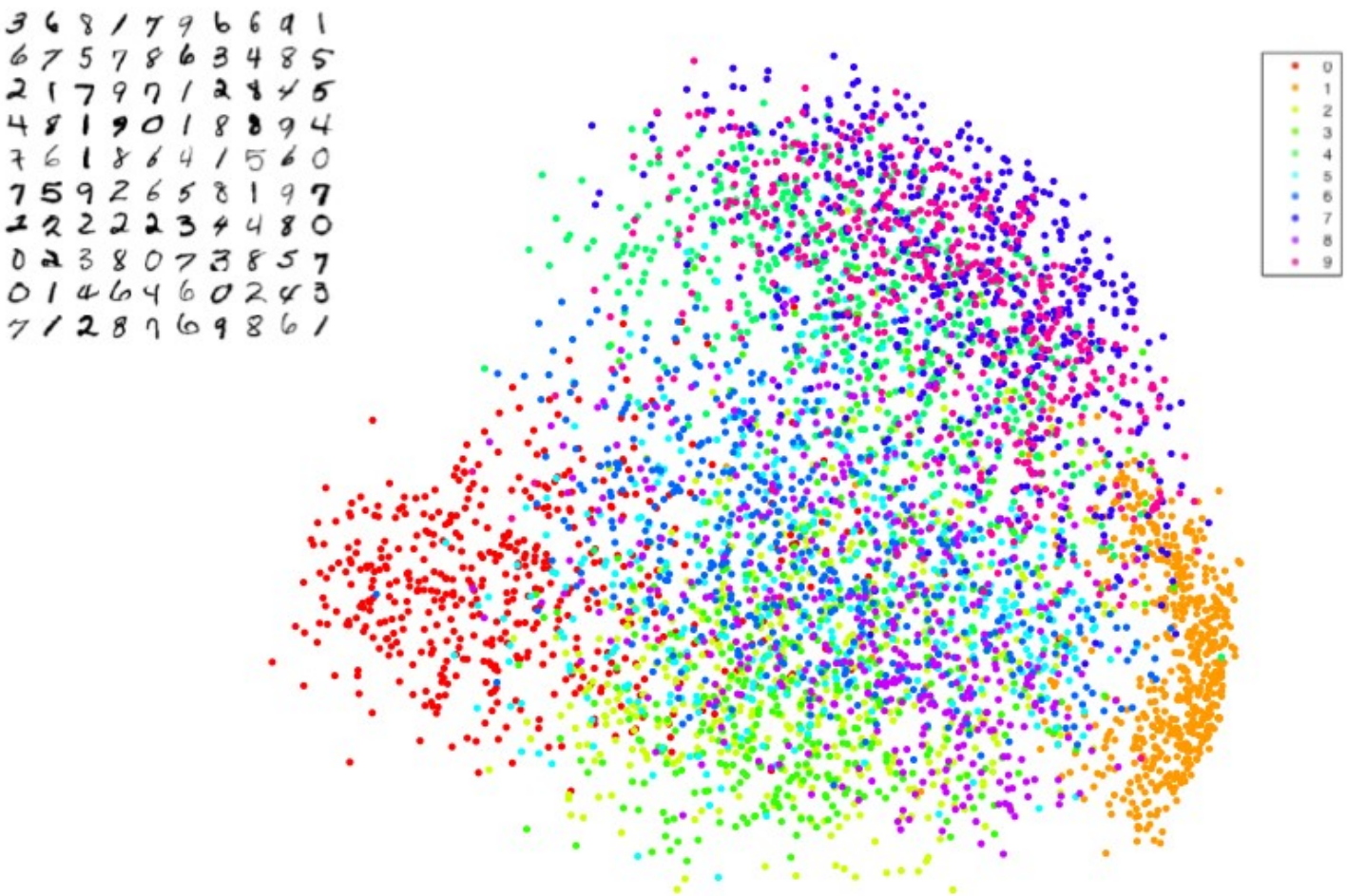
Advanced Data Analytics
Week 1: SNE

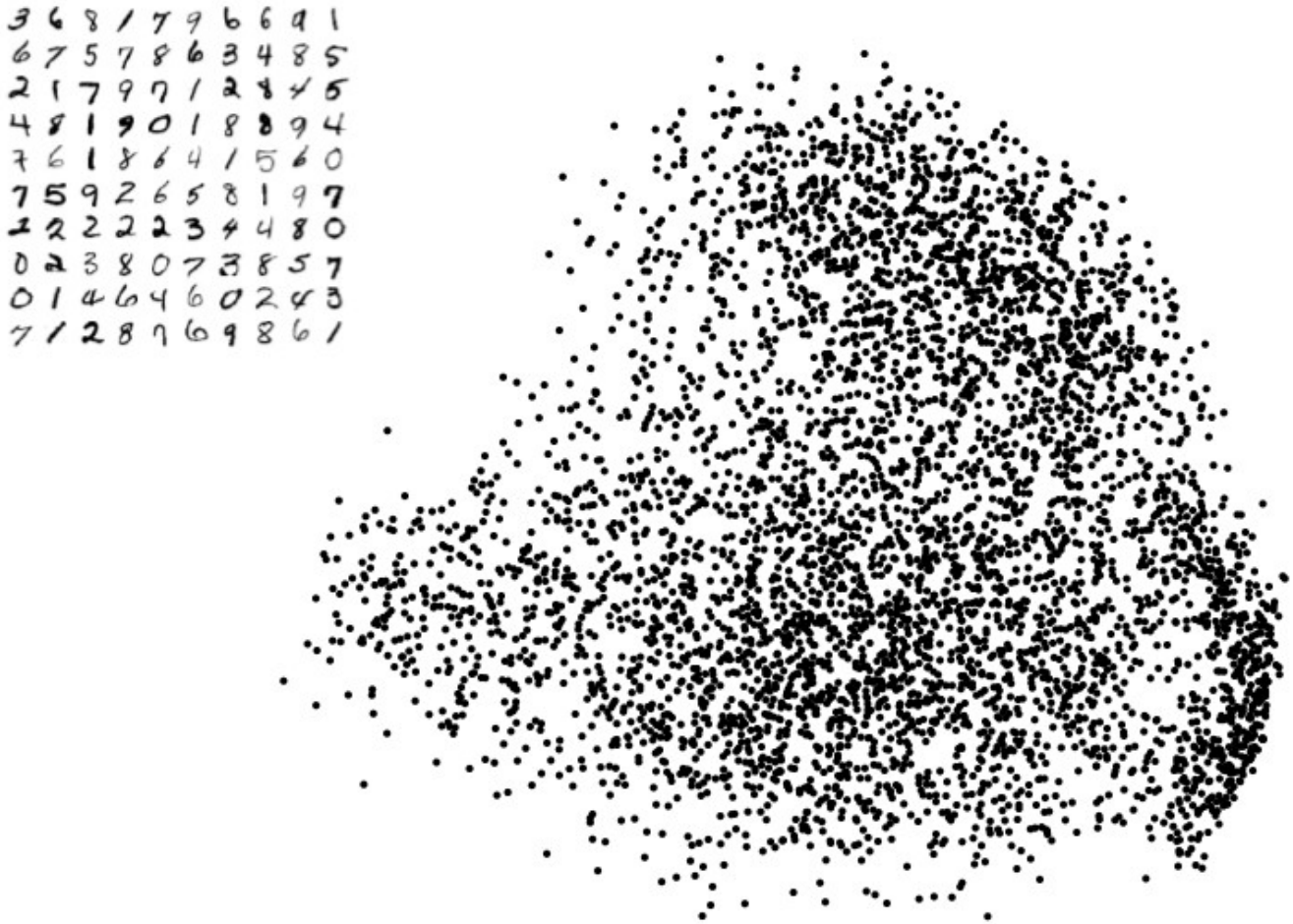
Ian T. Nabney

University of BRISTOL

Limitations of PCA: embedding MNIST

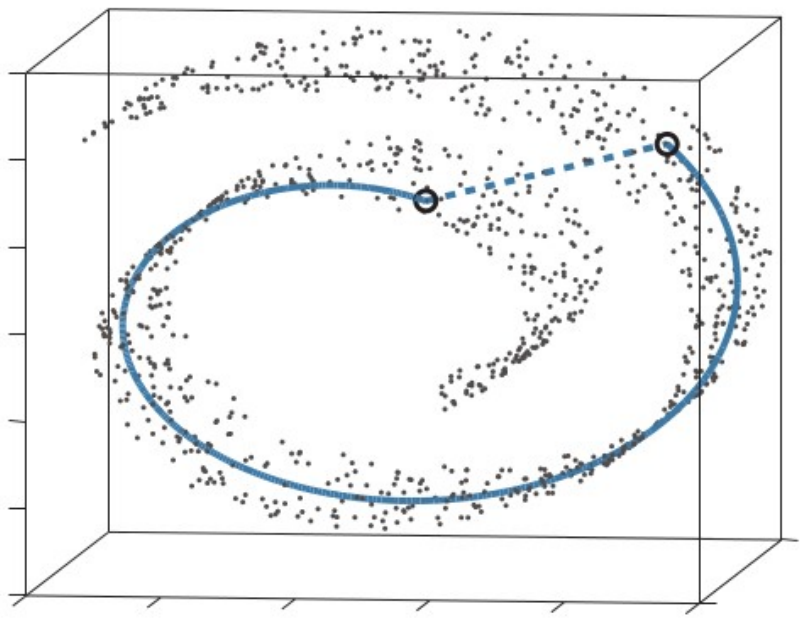


PCA without colour coding



Local structure

- Stress metric measures all distances, but squared terms mean that large distances dominate.
- Similarly, techniques such as PCA consider the global representation.
- But is this how we see the data? Think back to Ware's book.
- Arguable that we actually understand neighbourhoods best.



