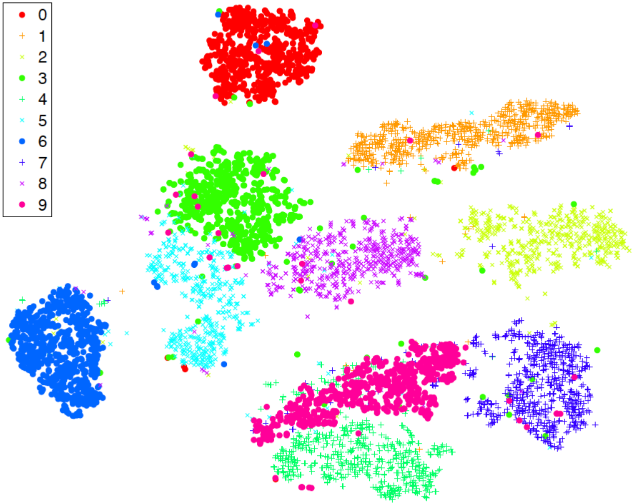
Data Visualization with t-SNE in Theory and Practice

Solveig Tränkner, Supervisor: Prof. Dr. Jochen Garcke

19th December 2024



Outline

Dimensionality Reduction

Goal: Map high-dimensional data to a lower dimension $\mathbb{R}^d \to \mathbb{R}^s$, $s \ll d$ while perserving intrinsic structure.

Dimensionality Reduction

Goal: Map high-dimensional data to a lower dimension $\mathbb{R}^d \to \mathbb{R}^s$, $s \ll d$ while perserving intrinsic structure.

Applications

- Compression
- Feature extraction
- Data visualization

Dimensionality Reduction

Goal: Map high-dimensional data to a lower dimension $\mathbb{R}^d \to \mathbb{R}^s$, $s \ll d$ while perserving intrinsic structure.

Applications

- Compression
- Feature extraction
- Data visualization

Methods

- Linear methods: Principal Component Analysis (PCA), Multidimensional Scaling (MDS)
- Nonlinear methods: Isomap, t-Stochastic Neighbor Embedding (t-SNE)

Challenges of High Dimensions

Curse of Dimensionality

- Volume of a hypercube (with side length 2) grows in $\mathcal{O}(2^d)$
- Data points become sparse

Challenges of High Dimensions

Curse of Dimensionality

- Volume of a hypercube (with side length 2) grows in $\mathcal{O}(2^d)$
- Data points become sparse

Distance Concentration

- In high-dimensional spaces $\frac{\|x_i - x_j\|}{\|x_i - x_k\|} \approx 1$ for most x_i, x_j, x_k.
- Euclidean distance becomes less meaningful

Challenges of High Dimensions

Curse of Dimensionality

- Volume of a hypercube (with side length 2) grows in $\mathcal{O}(2^d)$
- Data points become sparse

Distance Concentration

- In high-dimensional spaces $\frac{||x_i - x_j||}{||x_i - x_k||} \approx 1$ for most x_i, x_j, x_k.
- Euclidean distance becomes less meaningful

Crowding Problem

- High-dimensional points cannot be faithfully embedded in two or three dimensions
- Intrinsically distant points may cluster artificially due to limited space in the embedding

The Main Idea Behind t-SNE

Problem

Given a set of high-dimensional points $\mathcal{X} = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$ *find a "good" lower-dimensional representation* $\mathcal{Y} = \{y_1, y_2, ..., y_n\} \subset \mathbb{R}^s$ *of these points, where* $s = 2, 3$.

The Main Idea Behind t-SNE

Problem

Given a set of high-dimensional points $\mathcal{X} = \{x_1, x_2, ..., x_n\} \subset \mathbb{R}^d$ *find a "good" lower-dimensional representation* $\mathcal{Y} = \{y_1, y_2, ..., y_n\} \subset \mathbb{R}^s$ *of these points, where* $s = 2, 3$.

An Informal Overview of t-SNE [MH08, Van der Maaten, Hinton]

- Create an initial set of points $\mathcal{Y}$ in $\mathbb{R}^s$
- Turn $\mathcal{X}$ and $\mathcal{Y}$ into probability distributions reflecting pairwise similarities between datapoints
- Force these distributions to be as similar as possible by moving points in the lower dimension around

Measuring Similarity of Data Points

How can we measure similarity between points in a high dimension?

