t-SNE Exaggerates Clusters, Provably

Noah Bergam, Szymon Snoeck, Nakul Verma (Columbia University)

Overview

Motivation: t-SNE (t-distributed stochastic neighbor embedding) is one of the most widely-used data visualization techniques.

Previous (theory) analysis showed t-SNE produces "true positives":

well-clustered input implies well-clustered output.

We study t-SNE's failure modes, and prove that in general:

well-clustered output does NOT imply well-clustered input.

How t-SNE Works

Given $X = \{x_1, \dots, x_n\}$ and perplexity parameter $\rho \in [1, n-1]$,

$$\text{t-SNE}_\rho(X) = \text{argmin}_{Y=\{y_1,\dots,y_n\}\subset\mathbb{R}^2} \text{KL}(P(X,\rho) \,||\, Q(Y))$$

P and Q measure neighborhood structure in the input and output, resp.

$P(X,\rho)_{ij} \propto P_{i|j} + P_{j|i}$ where $P_{i|j} = \frac{e^{-||x_i - x_j||^2/\sigma_j^2}}{\sum_{l\neq j} e^{-||x_l - x_j||^2/\sigma_j^2}}$ σ_j^2 selected such that $\text{entropy}(P_{\cdot|j}) = \rho$

$Q(Y)_{ij} \propto \left(1 + ||y_i - y_j||^2\right)^{-1}$ (P ≈ nearest neighbors, Q ≈ radius neighbors)

t-SNE Exaggerates Clusters

Theorem ["False positive" clustering]*

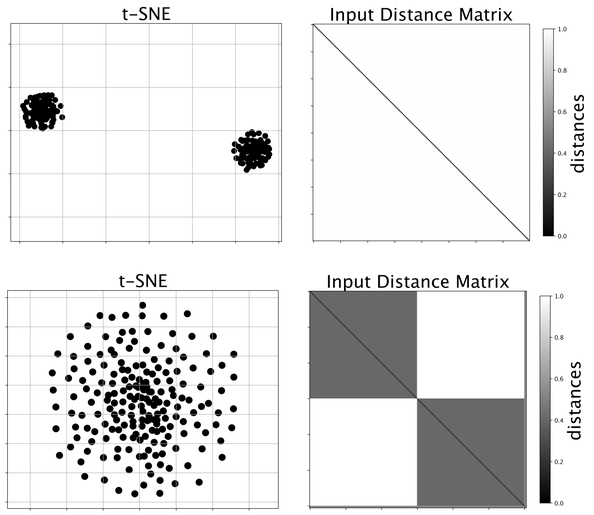
Any t-SNE output can be generated by an "impostor" dataset where all distances are arbitrarily close to uniform.

Theorem ["False negative" clustering]*

There exist "high-dimensional" X and "poison point" x_0 such that: $\text{t-SNE}_\rho(X)$ is clustered but $\text{t-SNE}_\rho(X \cup x_0)$ is unclustered.

Theorem [Instability]*

For "high-dimensional" X, perturbations can arbitrarily change output.



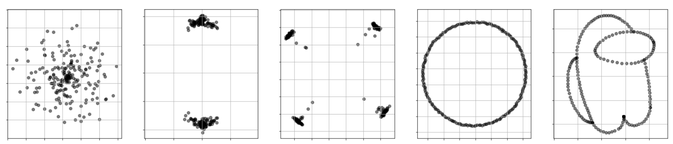


Figure: Various 2D t-SNE visualizations produced by adversarial perturbations of a 200-point unit regular simplex

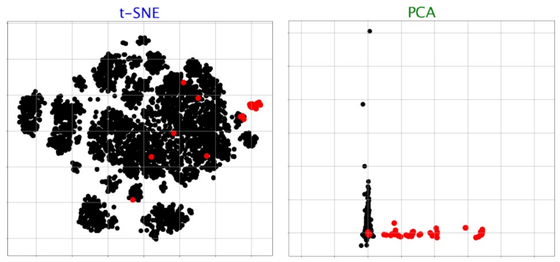
Common Proof Idea: taking advantage of t-SNE's "additive invariance" property.

- **Fact**: If D_0, D_1 are squared Euclidean distance matrices (EDMs), then $D_0 + D_1$ is also EDM. (EDMs form a cone).
- Let $D(X)$ denote the EDM of a dataset X. Let X(D) denote the point cloud realization of EDM D.
- Let $D_{\text{simp}} = 1_{n\times n} - I_n$ be the squared distance matrix of the regular simplex.
- **Fact**: $\text{t-SNE}(X) = \text{t-SNE}(X(D(X) + cD_{\text{simp}}))$ for all datasets X and $c > 0$.

*Holds empirically for UMAP

t-SNE Suppresses Outliers

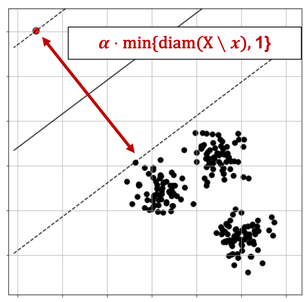
Another form of "false positive" clustering: t-SNE absorbs outliers into cluster structure!



e.g. t-SNE versus PCA on credit card activity dataset (fraudulent users = red)

We formalize this!

For a dataset X, let $\alpha(X)$ denote outlier strength (visual definition below)



Theorem [outlier suppression]*

For all X and ρ,

$$\alpha\left(\text{t-SNE}_\rho(X)\right) \leq 3.622.$$

In practice, $\alpha\left(\text{TSNE}_\rho(X)\right) \approx 0$ even as $\alpha(X) \to \infty$.

Experiment: Let X be a 500 points from a 50-dimensional Gaussian-distributed dataset with a single extreme outlier.

Vary $\alpha(X)$ by changing the position in the outlier. Observe $\alpha(\text{PCA}(X)) \approx \alpha(X)$ while $\alpha(\text{t-SNE}(X)) \approx 0$