BAINSA

Visualizing Data using t-SNE

By: Albertini Federico
Bertamini Riccardo
Calamita Corrado

INDEX

Visualizing Data using t-SNE

Algorithm 1: Simple version of t-Distributed Stochastic Neighbor Embedding.

Data: data set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$,
cost function parameters: perplexity $Perp$,
optimization parameters: number of iterations T, learning rate η, momentum $\alpha(t)$.
Result: low-dimensional data representation $\mathcal{Y}^{(T)} = \{y_1, y_2, \ldots, y_n\}$.
begin
compute pairwise affinities $p_{j|i}$ with perplexity $Perp$ (using Equation 1)
set $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$
sample initial solution $\mathcal{Y}^{(0)} = \{y_1, y_2, \ldots, y_n\}$ from $\mathcal{N}(0, 10^{-4} I)$
for *t=1* **to** T **do**
compute low-dimensional affinities q_{ij} (using Equation 4)
compute gradient $\frac{\delta C}{\delta \mathcal{Y}}$ (using Equation 5)
set $\mathcal{Y}^{(t)} = \mathcal{Y}^{(t-1)} + \eta \frac{\delta C}{\delta \mathcal{Y}} + \alpha(t) \left(\mathcal{Y}^{(t-1)} - \mathcal{Y}^{(t-2)} \right)$
end
end

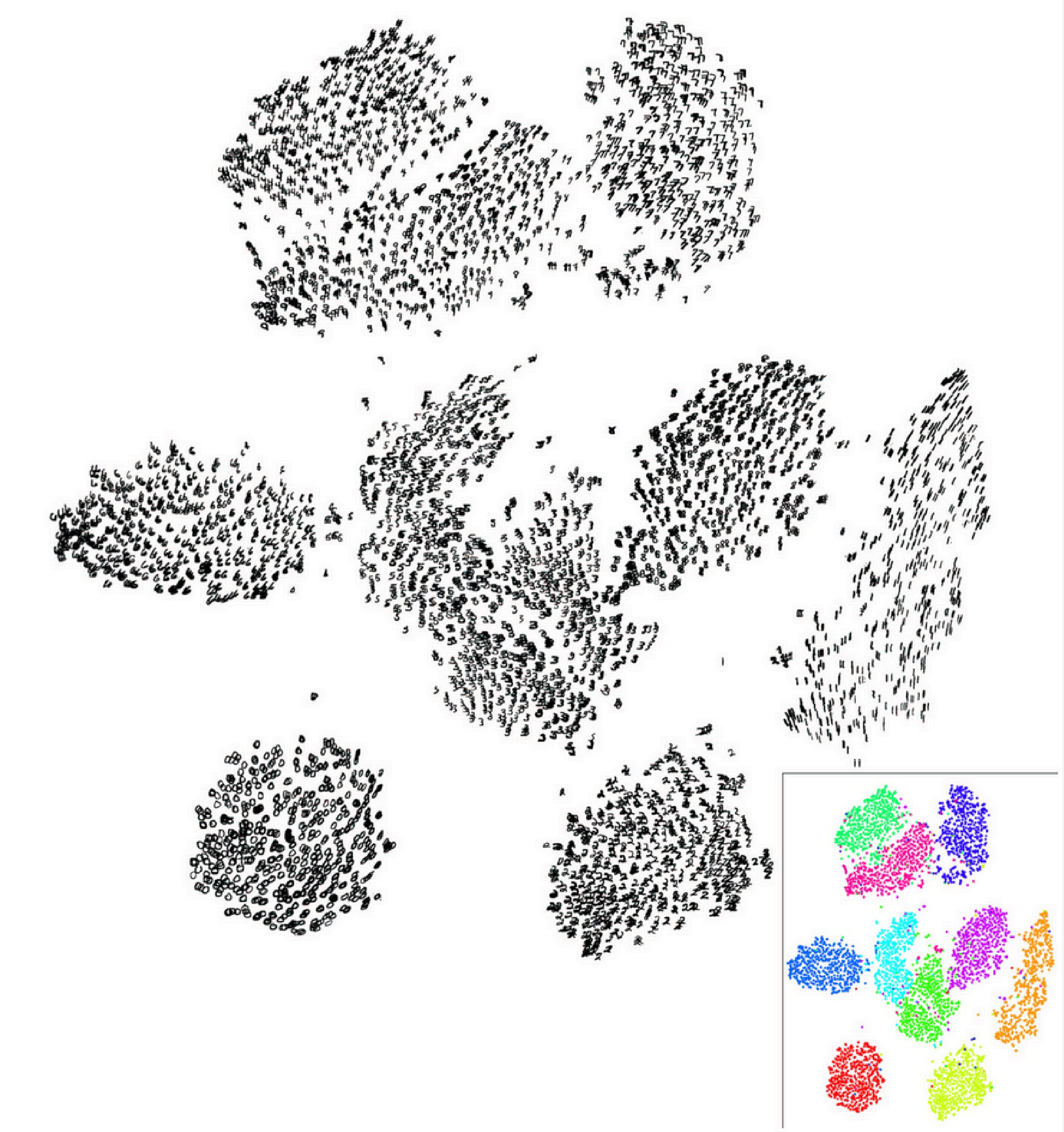
Goal:

