Visualizing Data with t-SNE

Lauren van der Maaten, Geoffrey Hinton

Presented by Nicolas Rondan and Biagio Antonelli

Objectives

The main objectives of t-SNE are the following:

- Powerful data visualization
- Dimensionality reduction
- Capture the local structure and preserve global structure as clusters

What is t-SNE

t-SNE is a variation of Stochastic Neighbor Embedding (SNE). It transforms high-dimensional dataset in 2 or 3 dimensional visualization maps, based on pairwise similarity between points. It is a very powerful data visualization technique.

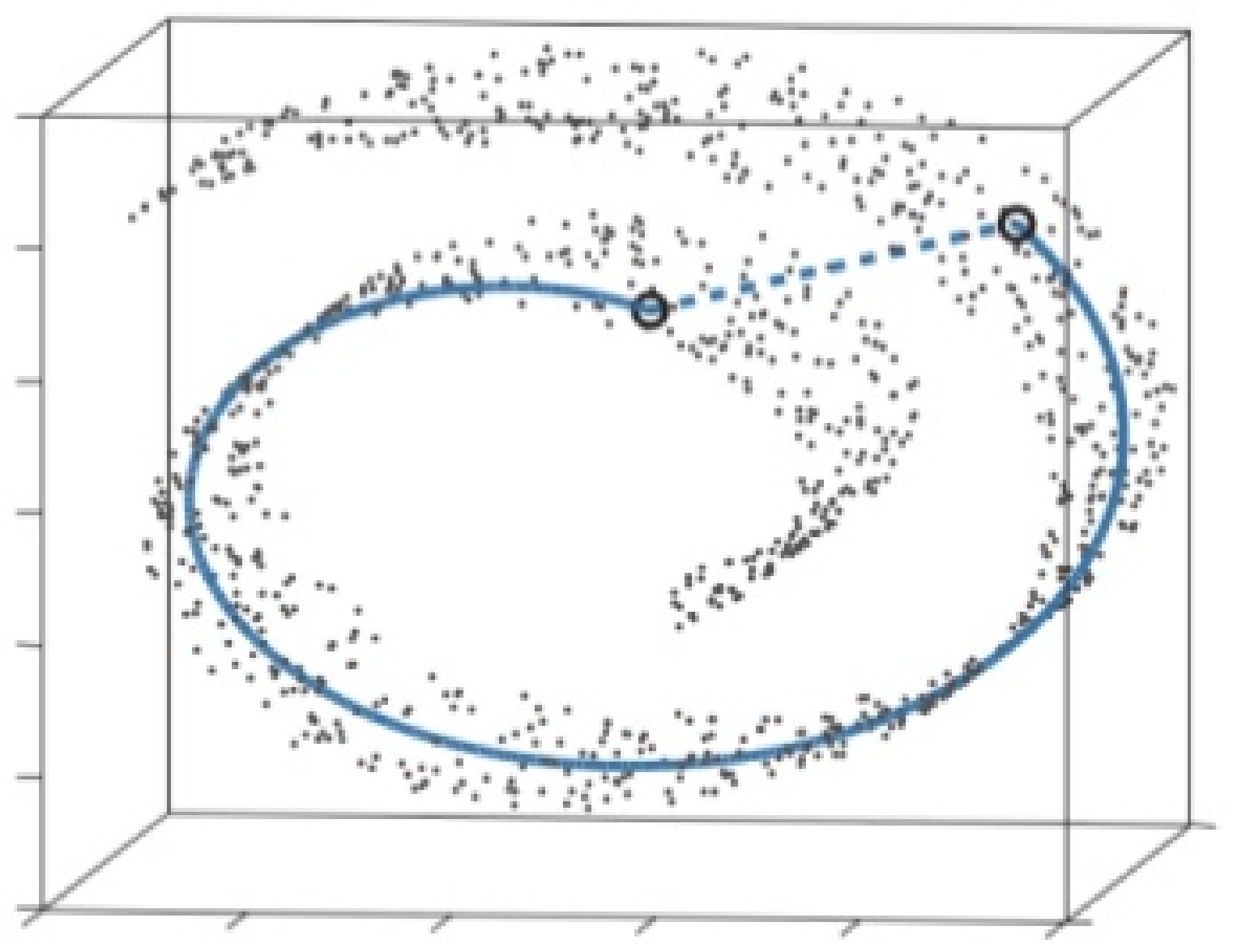