Measuring Similarity of Data Points

How can we measure similarity between points in a high dimension?

- Compute a joint probability distribution over points x_i and x_j $(i \neq j)$ via

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}, \ p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

where σ_i denotes the bandwidth of the Gaussian kernel centered at x_i.

Measuring Similarity of Data Points

How can we measure similarity between points in a high dimension?

- Compute a joint probability distribution over points x_i and x_j $(i \neq j)$ via

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}, \; p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

  where σ_i denotes the bandwidth of the Gaussian kernel centered at x_i.

Should we measure similarity in $\mathbb{R}^2$ in the same way?

Measuring Similarity of Data Points

How can we measure similarity between points in a high dimension?

- Compute a joint probability distribution over points x_i and x_j $(i \neq j)$ via

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}, \; p_{ij} = \frac{p_{i|j} + p_{j|i}}{2n}$$

where σ_i denotes the bandwidth of the Gaussian kernel centered at x_i.

Should we measure similarity in $\mathbb{R}^2$ in the same way?

- To avoid overcrowding, we define a similarity measure between points in the low dimensional embedding y_i and y_j $(i \neq j)$ via

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_l \sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}}$$

using a Student's t-distribution with one degree of freedom (Cauchy distribution) which is heavy-tailed compared to a Gaussian.

Defining a Loss Function

Definition (Kullback-Leibler divergence)

Let P and Q be discrete probability distributions defined on a sample space $\mathcal{X}$. The Kullback-Leibler divergence between P and Q is defined as

$$\mathrm{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

Defining a Loss Function

Definition (Kullback-Leibler divergence)

Let P and Q be discrete probability distributions defined on a sample space $\mathcal{X}$. The Kullback-Leibler divergence between P and Q is defined as

$$\mathrm{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

- We want to find points $\{y_1, ..., y_n\}$ which minimize the Kullback-Leibler divergence between the distributions P and Q defined above.

Defining a Loss Function

Definition (Kullback-Leibler divergence)

Let P and Q be discrete probability distributions defined on a sample space $\mathcal{X}$. The Kullback-Leibler divergence between P and Q is defined as

$$\mathrm{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

- We want to find points $\{y_1, ..., y_n\}$ which minimize the Kullback-Leibler divergence between the distributions P and Q defined above.
- The loss function is given by

$$C(\mathcal{Y}) = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Defining a Loss Function

Definition (Kullback-Leibler divergence)

Let P and Q be discrete probability distributions defined on a sample space $\mathcal{X}$. The Kullback-Leibler divergence between P and Q is defined as

$$\mathrm{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)}.$$

- We want to find points $\{y_1, ..., y_n\}$ which minimize the Kullback-Leibler divergence between the distributions P and Q defined above.
- The loss function is given by

$$C(\mathcal{Y}) = KL(P||Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- KL divergence is asymmetric, i.e. in general $\mathrm{KL}(P||Q) \neq \mathrm{KL}(Q||P)$.

Optimization via Gradient Descent

- We can minimize $C(\mathcal{Y})$ using Gradient Descent, where

$$\frac{\partial C}{\partial y_i} = 4\sum_{j \neq i}(p_{ij} - q_{ij})q_{ij}Z(y_i - y_j) \text{ with } Z = \sum_{l}\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}.$$

Optimization via Gradient Descent

- We can minimize $C(\mathcal{Y})$ using Gradient Descent, where

$$\frac{\partial C}{\partial y_i} = 4\sum_{j\neq i}(p_{ij} - q_{ij})q_{ij}Z(y_i - y_j) \text{ with } Z = \sum_{l}\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}.$$

- Standard optimization techniques for gradient descent can be employed, such as including a momentum term

$$y_i^{(t+1)} = y_i^{(t)} - h \cdot \frac{\partial C}{\partial y_i^{(t)}} + \beta(y_i^{(t)} - y_i^{(t-1)}) \text{ with } 0 \leq \beta < 1, h > 0.$$

Optimization via Gradient Descent

- We can minimize $C(\mathcal{Y})$ using Gradient Descent, where

$$\frac{\partial C}{\partial y_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij}) q_{ij} Z (y_i - y_j) \text{ with } Z = \sum_{l} \sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}.$$

- Standard optimization techniques for gradient descent can be employed, such as including a momentum term

$$y_i^{(t+1)} = y_i^{(t)} - h \cdot \frac{\partial C}{\partial y_i^{(t)}} + \beta (y_i^{(t)} - y_i^{(t-1)}) \text{ with } 0 \leq \beta < 1, h > 0.$$