analyse

visualize

understand

Dimensionality reduction



Modern datasets are extremely high-dimensional, with high correlated data

Problems

overfitting

computational complexity

unintuitive visualization

How to deal with it?

BAINSA

Linear methods

PCA

Non linear methods

t-distributed Stochastic Neighbor Embedding

Stochastic Neighbor Embedding

given x_i what is the probability to pick x_j as it's neighbor?

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

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

→ we are interested in modelling pairwise similarities

Cost function

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

idea:

if the map points y_i and y_j correctly model the similarity between the high dimensional data points x_i and x_j, the conditional probabilities will be equal

uses a Kullback-Leinbler divergence

we want to minimize it

Minimisation of the cost function

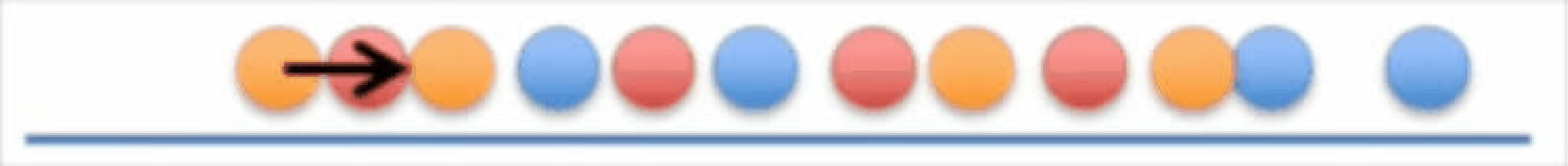
It is performed using a gradient descent method where the gradient has a surprisingly simple form:

$$\frac{\delta C}{\delta y_i} = 2 \sum_j (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j)$$

Phisically it’s like all the map points are connected by some springs that exert a repelling or attracting force