University of BRISTOL

Stochastic Neighbourhood Embedding

- Start by converting Euclidean distances into conditional probability that represent similarities

$$p_{j|i} = \frac{\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2\right)} \tag{1}$$

  where σ_i^2 is the variance of a Gaussian centred on $\mathbf{x}_i$. We set $p_{i|i} = 0$.

- We create a similar metric in the low-dimensional space

$$q_{j|i} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k\neq i}\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)} \tag{2}$$

- If the mapped points $\mathbf{y}_i$ and $\mathbf{y}_j$ correctly model the similarity between the high-dimensional datapoints $\mathbf{x}_i$ and $\mathbf{x}_j$, the conditional probabilities $p_{j|i}$ and $q_{j|i}$ will be equal. Measure the mismatch using the Kullback-Leibler divergence

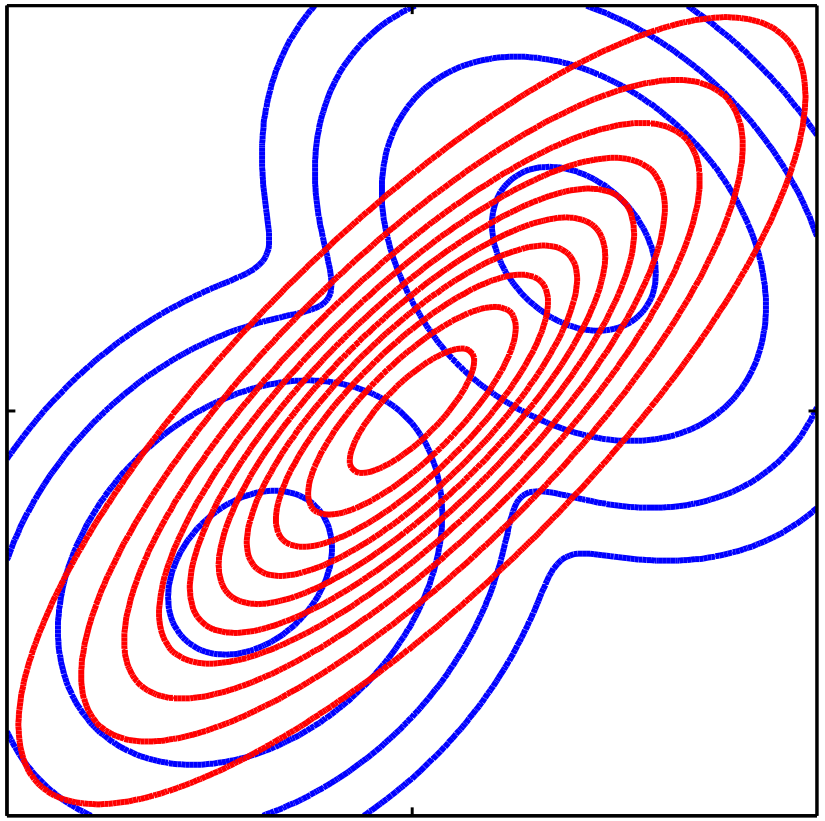
$$C = \sum_i KL(P_i\|Q_i) = \sum_i\sum_j p_{j|i}\log\frac{p_{j|i}}{q_{j|i}}, \tag{3}$$

  where P_i represents the conditional probability over all other datapoints given data point $\mathbf{x}_i$ and similarly for Q_i. Minimise the cost function using gradient descent (and partial derivatives).

University of BRISTOL

Interpreting the cost function

- Different types of error in the pairwise distances in the low-dimensional map are not weighted equally.
- There is a large cost for using widely separated map points to represent nearby datapoints (i.e., for using a small $q_{j|i}$ to model a large $p_{j|i}$).
- There is only a small cost for using nearby map points to represent widely separated datapoints.



Blue contours show a bimodal distribution $p(\mathbf{X})$ and the red contours correspond to a single Gaussian distribution $q(\mathbf{Y})$ that minimises the KL-divergence $KL(p||q)$.

University of BRISTOL

Optimising the cost function

- The gradient has a surprisingly simple form

$$\frac{\partial C}{\partial y_i} = \sum_{j \neq i} (p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})(y_i - y_j) \tag{4}$$

and the update equation with a momentum term

$$\mathbf{Y}^{(t)} = \mathbf{Y}^{(t-1)} - \eta \frac{\partial C}{\partial y_i} + \beta(t)(\mathbf{Y}^{(t-1)} - \mathbf{Y}^{(t-2)}) \tag{5}$$

- Physically, the gradient may be interpreted as the resultant force created by a set of springs between the map point $\mathbf{y}_i$ and all other map points $\mathbf{y}_j$. All springs exert a force along the direction $(\mathbf{y}_i - \mathbf{y}_j)$.
- The force exerted by the spring between $\mathbf{y}_i$ and $\mathbf{y}_j$ is proportional to its length, and also proportional to its stiffness, which is the mismatch $(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})$ between the pairwise similarities of the data points and the map points.