- Rewriting the gradient yields interpretation via attractive and repulsive forces:

$$-\frac{1}{4} \frac{\partial C}{\partial y_i} = \underbrace{\sum_{j \neq i} p_{ij} q_{ij} Z (y_j - y_i)}_{\text{attractive force}} - \underbrace{\sum_{j \neq i} q_{ij}^2 Z (y_j - y_i)}_{\text{repulsive force}}$$

Initialization

- It is best to use "informative initialization" instead of random initialization for t-SNE.
- PCA initialization better preserves global structure.

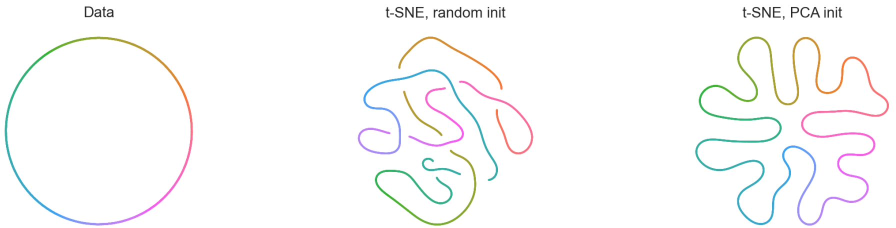


Figure: The t-SNE algorithm only produces a faithful representation of the circle with informative initialization. Visualization reproduced from [KL21]

.

Perplexity

- Recall

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

How do we choose the bandwidth σ_i of the Gaussian kernel centered at x_i?

Perplexity

- Recall

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

  How do we choose the bandwidth σ_i of the Gaussian kernel centered at x_i?
- It depends on the data! In dense regions, we want a smaller σ_i than in sparse ones.

Perplexity

- Recall

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

How do we choose the bandwidth σ_i of the Gaussian kernel centered at x_i?

- It depends on the data! In dense regions, we want a smaller σ_i than in sparse ones.

Definition (Shannon Entropy)

Let X be a discrete random variable taking values in $\mathcal{X}$ which is distributed according to $p : \mathcal{X} \to [0,1]$, then the Shannon Entropy is $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log_2(p(x))$.

- We define a constant **perplexity** value

$$\text{Perp}(P_i) = 2^{H(P_i)} = 2^{-\sum_j p_{j|i}\log p_{j|i}}$$

where P_i is the probability distribution induced by σ_i.

Perplexity

- Recall

$$p_{i|j} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

  How do we choose the bandwidth σ_i of the Gaussian kernel centered at x_i?
- It depends on the data! In dense regions, we want a smaller σ_i than in sparse ones.

Definition (Shannon Entropy)

Let X be a discrete random variable taking values in $\mathcal{X}$ which is distributed according to $p : \mathcal{X} \to [0,1]$, then the Shannon Entropy is $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2(p(x))$.

- We define a constant **perplexity** value

$$\text{Perp}(P_i) = 2^{H(P_i)} = 2^{-\sum_j p_{j|i} \log p_{j|i}}$$

  where P_i is the probability distribution induced by σ_i.
- Perplexity effectively measures the number of neighbours we wish to consider.

Perplexity

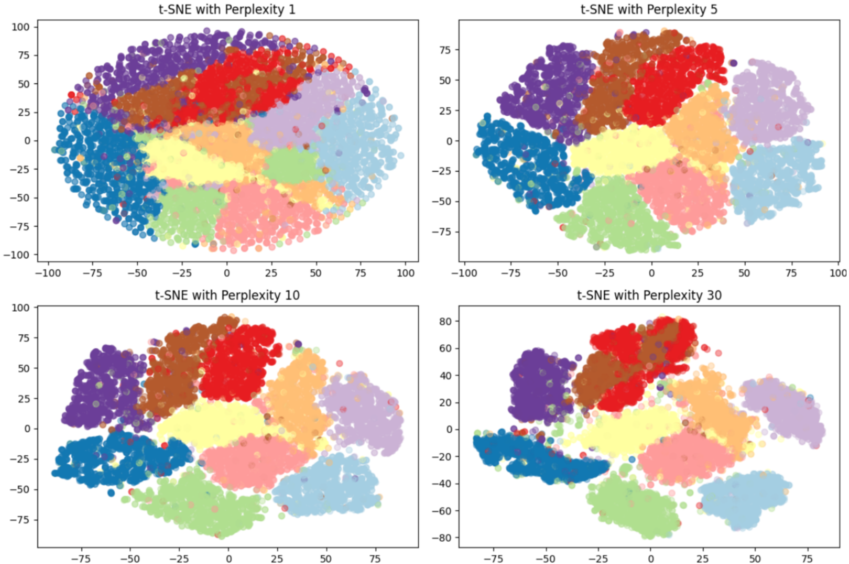


Figure: Different perplexity values on the MNIST dataset.

