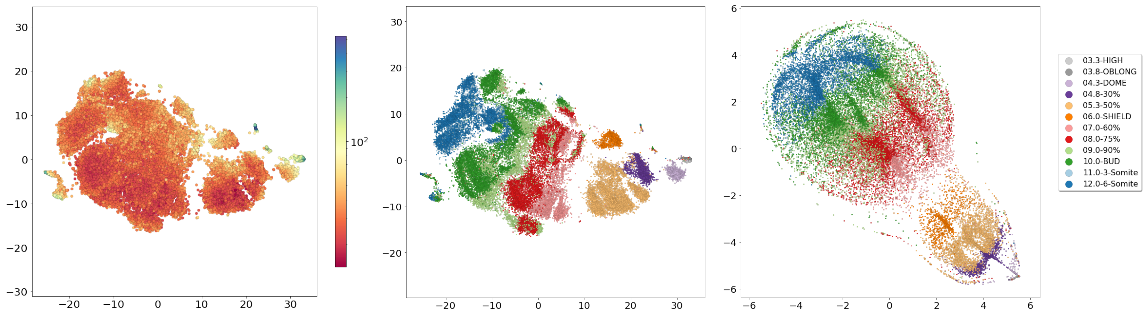
(b) Standard t-SNE with perplexities, $\eta = 10, 25, 50, 100, 250$.

(c) Scaled Diffusion t-SNE with perplexity, $\eta = 25$, and time step, $t = 5, 10, 20, 50, 100$.

Figure 5: Swiss roll (a) embedding. Comparing effect of varying perplexity value in standard t-SNE (b) and varying time step parameter in Scaled Diffusion t-SNE (c). In addition to recovering the global structure like Diffusion t-SNE, teh Scaled Diffusion t-SNE also respects the varying density across the manifold.

Zebrafish embryogenesis The effectiveness of diffusion t-SNE can be illustrated on another scRNA-seq dataset generated by Farrell and co-authors in a study describing the developmental trajectories during zebrafish embryogenesis [3].

We show the standard t-SNE embeddings in Fig. 6 colored by estimated bandwidths in (a) and cell stage (b). Diffusion t-SNE ($t = 10$), (c) is better at recovering the embryonic developmental process showing a more consistent progression through different stages unlike the standard t-SNE which outputs disconnected patches.



(a) Standard t-SNE, bandwidths

(b) Standard t-SNE, cell stage

(c) Diffusion t-SNE

Figure 6: Comparison of standard and diffusion t-SNE. Single cell RNA-seq dataset collected by Farrell et al. [3] to study the zebrafish embryogenesis. Standard t-SNE embedding colored by estimated bandwidths (a) and cell stage (b). Diffusion t-SNE ($t = 10$) (c).