Optimising σ_i

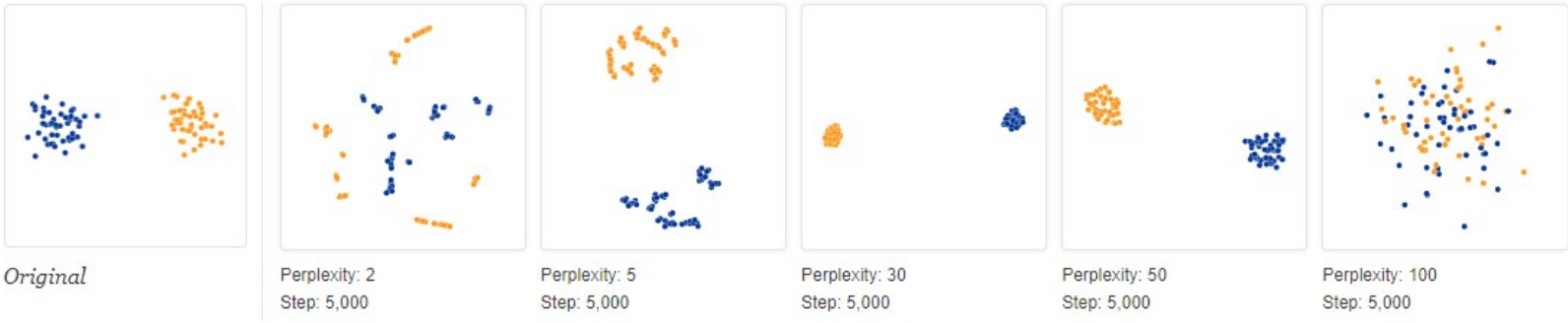
- It is not likely that there is a single value of σ_i that is optimal for all datapoints in the data set because the density of the data is likely to vary.
- In dense regions, a smaller value of σ_i is usually more appropriate than in sparser regions.
- SNE performs a binary search for the value of σ_i that produces a P_i with a fixed perplexity that is specified by the user.
- Perplexity is defined as

$$Perp(P_i) = 2^{H(P_i)} \quad \text{where} \quad H(P_i) = -\sum_j p_{j|i} \log_2 p_{j|i}. \tag{6}$$

- The perplexity can be interpreted as a smooth measure of the effective number of neighbours. The performance of SNE is fairly robust to changes in the perplexity, and typical values are between 5 and 50.

University of BRISTOL

Effect of perplexity



Crowding problem

- In high dimension we have more room, so points can have a lot of different neighbours while this is more difficult in lower dimension.
- This is the crowding problem – we don't have enough room to accommodate all neighbours. This is one of the biggest problems with SNE.
- t-SNE solution: Change the Gaussian in Q to a heavy-tailed distribution.
- If Q changes slower, we have more 'wiggle room'to place points at.

Symmetric SNE

- Introduce two elements to address the crowding problem.
- Replace conditional probabilities $p_{j|i}$ by joint probabilities p_{ji}.

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \quad \text{and} \quad q_{ij} = \frac{\exp\left(-\|\mathbf{y}_i - \mathbf{y}_j\|^2\right)}{\sum_{k \neq l} \exp\left(-\|\mathbf{y}_k - \mathbf{y}_l\|^2\right)} \tag{7}$$

- The gradient is simpler and faster to compute

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j). \tag{8}$$

University of BRISTOL

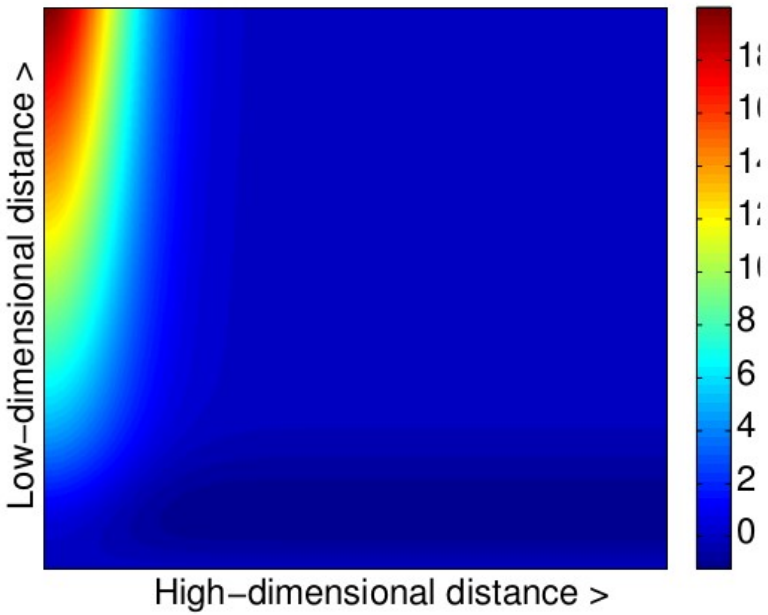
t-distribution Stochastic Neighbourhood Embedding