Gradient descent method

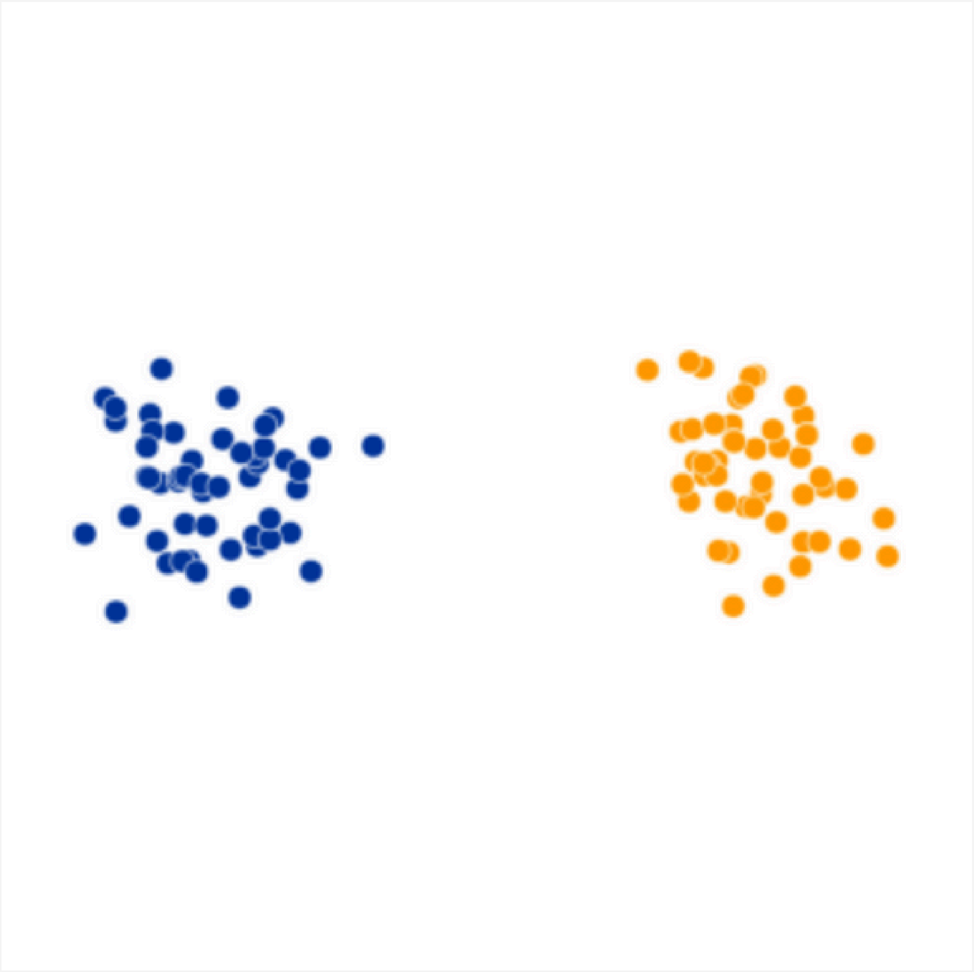
reducing from 2D to 1D



(same color means the point belong to the same cluster)

At each step, a point on the line is attracted to points it is near in the scatter plot, and repelled by points it is far from

Original



t-sne embedding

Sigma value σ_i

SNE performs a binary search to find the right value that produces a P_i with a fixed perplexity that is specified by the user

$$Perp(P_i) = 2^{H(P_i)}$$

The crowding problem in SNE

1

difficult to preserve distances in some cases

2

even the small (not relevant) forces of the springs, summed for the number of data points creates a strong force that crushed the points in the center of the map

3

we are trying to embed very high dimensional data into a low dimensional space

t-SNE (differences)

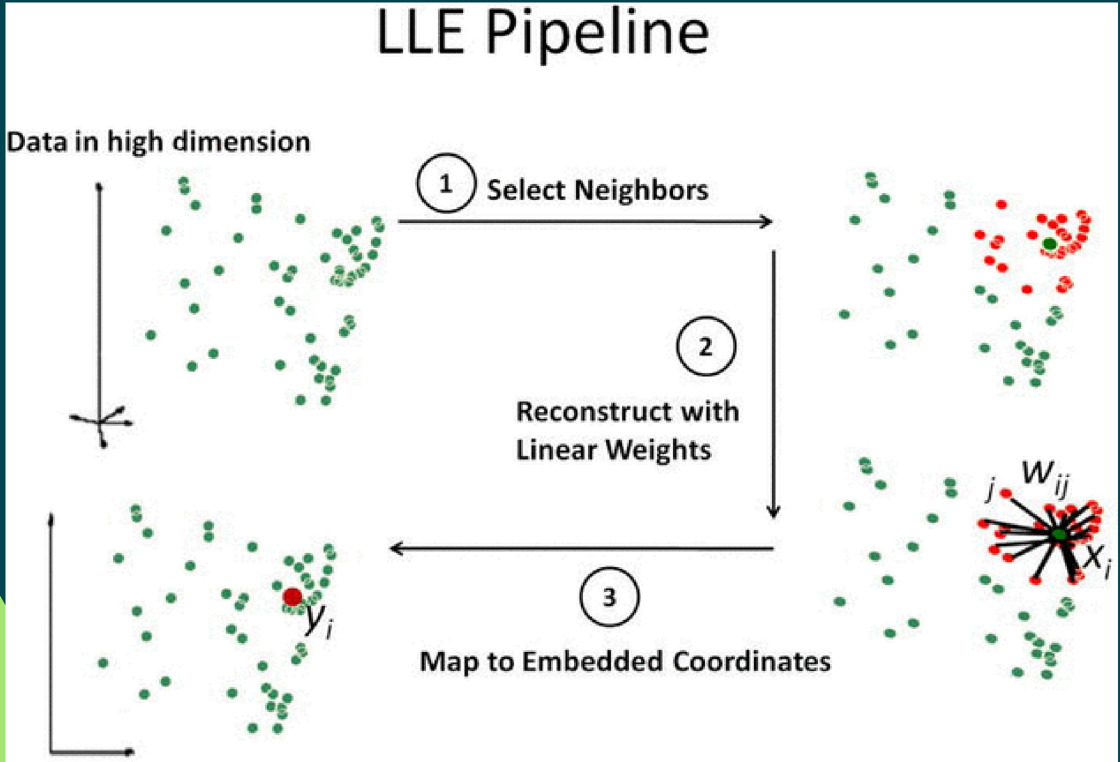
t-	symmetric
it uses a Student t-distribution with one degree of freedom, also called Cauchy distribution in the low dimensional space	$p_{ij} = p_{ji}$ and $q_{ij} = q_{ji}$
$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$ inverse square law	symmetrized conditional probabilities and it uses a simpler form of gradient which is faster to compute

