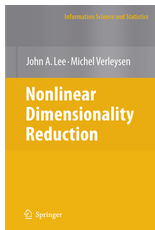


Visualisation et réduction de dimension

Stéphane Canu

asi.insa-rouen.fr/enseignants/~scanu
scanu@insa-rouen.fr

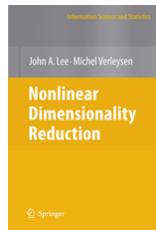


École d'été BasMatI 2018

Porquerolles, june 6, 2018

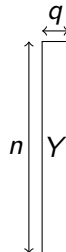
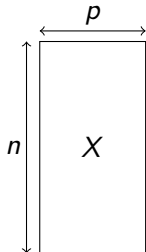
Lecture road map

- 1 Introduction to hidden variables
- 2 PCA: principal component analysis
- 3 Distance preservation approaches (global methods)
- 4 Local approaches

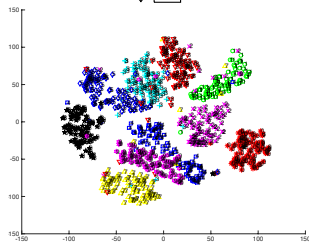


A dimensionality reduction problem

Given X a n by p data matrix, find a Y $n \times q$ matrix with $q < p$



$p = 784$



$q = 2$

Dimensionality reduction: what for ?

- Visualize ($q = 2$ ou 3)
 - ▶ validate data coding
 - ▶ detect outliers and liss labeled data
 - ▶ visualize classes
- Represent ($q < p$)
 - ▶ summarize (remove noise)
 - ▶ preprocessing: brings statistic and computation efficiency
 - ▶ the hidden variable hypothesis

Coding/decoding functions

$$\begin{aligned} \text{cod} : \mathbb{R}^p &\longrightarrow \mathbb{R}^q, \quad \mathbf{x} \longmapsto \mathbf{y} = \text{cod}(\mathbf{x}) \\ \text{dec} : \mathbb{R}^q &\longrightarrow \mathbb{R}^p, \quad \mathbf{y} \longmapsto \mathbf{x} = \text{dec}(\mathbf{y}) \end{aligned}$$

This problem is ill posed: what is the criteria to be optimized?

The curse of dimensionality

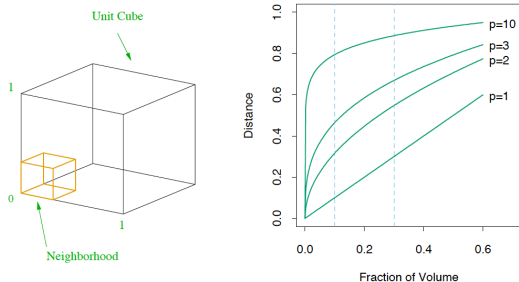


FIGURE 2.6. The curse of dimensionality is well illustrated by a subcubical neighborhood for uniform data in a unit cube. The figure on the right shows the side-length of the subcube needed to capture a fraction r of the volume of the data, for different dimensions p . In ten dimensions we need to cover 80% of the range of each coordinate to capture 10% of the data.

In large dimension, intuitions we have on distances in low dimension (2 or 3) no longer apply.

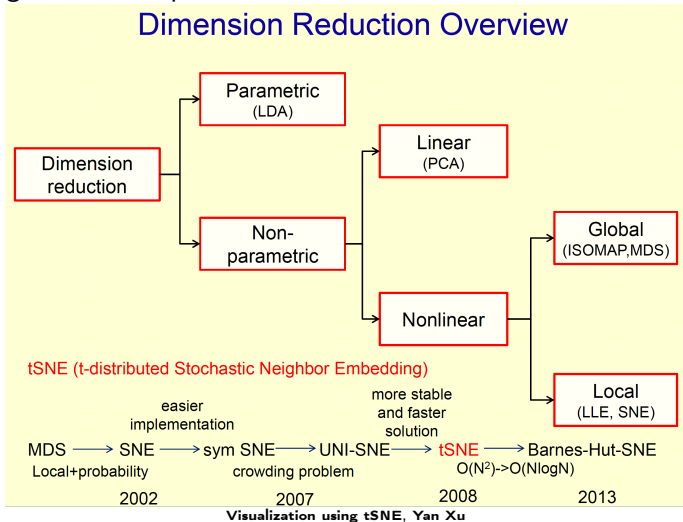
Dimensionality reduction: the big picture

paramétrique non paramétrique

linéaire non linéaire

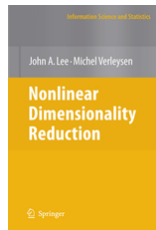
métrique globale métrique locale

Dimension Reduction Overview



Lecture road map

- 1 Introduction to hidden variables
- 2 PCA: principal component analysis
- 3 Distance preservation approaches (global methods)
- 4 Local approaches



PCA: principal component analysis

Model: data = information + noise

$$X = YV^T + B$$

linear coding: $cod : \mathbb{R}^p \longrightarrow \mathbb{R}^q$, $X \longmapsto Y = XV$
 $dec : \mathbb{R}^q \longrightarrow \mathbb{R}^p$, $Y \longmapsto YV^T$