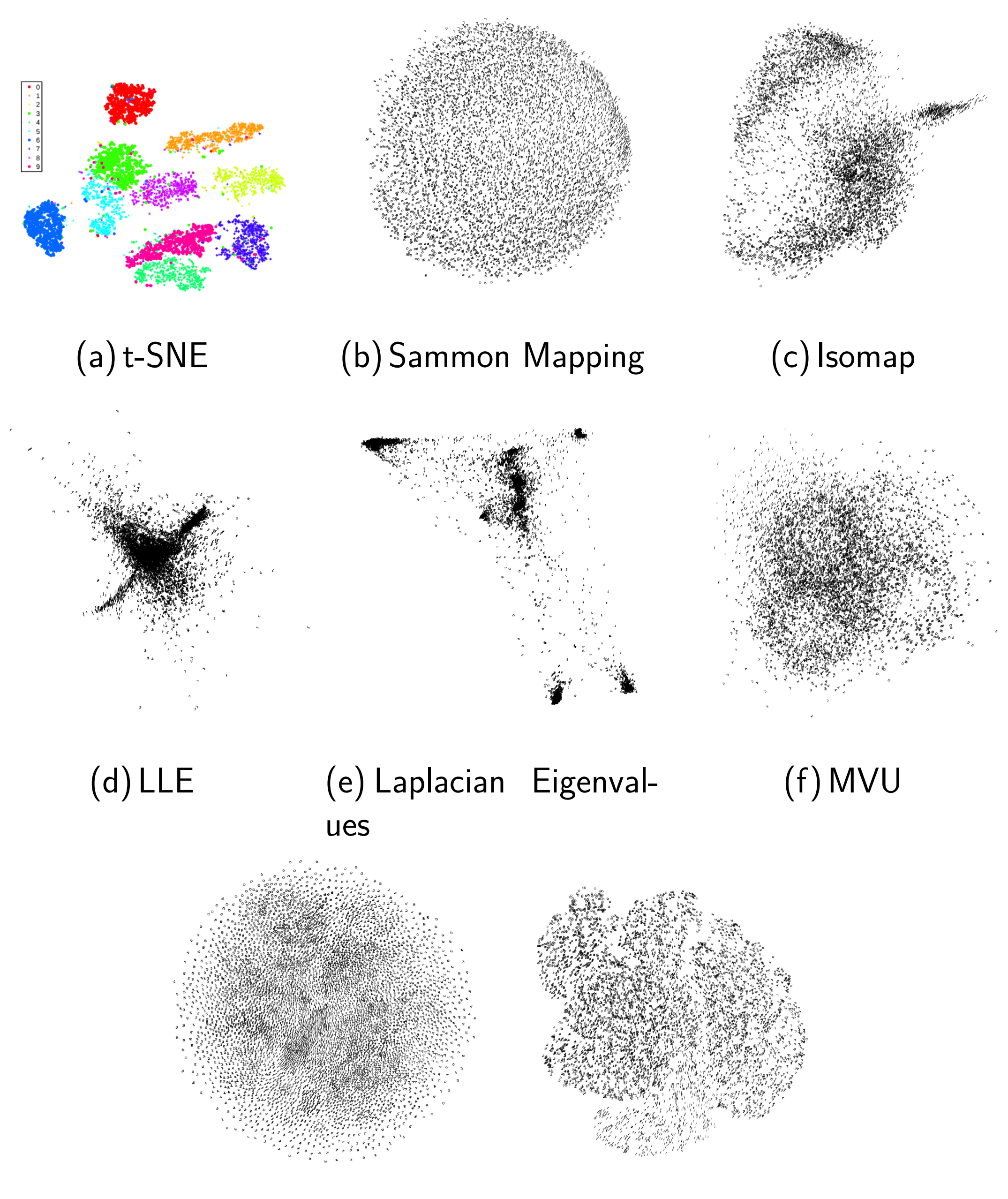


Figure 1: Swiss roll dataset

In contrast with other dimension reduction techniques, t-SNE is very good at revealing the structure of multidimensional data at different scales. The problem with linear techniques (e.g. PCA, MDS) is that they fail to represent the structure of the data when they lie on a non-linear high-dimensional manifold. The swiss roll dataset in Fig. 1 in an example where linear techniques fail in the task.

Compared to other non-linear techniques the visualizations produced by t-SNE are better on almost all the datasets and the algorithm is easy to optimize.