- Student-t density $p(x) \propto (1 + \frac{x^2}{\nu})^{-(\nu+1)/2}$ which goes to zero much slower than a Gaussian. Choice of $\nu = 1$ is equivalent to the Cauchy distribution.
- Equivalent to averaging Gaussians with a specific prior over σ^2. Removes need to optimise σ_i.
- Redefine q_{ij} but leave p_{ij} the same. This allows a moderate distance in the high-dimensional space to be faithfully modeled by a much larger distance in the map and so it eliminates the unwanted attractive forces between map points that represent moderately dissimilar datapoints.

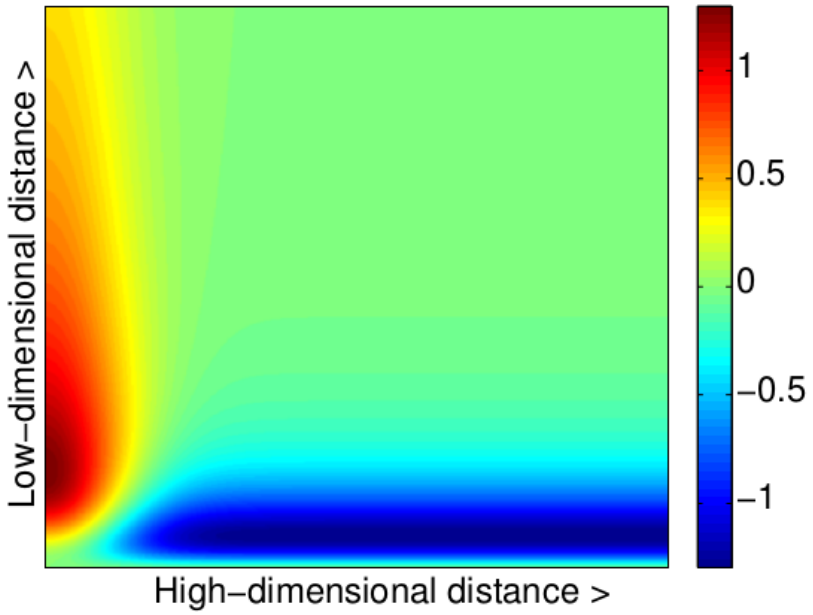
$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}} \tag{9}$$

t-SNE Gradient

$$\frac{\partial C}{\partial \mathbf{y}_i} = 4 \sum_{j \neq i} (p_{ij} - q_{ij})(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}(\mathbf{y}_i - \mathbf{y}_j) \tag{10}$$



Gradient of SNE



Gradient of t-SNE

Experiments

MNIST

- Randomly selected 6,000 images
- $28 \times 28 = 784$ pixels

Olivetti faces

- 400 images (10 per individual)
- $92 \times 112 = 10,304$ pixels

COIL-20

- 20 different objects and 72 equally spaced orientations, yielding a total of 1,440 images
- $32 \times 32 = 1024$ pixels

Use PCA to reduce dimensionality of data to 30.