Perplexity

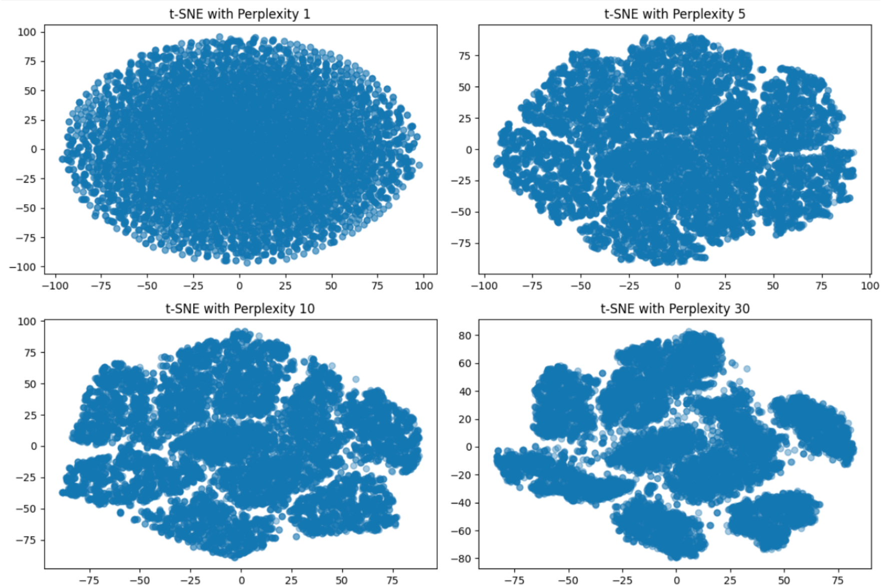


Figure: Different perplexity values on the MNIST dataset without ground truth labels.

Early Exaggeration

- **Idea**: in early iterations of t-SNE we want to focus on tight cluster formation and only worry about "nice" visualization later on

Early Exaggeration

- **Idea**: in early iterations of t-SNE we want to focus on tight cluster formation and only worry about "nice" visualization later on
- We multiply all p_{ij} by a certain factor α (standard value is $\alpha = 12$) for the first few iterations (standard: 250).
- Clusters can more easily move around in space later on.

Early Exaggeration

- **Idea**: in early iterations of t-SNE we want to focus on tight cluster formation and only worry about "nice" visualization later on
- We multiply all p_{ij} by a certain factor α (standard value is $\alpha = 12$) for the first few iterations (standard: 250).
- Clusters can more easily move around in space later on.
- In the attraction-repulsion framework:

$$\frac{1}{4}\frac{\partial C}{\partial y_i} = \sum_{i \neq j} \alpha p_{ij} q_{ij} Z(y_i - y_j) - \sum_{j \neq i} q_{ij}^2 Z(y_i - y_j)$$

Early Exaggeration

- **Idea**: in early iterations of t-SNE we want to focus on tight cluster formation and only worry about "nice" visualization later on
- We multiply all p_{ij} by a certain factor α (standard value is $\alpha = 12$) for the first few iterations (standard: 250).
- Clusters can more easily move around in space later on.
- In the attraction-repulsion framework:

$$\frac{1}{4}\frac{\partial C}{\partial y_i} = \sum_{i \neq j} \alpha p_{ij} q_{ij} Z(y_i - y_j) - \sum_{j \neq i} q_{ij}^2 Z(y_i - y_j)$$

- **Question**: How do we find good values for α and for how many iterations should we keep early exaggeration on?