Dimension reduction techniques



(a) t-SNE (b) Sammon Mapping (c) Isomap

(d) LLE (e) Laplacian Eigenvalues (f) MVU

(g) CCA (h) SNE

Figure 2: Visualizations of 6,000 handwritten digits from the MNIST data set.

SNE

Stochastic Neighbour Embedding (SNE) maps a high dimensional dataset χ into a 2 or 3 dimensions dataset Y by matching the pairwise similarity of points in χ and Y transforming distances into probabilities

$$\chi = \{x_1, x_2, ..., x_n\} \quad x_n \epsilon \mathbb{R}^d$$
$$Y = \{y_1, y_2, ..., y_n\} \quad y_n \epsilon \mathbb{R}^2$$

Similarity of x_j to x_i in High dimension is defined as probability $p_{j|i}$

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

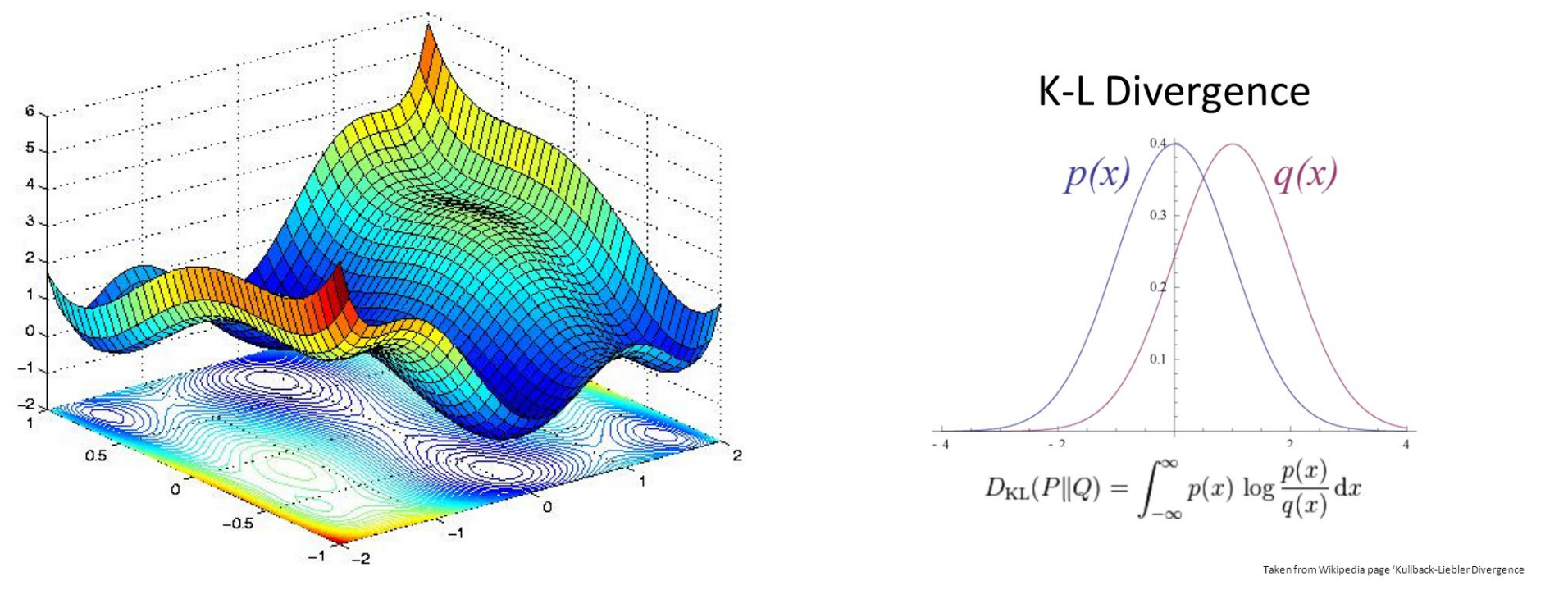
Similarity of y_j to y_i is defined as probability $q_{j|i}$

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

Perplexity of P_i is user fixed and determines σ_i

$$Perp(P_i) = 2^{H(P_i)} \quad H(P_i) = -\sum_j p_(j|i) \log_2 p_(j|i)$$

To maintain the data structure the difference between $p_{j|i}$ and $q_{j|i}$ is minimized. The cost function C is the KL-divergence between both conditional distributions