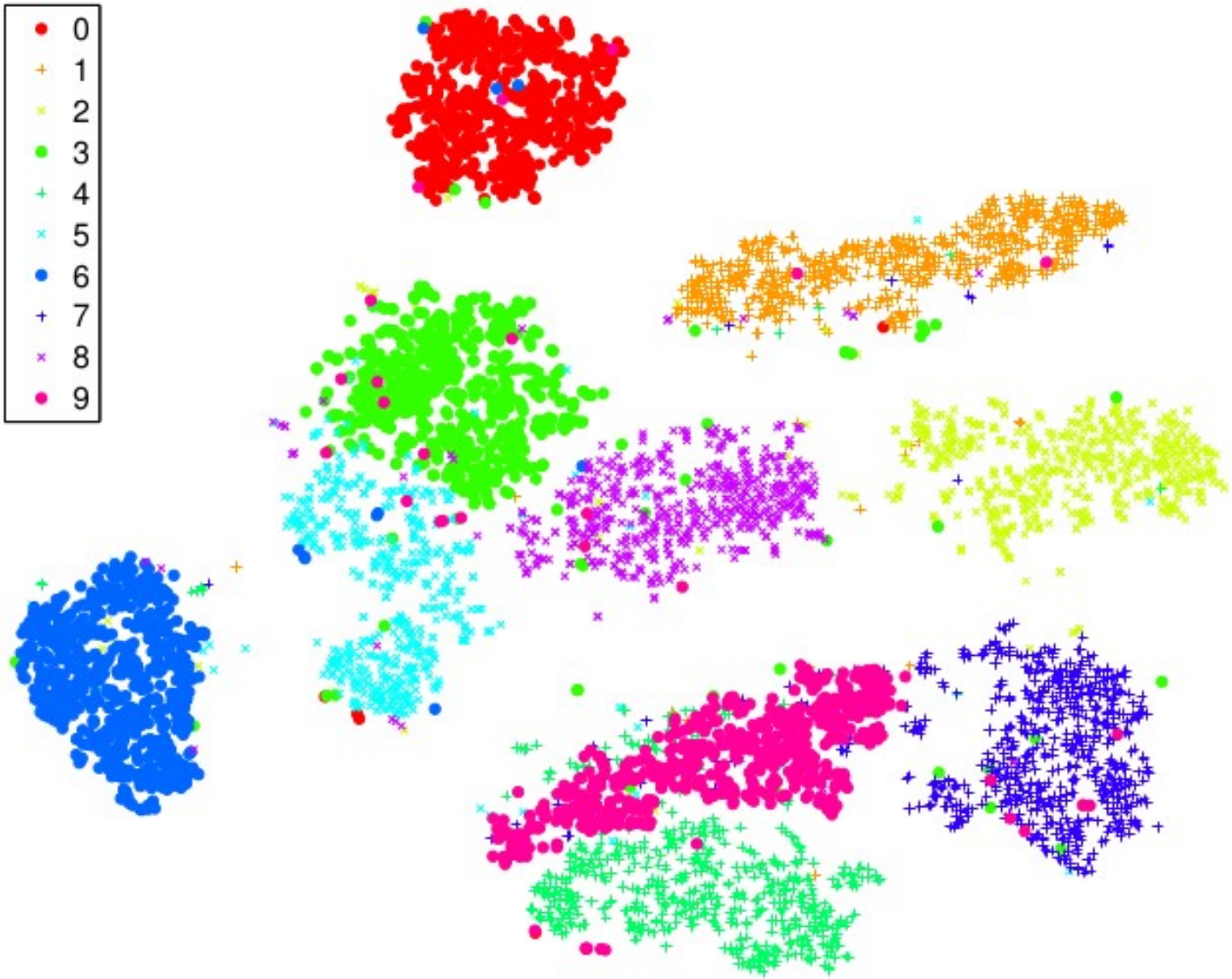
University of BRISTOL

Experimental parameters

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

Table 1: Cost function parameter settings for the experiments.