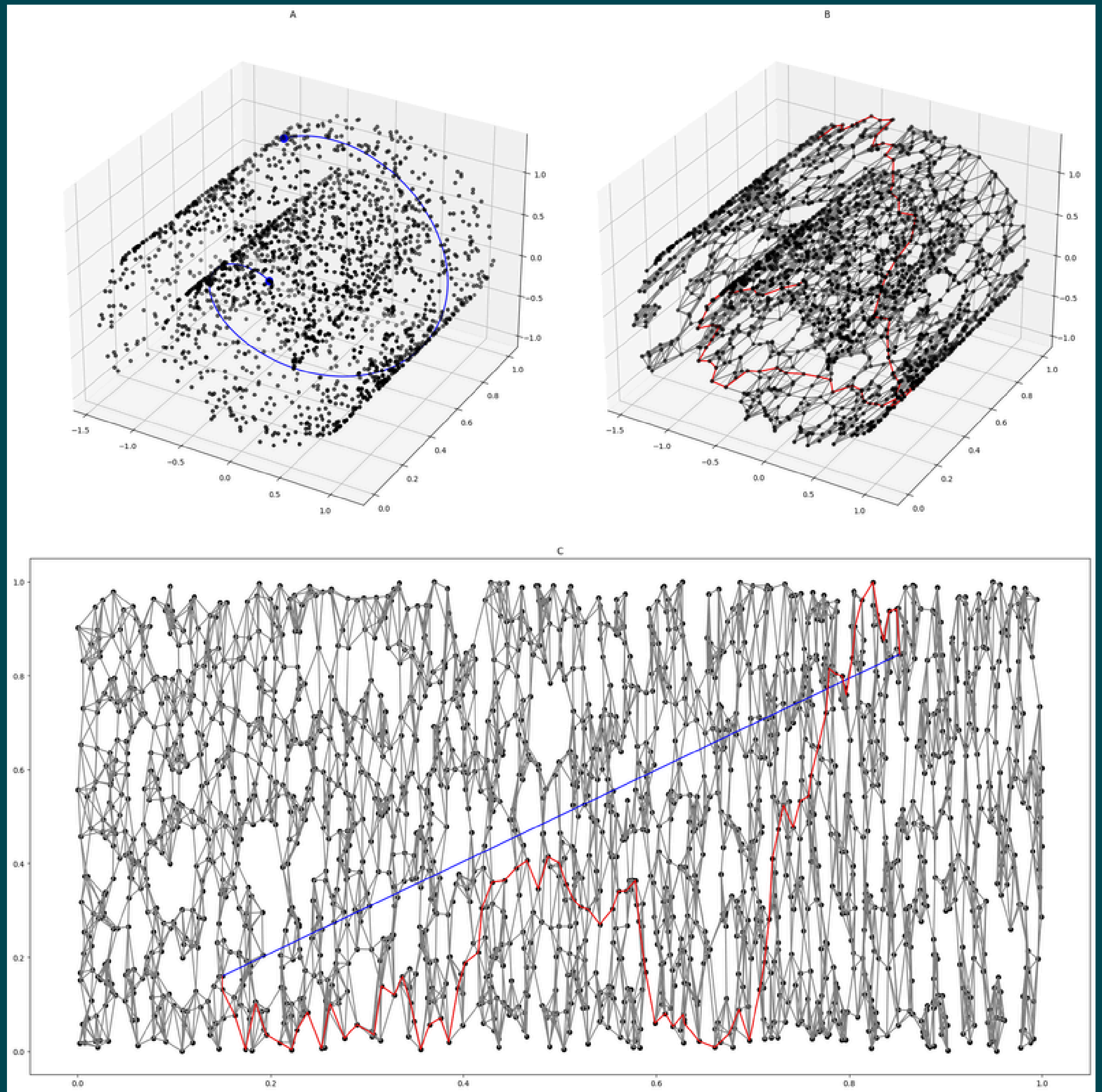
Experiments

$$E = \frac{1}{\sum_{i<j} d^*_{ij}} \sum_{i<j} \frac{(d^*_{ij} - d_{ij})^2}{d^*_{ij}}.$$

Performance Evaluation of:

1. t-SNE
2. Sammon Mapping
3. Isomap
4. LLE





Data-Sets

MNIST data-set

60 000 grayscale images of handwritten digits and pictures are 28x28

Olivetti faces data-set

400 images of 40 individuals, each image has a unique viewpoint (and in some cases also glasses) and pictures are 92x112 pixels

Coil-20

1440 images of 20 objects from 72 equally spaced orientations, pictures are 32x32 pixels