(a) 3D KL divergence (b) Diference Between Distributions

Figure 3: t-SNE cost function.

t-SNE

Symmetric Cost Function based on symmetric joint probability p_{ij} and a heavy tailed low dimensional joint probability q_{ij}. q_{ij} is a Student t-distribution with one degree of freedom:

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

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

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} log \frac{p_{ij}}{q_{ij}}$$

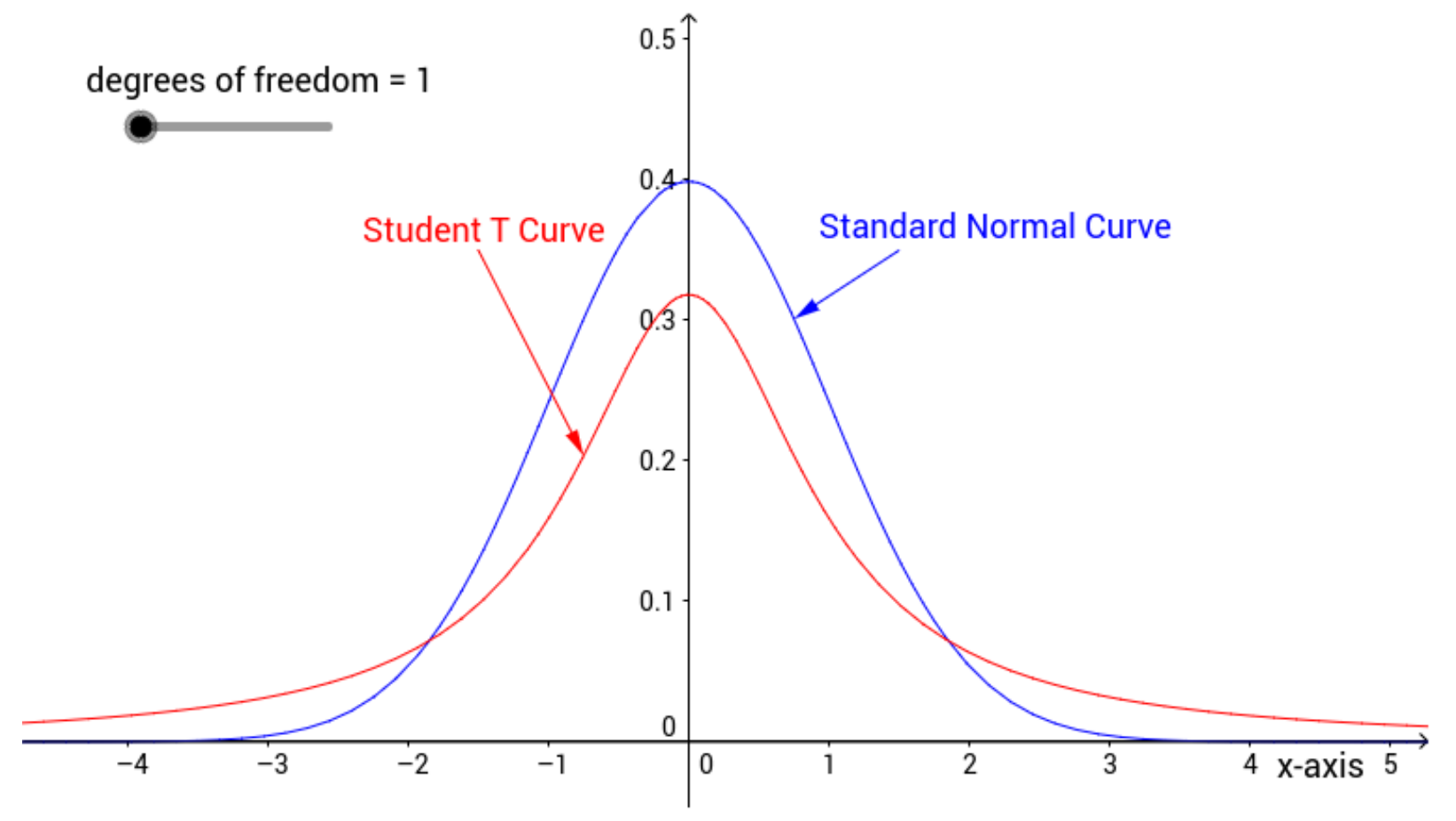


Figure 4: t-student vs normal distribution

Conclusion

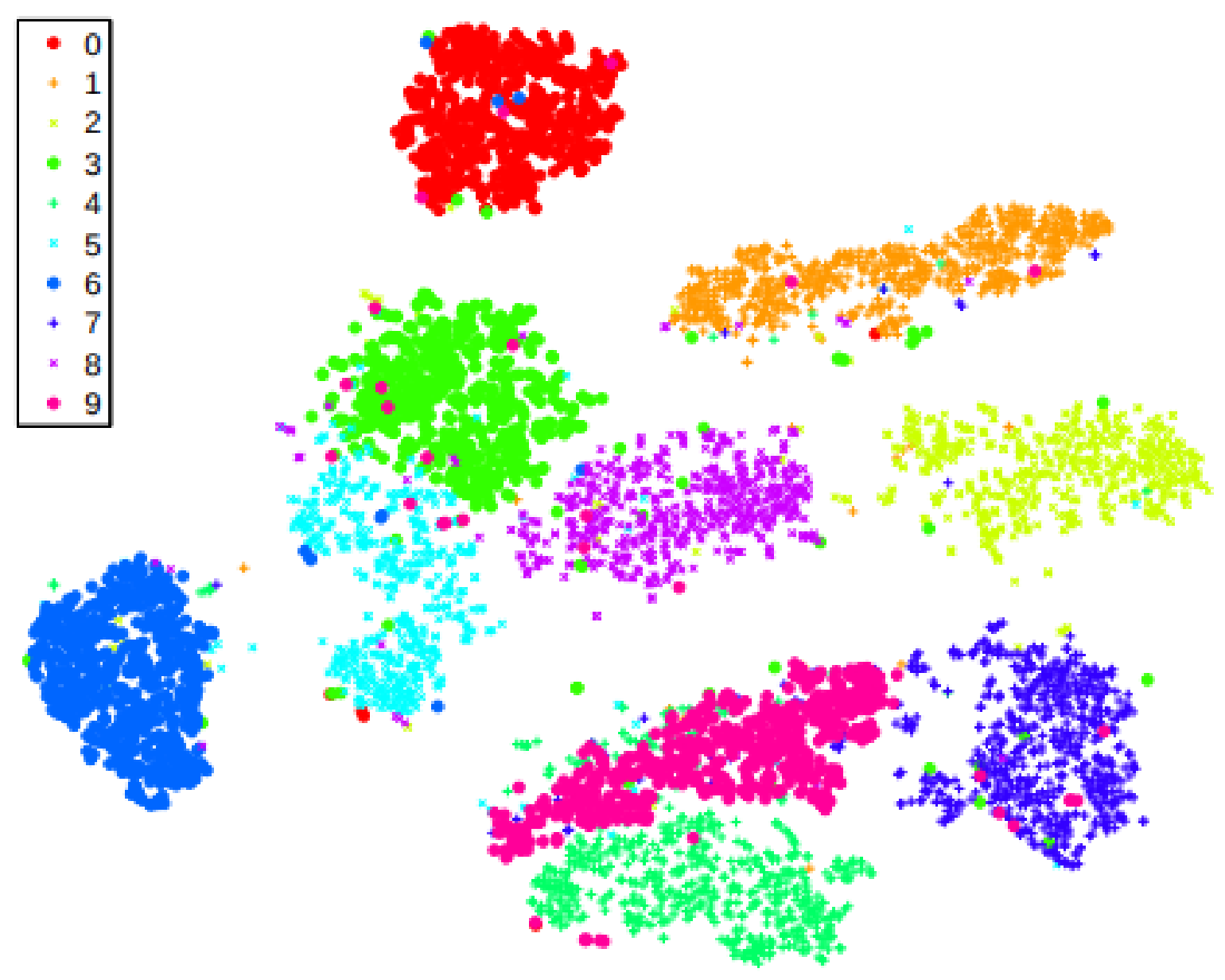


Figure 5: Visualization of 6,000 handwritten digits from the MNIST dataset.

t-SNE is able to effectively represent the high dimensional data structure in just few dimension. Both the computational and the memory complexity of t-SNE is $O(N^2)$.

Limitations:

- Computational cost $O(N^2)$
- Performance on general dimension reduction tasks unclear.
- Non-convex cost function: the convergence of the convergence of the algorithm is not guaranteed. While most of the state of the art algorithms do have a convex cost function.

References

[1] Geoffrey E Hinton and Sam T Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

THE UNIVERSITY of EDINBURGH