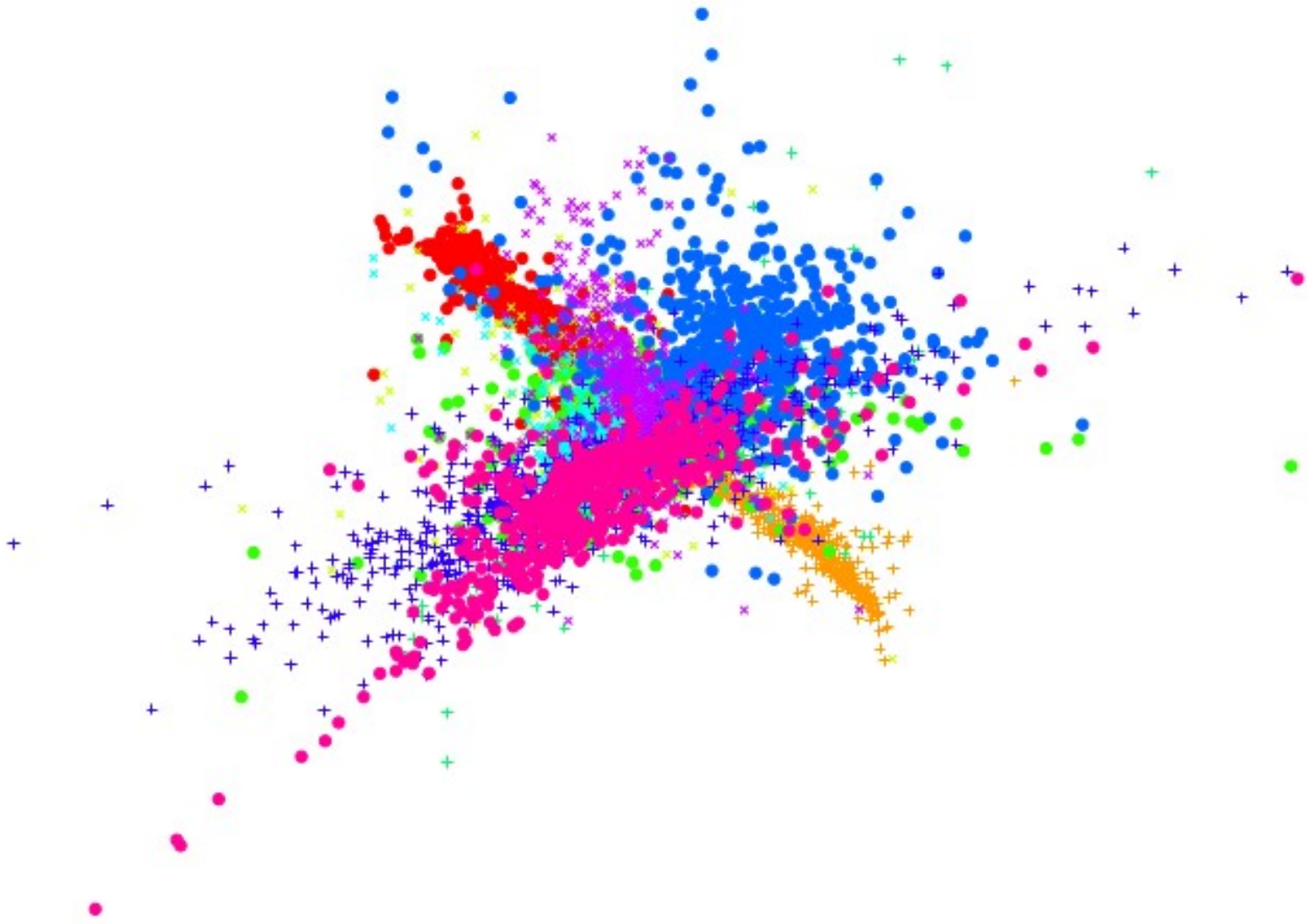
MNIST t-SNE



MNIST Sammon

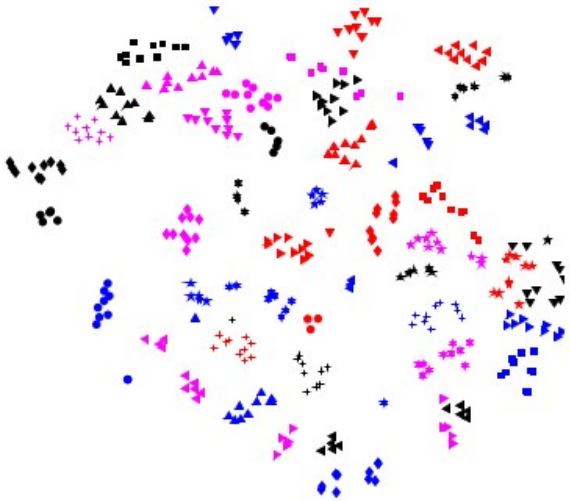
MNIST Isomap

MNIST local linear embedding (LLE)