Experimental Setup

use PCA to reduce the data to 30 dimensionalities

Use the dimensionality reduction technique to go from 30d to 2d and plot results

The Scatterplots:

- information about each single datapoint
- class information used to select colors/symbols, not to determine spatial coordinates of the map points
- coloring used to evaluate how well the map preserves similarities within each class

Technique	*Cost function parameters*
t-SNE	$Perp = 40$
Sammon mapping	none
Isomap	$k = 12$
LLE	$k = 12$

Perp -> is the perplexity of the conditional probability distribution induced by a Gaussian Kernel (p)

$$\text{Perplexity} = 2^{H(P)}$$

Perplexity as a function of entropy

$$H(P_i) = -\sum_{j \neq i} p_{j|i} \log_2 p_{j|i}$$

entropy

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

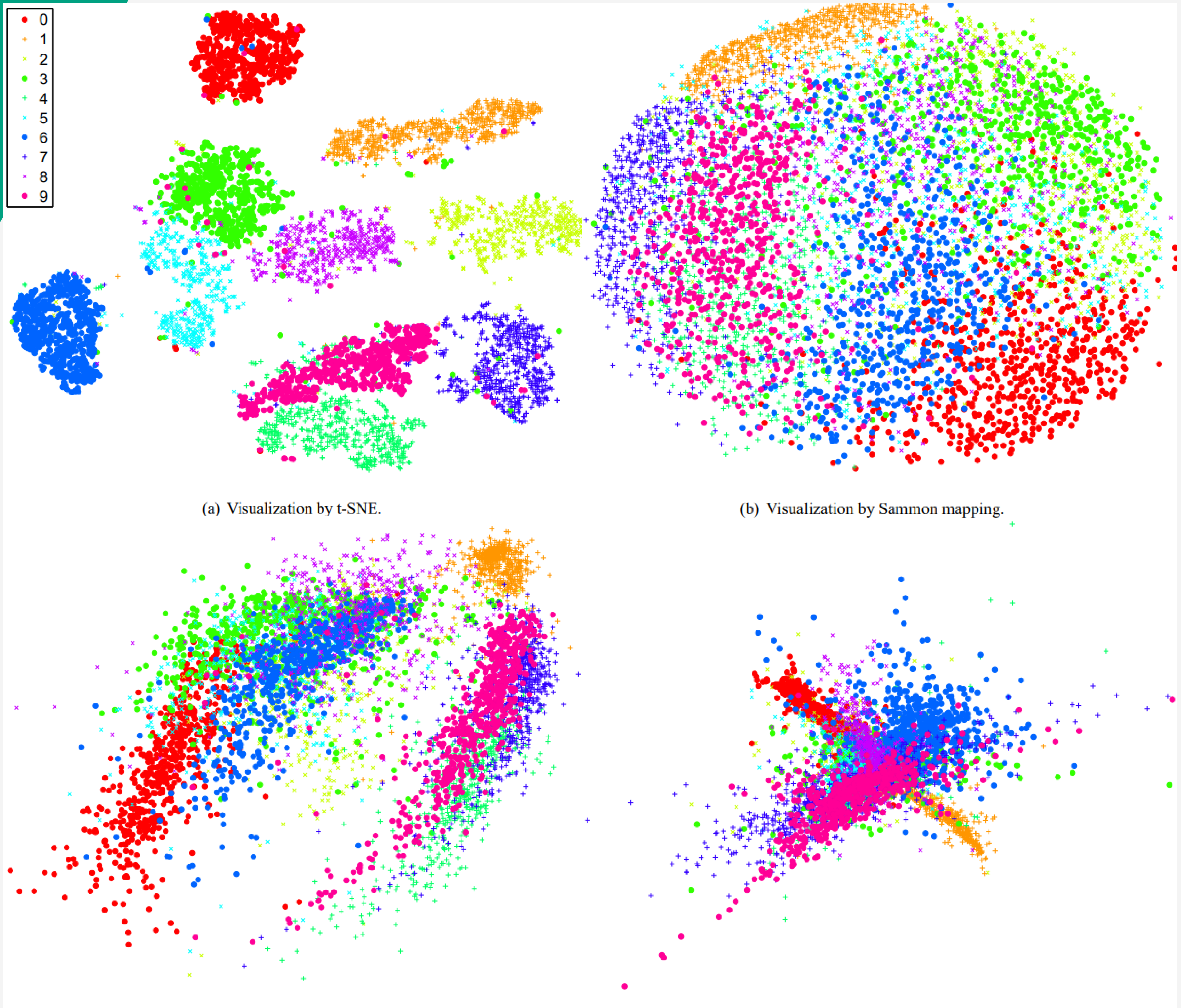
sigma is the scale parameter of the Gaussian Kernel

The Perplexity is hence used to represent the effective number of neighbors each point has and:

- sigma very small -> distribution becomes nearly deterministic, low perplexity
- sigma very large -> distribution becomes nearly uniform over all points, leading to high perplexity

Results

MNIST DATA-SET



(a) Visualization by t-SNE.