Automated Optimized Parameters

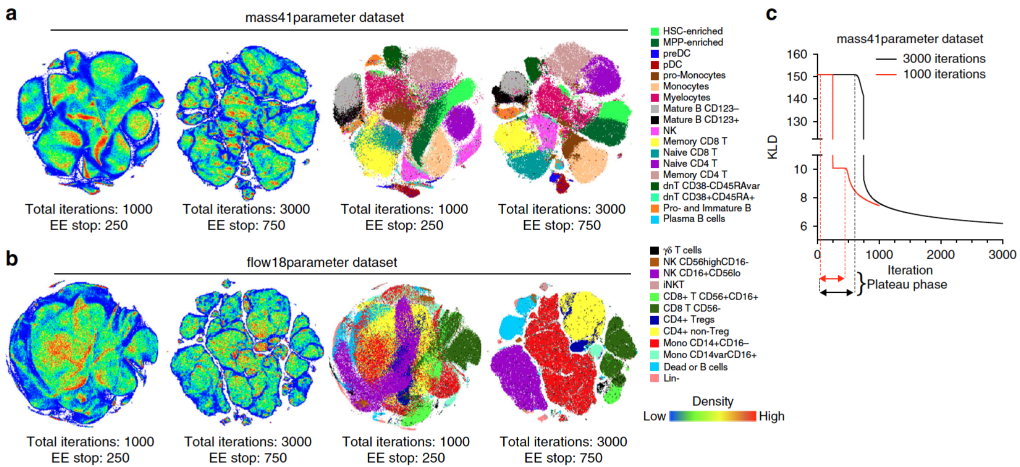


Figure: Performance of t-SNE for cytometry data visualization, see [Bel+19].

Theoretical Work on t-SNE

Theorem (Cluster Formation [LS22])

Let $\mathcal{X}$ be "clustered" and initialize $\mathcal{Y} \subset [-0.01, 0.01]^2$. Choose α and h such that for some $1 \leq i \leq n$

$$0.01 \leq \alpha h \sum_{\substack{j \neq i \\ \text{same cluster}}} p_{ij} \leq 0.9$$

Then, the diameter of the embedded cluster $\{y_i : 1 \leq j \leq n \wedge \pi(j) = \pi(i)\}$ decays exponentially until its diameter satisfies, for some universal $c > 0$,

$$\mathit{diam}\{y_j : 1 \leq j \leq n \wedge \pi(j) = \pi(i)\} \leq ch \left(\alpha \sum_{\substack{j \neq i \\ \text{same cluster}}} p_{ij} + \frac{1}{n} \right).$$

Theoretical Work on t-SNE

Theorem (Cluster Formation [LS22])

Let $\mathcal{X}$ be "clustered" and initialize $\mathcal{Y} \subset [-0.01, 0.01]^2$. Choose α and h such that for some $1 \leq i \leq n$

$$0.01 \leq \alpha h \sum_{\substack{j \neq i \\ \text{same cluster}}} p_{ij} \leq 0.9$$

Then, the diameter of the embedded cluster $\{y_i : 1 \leq j \leq n \wedge \pi(j) = \pi(i)\}$ decays exponentially until its diameter satisfies, for some universal $c > 0$,

$$diam\{y_j : 1 \leq j \leq n \wedge \pi(j) = \pi(i)\} \leq ch \left(\alpha \sum_{\substack{j \neq i \\ \text{same cluster}}} p_{ij} + \frac{1}{n} \right).$$

- Other theoretical results study connections to spectral clustering ([CM22]) and limiting behaviour for $n \to \infty$ ([MP24]).

Weaknesses of t-SNE

- No convergence to global optimum guaranteed due to non-convex loss function

Weaknesses of t-SNE

- No convergence to global optimum guaranteed due to non-convex loss function
- Results heavily depend on choice of parameters

Weaknesses of t-SNE

- No convergence to global optimum guaranteed due to non-convex loss function
- Results heavily depend on choice of parameters
- Visualizations can be misleading

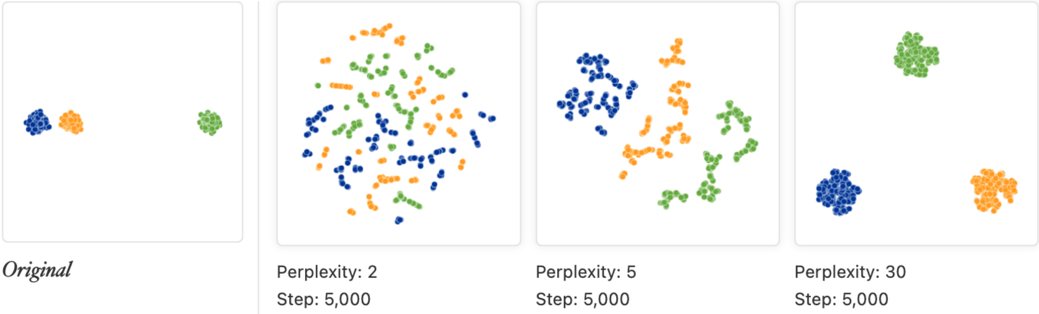


Figure: t-SNE does not preserve distance between clusters, see [WVJ16]