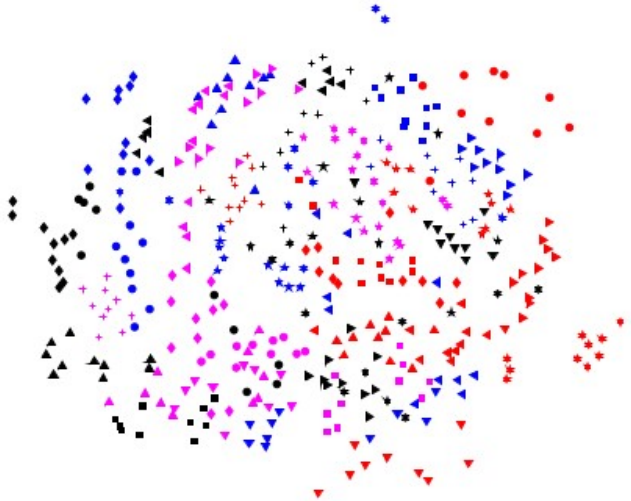


University of BRISTOL

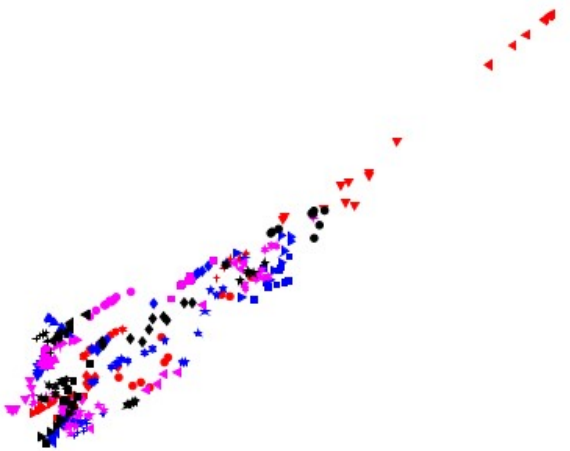
Olivetti faces



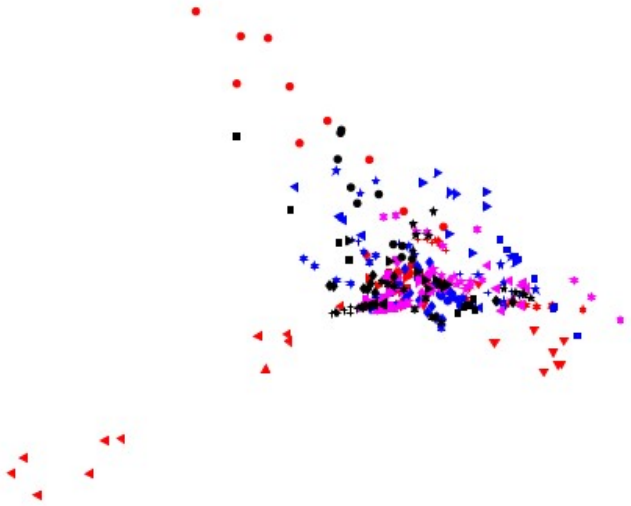
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



(c) Visualization by Isomap.

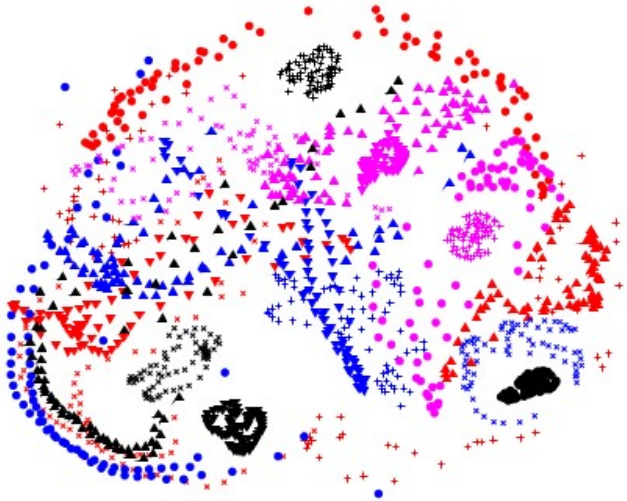


(d) Visualization by LLE.

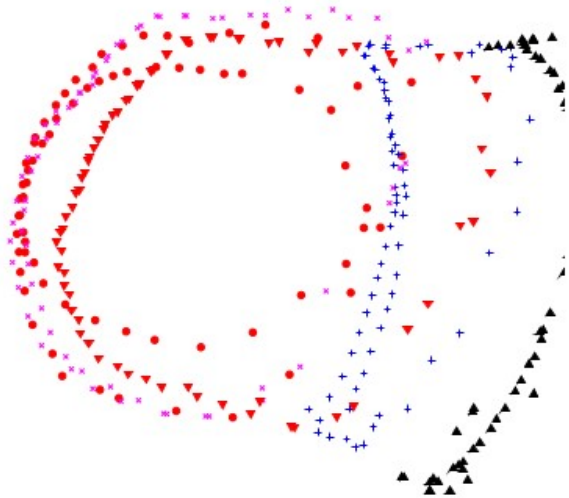
University of BRISTOL

COIL-20 results

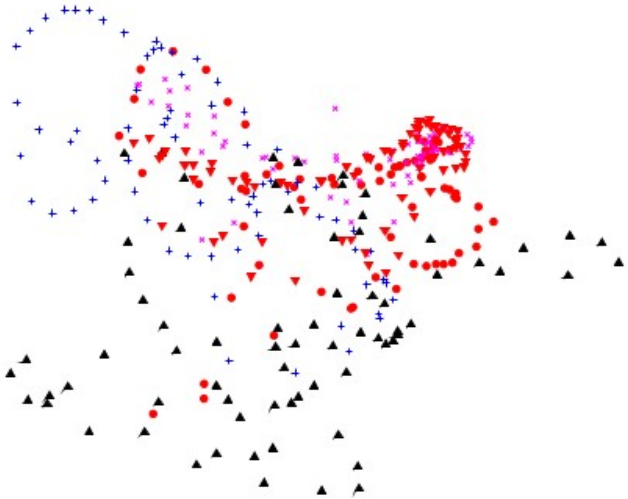
(a) Visualization by t-SNE.



(b) Visualization by Sammon mapping.



(c) Visualization by Isomap.



(d) Visualization by LLE.

Conclusions

- Both the computational and the memory complexity of t-SNE are $O(n^2)$, but the technique makes it possible to successfully visualize large real-world data sets with limited computational demands.
- We can reduce complexity from $O(N^2)$ to $O(N \log N)$ via Barnes Hut (tree-based) algorithm.
- Experiments on a variety of data sets show that t-SNE outperforms existing state-of-the-art techniques for visualizing a variety of real-world data sets.
- No functional mapping.

University of BRISTOL