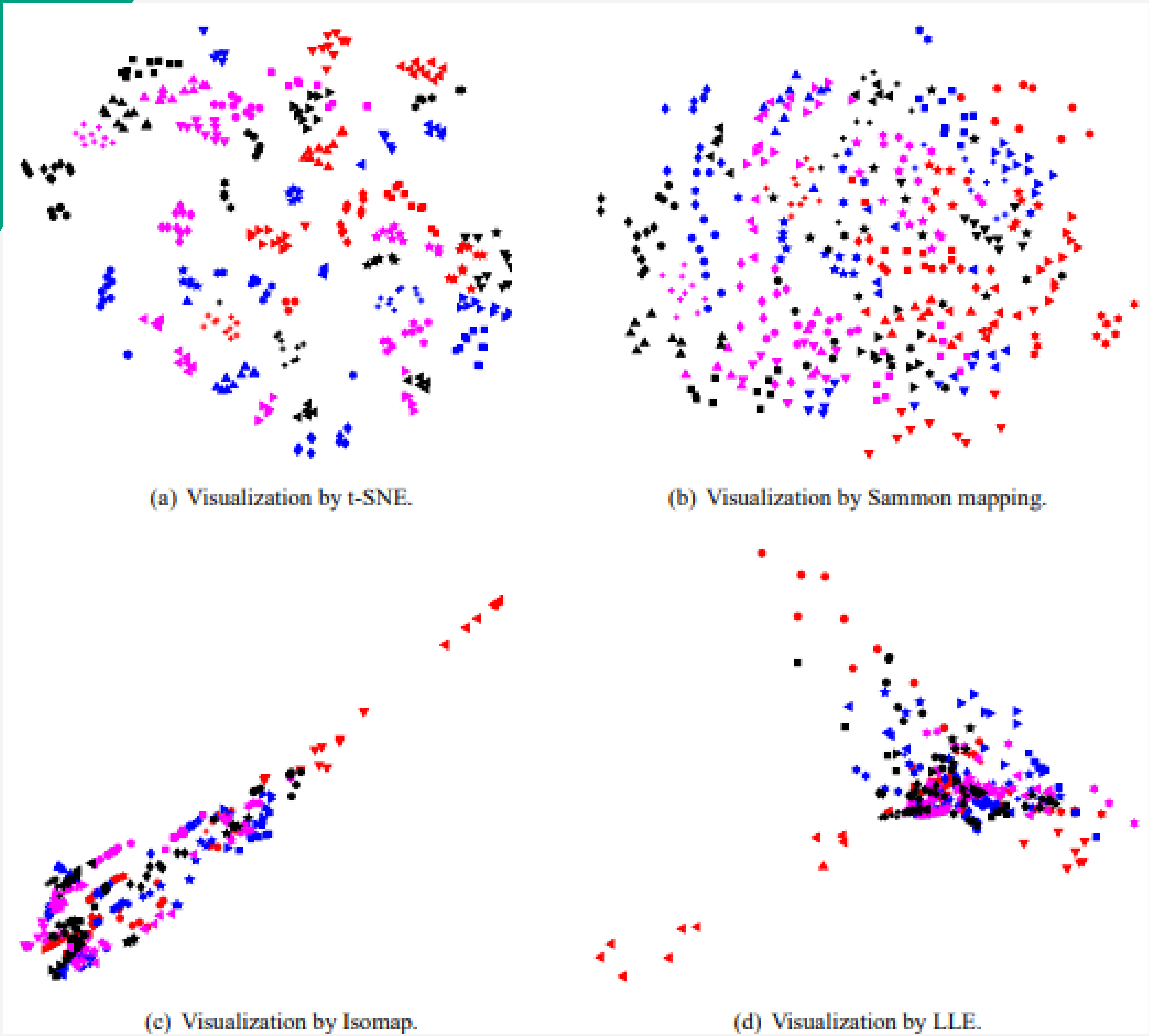
(b) Visualization by Sammon mapping.

(a) Visualization by Isomap.

(b) Visualization by LLE.

