Visualization

CH11: t-SNE

Jong-June Jeon

University of Seoul, Department of Statistics

Contents

- Introduction
- Stochastic Neighbor Embedding
- t-SNE
- Conclusion

Introduction

Multidimensional Scaling (MDS)

- Goal: Find low-dimensional embedding that preserves pairwise distances
- Rougly spealing, classical MDS minimizes:

$$\sum_{i<j} \left(\|x_i - x_j\|^2 - \|y_i - y_j\|^2 \right)^2$$

- Distance-based embedding, assumes Euclidean geometry

Limitations of MDS: Overview

- Global–local trade-off: stress weights all pairs equally $\rightarrow$ local neighborhoods distort
- Crowding problem: high-dimensional volume vs. 2-dimensional area
- Optimization complexity: stress minimization is NP-hard
- Statistical instability in the presence of noise
- Lower-bound theory: JL lemma $\rightarrow$ unavoidable distortion when $k = 2$

Global-Stress Objective Distorts Locals

- Raw-stress (Kruskal, 1964):

$$\mathcal{S}(Y) = \sum_{i<j} (\|x_i - x_j\|^2 - \|y_i - y_j\|^2)^2$$

- All pairs equally weighted $\Rightarrow$ a few large distances dominate optimisation.
- Stress <0.05 "excellent", >0.20 "poor" (rule of thumb).

Crowding Problem: Volume Argument I

- Volume of ball in $\mathbb{R}^d$ with radius r: $V_d(r) = C_d r^d$.
- Space per point (volume argument):

$$\mathrm{S}^{(d)}(r) = \frac{V_d(r)}{n}$$

- Find the typical spacing s_d by equating a d-ball volume to $\mathrm{S}^{(d)}(r)$:

$$C_d s_d^d = \frac{C_d r^d}{n} \implies s_d(r) = rn^{-1/d}.$$

Also, $s_2(r') = r' n^{-1/2}$

Crowding Problem: Volume Argument II

- For convenience let $r = 1$ denote $s_d(1)$ by s_d. If we maintain the distance between adjacent points on $\mathbb{R}^2$ ($s_2(r') = r' n^{-1/2} = n^{-1/d} = s_d$) then

$$r' = n^{\frac{d-2}{2d}}$$

  That is, to main the distance equally, we have to scatter n points on the circle with $V_2(r') = C_2 n^{\frac{d-2}{d}}$

- Consider the scaling for the unit circle then the shrink factor rendering the volumn on 2-D to be 1 is given by $\frac{1}{\sqrt{C_2}} n^{-\frac{d-2}{2d}}$

Crowding Problem: Volume Argument III

- Before the shrinkage, $s_2(r') = s_d$ Since $s_2(r') \propto r'$, we know that $s_2(kr')/s_2(r') = k$.
- Thus, the distance between adjacent points on $\mathbb{R}^2$ with the shrinkage factor $\frac{1}{\sqrt{C_2}} n^{-\frac{d-2}{2d}}$ becomes

$$s_d \frac{1}{\sqrt{C_2}} n^{-\frac{d-2}{2d}}$$

Moreover, most of data on the high dimensional space lie on the boundary of the hyperspheres. Thus, A vast number of points live in the "moderate distance" shell in high-D. In 2-D there is simply not enough area to place them without severe overlap.

Optimisation Complexity of MDS

- Stress minimisation (a.k.a. Kamada–Kawai) shown NP-hard for $k = 2$ (Favoni Huang Lee 2021).
- Gradient descent may converge to poor local minima; guarantees only for special graphs.

Motivation: Using Probabilities for Similarity

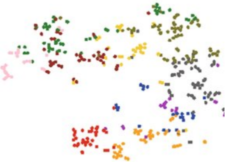
- Instead of preserving pairwise distances directly, preserve pairwise **similarities.**
- Define similarity between x_i and x_j as a conditional probability: "Given x_i, how likely is x_j a neighbor?"
- High similarity $\Rightarrow$ high probability $p_{j|i}$

Low dimensional Embedding I

(a)

(b)

(c)



(d)

Stochastic Neighbor Embedding

SNE: Basic Idea I

Stochastic Neighbor Embedding (SNE)

- Map high-dimensional points $x_i \in \mathbb{R}^D$ to low-dimensional $y_i \in \mathbb{R}^d$ ($d = 2$ or 3).
- Preserve *local* structure by matching neighbourhood probabilities.
- For any distance kernel $k(x) > 0$ that decreases with x (distance) and assume that k is gaussian.
- For each point x_i, define a conditional similarity:

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

SNE: Basic Idea II

In the map space we use a fixed-scale kernel:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

- Goal: make $q_{j|i}$ as close as possible to $p_{j|i}$ for every i (estimation of y_is called the low dimensional embedding).
- Cost function (sum of KL divergences):

$$C_{\mathsf{SNE}} = \sum_i \mathrm{KL}\left(P_i \,\|\, Q_i\right) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}.$$

SNE: Stochastic Neighbor Embedding

- Goal: Embed high-dimensional data into a low-dimensional space while preserving local structure.
- Conditional probability for similarity in high dimensions:

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

- In low-dimensional space:

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|y_i - y_k\|^2\right)}$$

- Cost function: Kullback-Leibler divergence

$$C = \sum_i KL(P_i\|Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

Limitations of SNE

- Asymmetry: $p_{j|i} \neq p_{i|j}$
- Crowding problem: Hard to preserve all pairwise distances in low-dimensional space
- Optimization challenges: Multiple local minima

Recall: the Crowding Problem

- When mapping $\mathbb{R}^D$ ($D \gg d$) to $\mathbb{R}^2$, **moderately distant neighbours** of a point cannot all fit at proper relative distances.
- They are pushed to the same narrow annulus, making clusters overlap or distort into "rings".
- Visual artefacts: blurred cluster borders, false neighbours.
- Moreover, most of data on the high dimensional space lie on the boundary of the hyperspheres. Thus, A vast number of points live in the "moderate distance" shell in high-D.