Weaknesses of t-SNE

- No guaranteed convergence to global optimum due to non-convex loss function
- Results heavily depend on choice of parameters
- Visualizations can be misleading

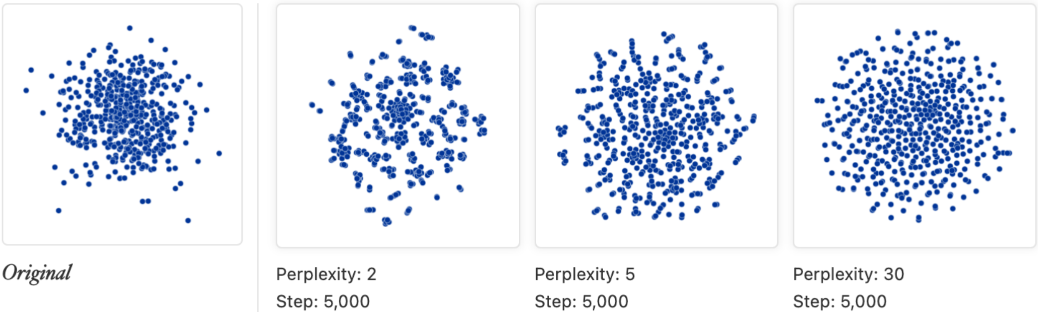


Figure: Random Gaussian noise does not always look random, see [WVJ16]

.

Questions for Future Work

- How realiable are theoretical results on t-SNE? Do the assumptions hold for real-world datasets?

Questions for Future Work

- How realiable are theoretical results on t-SNE? Do the assumptions hold for real-world datasets?
- Do parameters suggested in theoretical results lead to better performance in practice?

Questions for Future Work

- How realiable are theoretical results on t-SNE? Do the assumptions hold for real-world datasets?
- Do parameters suggested in theoretical results lead to better performance in practice?
- Which parameters are most important for obtaining good results?

Questions for Future Work

- How realiable are theoretical results on t-SNE? Do the assumptions hold for real-world datasets?
- Do parameters suggested in theoretical results lead to better performance in practice?
- Which parameters are most important for obtaining good results?
- How does initialization impact t-SNE results across different datasets?

Questions for Future Work

- How realiable are theoretical results on t-SNE? Do the assumptions hold for real-world datasets?
- Do parameters suggested in theoretical results lead to better performance in practice?
- Which parameters are most important for obtaining good results?
- How does initialization impact t-SNE results across different datasets?
- How can we deal with large datasets? Does rescaled t-SNE work in practice?

References I

[Bel+19] Anna C. Belkina et al. "Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets". In: *Nature Communications* 10.1 (2019).

[CM22] T. Tony Cai and Rong Ma. "Theoretical Foundations of t-SNE for Visualizing High-Dimensional Clustered Data". In: *Journal of Machine Learning Research* 23.301 (2022), pp. 1–54. URL: http://jmlr.org/papers/v23/21-0524.html.

[KL21] Dmitry Kobak and George C. Linderman. "Initialization is critical for preserving global data structure in both t-SNE and UMAP". In: *Nature Biotechnology* 39.2 (2021). DOI: 10.1038/s41587-020-00809-z.

[LS22] George C. Linderman and Stefan Steinerberger. "Dimensionality Reduction via Dynamical Systems: The Case of t-SNE". In: *SIAM Review* 64.1 (2022), pp. 153–178. DOI: 10.1137/21M1446769.

References II

[MH08] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

[MP24] Ryan Murray and Adam Pickarski. *Large data limits and scaling laws for tSNE*. 2024. arXiv: 2410.13063 [math.ST]. URL: https://arxiv.org/abs/2410.13063.

[WVJ16] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. “How to Use t-SNE Effectively”. In: *Distill* (2016). DOI: 10.23915/distill.00002. URL: http://distill.pub/2016/misread-tsne.