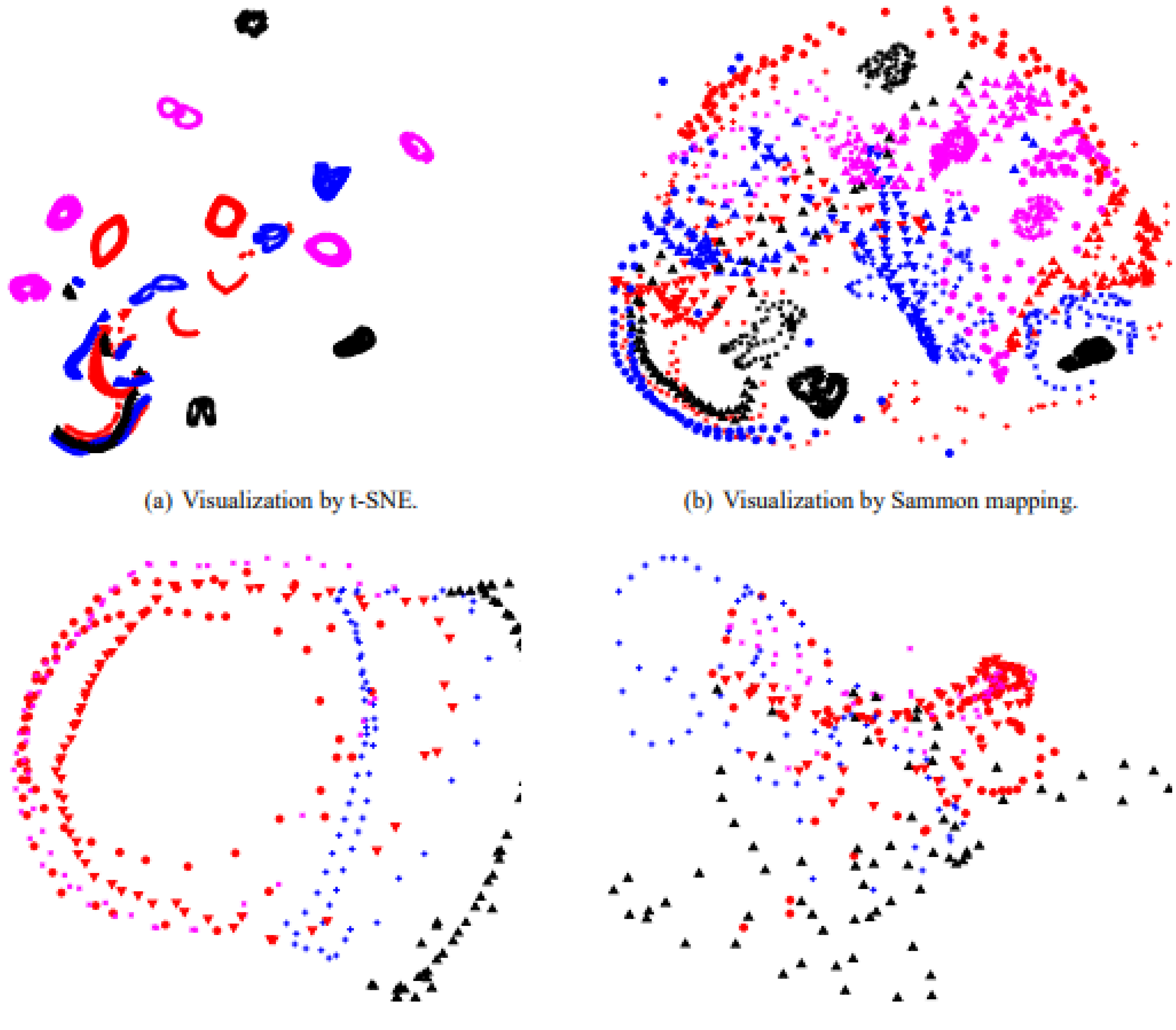
Results

OLIVETTI FACES DATA-SET



Results

COIL-20 DATA-SET



(a) Visualization by t-SNE. (b) Visualization by Sammon mapping.

(c) Visualization by Isomap. (d) Visualization by LLE.