Crowding Problem in Probabilistic View

- High-dim similarities use a Gaussian kernel $p_{ij} \propto \exp(-d_{ij}^2/2\sigma^2)$.
- Mid-range distances $\Rightarrow p_{ij} \approx 0$ but *there are many such pairs*.
- If the map also uses a Gaussian, q_{ij} for those pairs ≈ 0 too $\Rightarrow$ KL divergence forces them closer, intensifying crowding.

t-SNE

How t-SNE Resolves Crowding

Student-t kernel

$$q_{ij} \propto (1 + d_{ij}^2)^{-1}$$

Heavy tail $\sim d^{-2}$ keeps moderate neighbours at non-zero probability.

Early exaggeration

$$p_{ij} \leftarrow \alpha p_{ij}, \ \alpha \approx 12$$

Separates clusters early, then relaxes for fine-scale structure.

Perplexity tuning provides an additional trade-off handle.

t-SNE

$x_i \in \mathbb{R}^p$ for $i = 1, \cdots, n$: (high dimensional) data points. Define a relative closeness of x_j from x_i by

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

For symmetry let

$$p_{ij} = \frac{p_{i|j} + p_{j|i}}{2},$$

the relative closeness between x_i and x_j.
Roughly, if $\|x_i - x_j\|/\|x_i - x_k\| \simeq 1/C$ then $p_{i|j}/p_{i|k} \simeq \exp(C)$.

t-SNE

$y_i \in \mathbb{R}^q$ for $i = 1, \cdots, n$: (low dimensional) embedded data points. Define a relative closeness of y_j from y_i by

$$q_{i|j} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_k (1 + \|y_i - y_k\|^2)^{-1}}.$$

Roughly, if $\|y_i - y_j\| / \|y_i - y_k\| \simeq 1/C$ then $q_{i|j}/q_{i|k} \simeq C$.

t-SNE

Objective (loss) – symmetrised KL divergence

$$\mathcal{L} = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

Gradient w.r.t. a Map Point

Because only q_{ij} depends on $\mathbf{y}$,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij}) \frac{(\mathbf{y}_i - \mathbf{y}_j)}{1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2}.$$

- Attractive term when $p_{ij} > q_{ij}$ (pulls neighbours closer).
- Repulsive term when $p_{ij} < q_{ij}$ (pushes points apart).
- Heavy-tailed denominator mitigates the crowding problem.

Convergence Diagnostics

- Monitor $\mathcal{L}$ – should plateau smoothly after exaggeration ends.
- Watch the *KL gap* $\sum_{i \neq j} |p_{ij} - q_{ij}|$ for stagnation.
- Inspect embeddings every 100–200 iter.: unresolved crowding $\Rightarrow$ raise perplexity or run longer.

Visualization

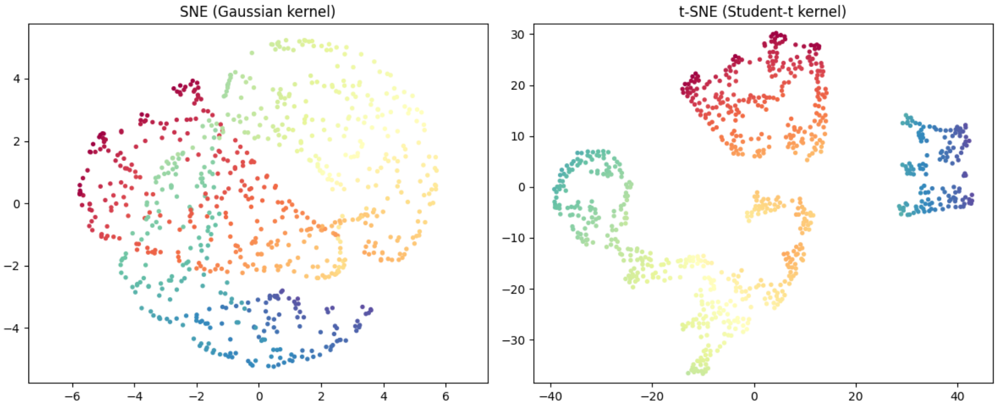


Figure: Left: SNE, Right: t-SNE

Conclusion

Key Takeaways

- **MDS** preserves global pairwise *distances* but suffers from crowding, uniform stress weighting, and NP-hard optimisation.
- **SNE** shifts the focus to *local similarities*, yet its asymmetric Gaussian kernel still crowds mid-range neighbours.
- **t-SNE** resolves crowding via a heavy-tailed Student-t kernel, early exaggeration, and perplexity-controlled neighbourhoods.
- Gradient can be written in closed form, enabling efficient momentum GD; convergence diagnosed by KL loss plateau.