LARGE DATASETS

O(n^2) Computational cost

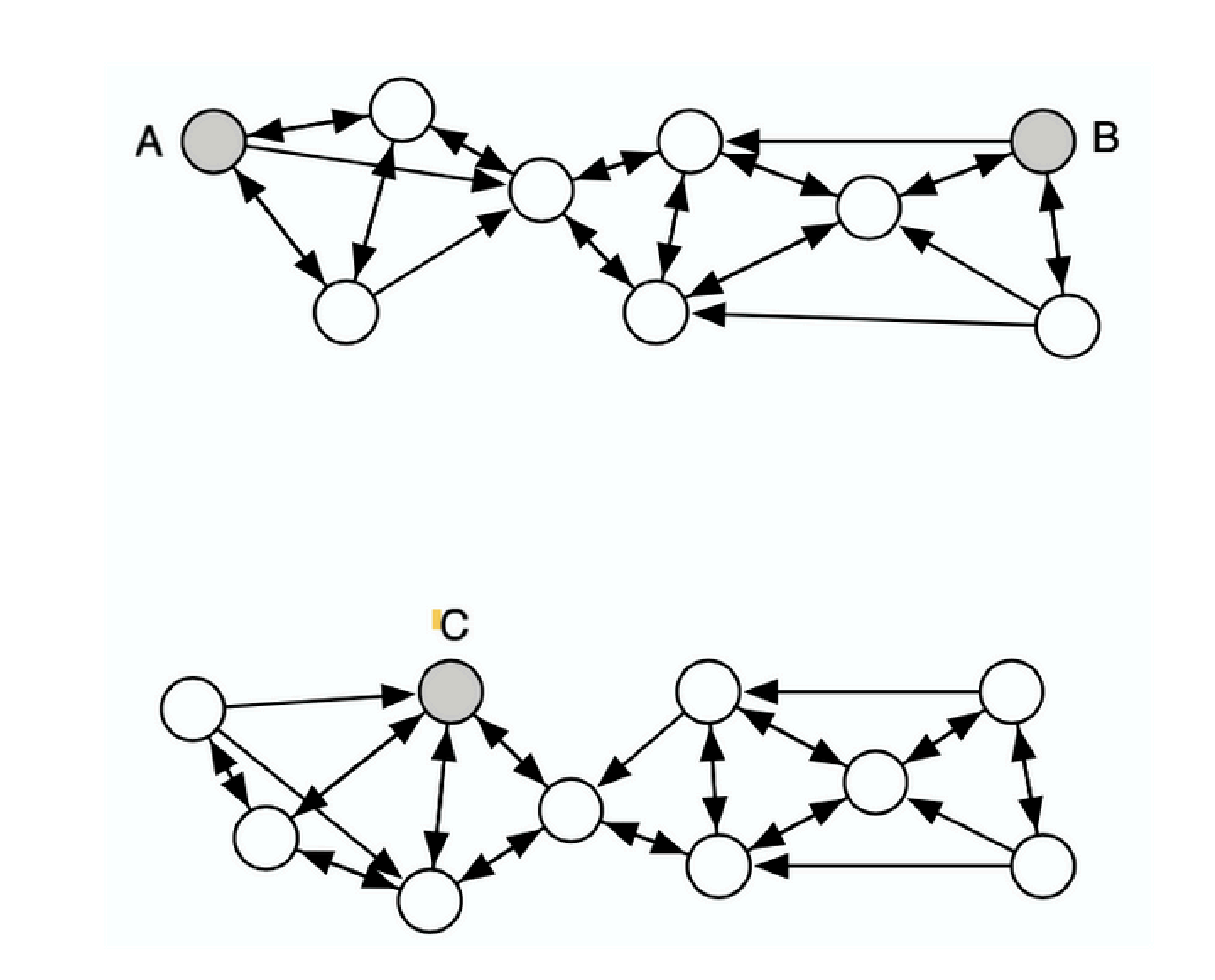
Infeasable on large datasets

Using subsets of the dataset leads to wrong results

Solutions:

1. Random Walk Approach
2. Analytical Approach

Random Walk



Select landmarks

Landmarks are a number of points arbitrarily selected

Create Neighbourhoods

Neighborhoods are created by selecting a hyperparameter k and creating a graph connecting each vertex to the k other closest vertices

---

Perform random walks

Perform random walks on the edges until two landmark masses connect where the probability of selecting an edge is $e^{-\|x_i - x_j\|^2}$

Analytical Solution

Solve system of equations

A system of sparse linear equations can be solved to find the pairwise similarities

Effectiveness

In the experiments presented in the paper there did not seem to be a significant difference between random walk and analytical solution

Pros and cons

The analytical solution is more computationally intensive, although it is more effective on high dimensional data that presents very sparse data

Comparison with other techniques

Preserving long distances

Classical (linear) scaling
Isomap
LLE
Diffusion maps

Preserving short distances

Sammon mapping
CCA
MVU

Weaknesses

1. Dimensionality reductions with more than 3 dimensions
2. Assumption of linearity of the manifold (Euclidean distance)
3. Non-convexity of the of the cost function

Future steps

1. Investigate using t-distributions with higher df
2. Combining t-SNE with neural networks to explicitly map manifolds