Objective: min. the reconstruction error between X et $dec(cod(X))$

$$\min_{Y \in \mathbb{R}^{n \times q}, V} \|X - YV^T\|_F^2$$

or maximize the variance of the projection

$$\max_{v \in \mathbb{R}^p} \|Xv\|^2 \quad \text{with } \|v\|^2 = 1 \text{ and } y = Xv$$

or minimize the reconstruction error of the covariance (Gram) matrix

$$\min_{y \in \mathbb{R}^n} \|XX^T - yy^T\|^2$$

PCA computation

Theorem (Eckart & Young, 1936)

The unique solution of

$$\min_{\mathbf{u}, \mathbf{v}} J(\mathbf{u}, \mathbf{v}) \quad \text{with} \quad J(\mathbf{u}, \mathbf{v}) = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$$

with $\|\mathbf{v}^*\| = 1$, is given by: \mathbf{v}^* and $\mathbf{u}^* = X \mathbf{v}^*$, where \mathbf{v}^* is the normalized eigen vector associated with λ the **largest eigen value** of $X^\top X$. Furthermore, we have: $\|\mathbf{u}^*\| = \sqrt{\lambda}$.

proof

$$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u}, \mathbf{v}) = -2X\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} = 0 \\ \nabla_{\mathbf{v}} J(\mathbf{u}, \mathbf{v}) = -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} = 0 \end{cases}$$

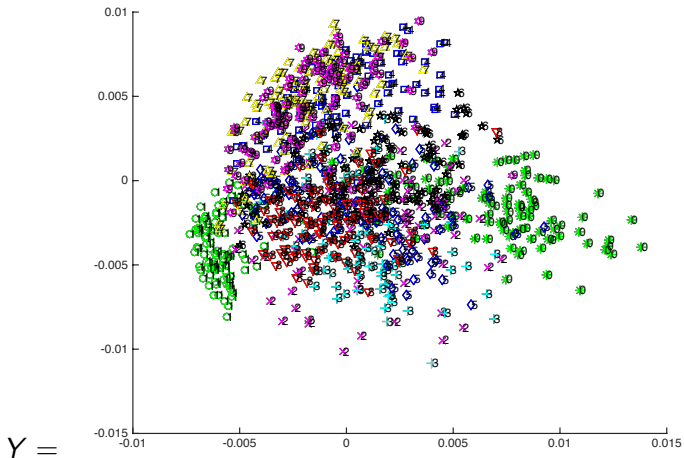
3 different ways to get Y

$$\text{svd}(X), \text{eig}(X^\top X), \text{eig}(XX^\top)$$

2d PCA on the MNIST data

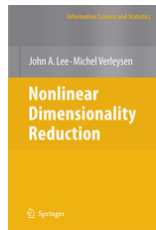
$p = 784$

$q = 2$



Lecture road map

- 1 Introduction to hidden variables
- 2 PCA: principal component analysis
- 3 Distance preservation approaches (global methods)
- 4 Local approaches



Distances

- symmetric : $d(x, y) = d(y, x)$
- separation : $d(x, y) = 0 \Leftrightarrow x = y$
- triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$

Example: the euclidian distance

$$d(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Distance and dot product

$$d(x, y)^2 = \|x\|^2 + \|y\|^2 + 2x^T y$$

hypermetrics and quasi distances

distances and probabilities

Gram matrix

$$G = XX^T$$

Dimensionality reduction: MDS

Given

$$d_X(i, j) = \|x_i - x_j\|$$

Distances conservation

$$\min_{Y \in \mathbb{R}^{n \times q}} \sum_{i=1}^n \sum_{j=1}^{i-1} \left(d_X(i, j)^2 - \underbrace{\|y_i - y_j\|^2}_{d_Y(i, j)^2} \right)^2$$

Related optimization problems (many variants):

- classical MDS (Torgerson, 1958)
- Kruskal -Shepard method (Kruskal, 1964)
- Sammon projection
- MDS INDSCAL (Carroll et Chang, 1970)
- ...

Classical MDS

Let H be the $n \times n$ centering projection matrix

$$H = I - \frac{1}{n}ee^T \quad \text{avec} \quad e = (1, 1, \dots, 1, \dots, 1)^T \in \mathbb{R}^n$$

- 1 given the distance matrix D_X
 - ▶ Y columns are the eigen vectors of HD_XH multiplied by the square root of their corresponding eigen values

$$Y_j = \sqrt{\lambda_i} u_i, \quad i = 1, q$$

- 2 if X_c is known, it is the PCA of the centered data matrix
 - ▶ Y columns are singular values of $X^c = HX$ multiplied by their singular values.

$$Y_j = \mu_i u_i, \quad i = 1, q$$

MDS and PCA

MDS = PCA

- if the data is lying on an hyperplane
→ in that case, distances are preserved

- if X is centered
and if D_X is doubly centered

MDS and PCA

let c be the col mean vector of X ,

$$c = \frac{1}{n} X^T e \quad \text{avec} \quad e = (1, 1, \dots, 1, \dots, 1)^T \in \mathbb{R}^n$$

let X^c be the centered data matrix

$$\begin{aligned} X^c &= X - ec^T \\ &= X - \frac{1}{n} ee^T X = HX \quad \text{with} \quad H = I - \frac{1}{n} ee^T \end{aligned}$$

recall the distance and scalar product formula $d_X(i, j)^2 = \|x_i - x_j\|^2 = \|x_i\|^2 + \|x_j\|^2 - 2x_i^T x_j$

let G be the Gram matrix $G = XX^T$ and $\delta = \text{diag}(G)$ with $\delta_i = \|x_i\|^2$
we have, with D_X the distances matrix of general term $d_X(i, j)^2$

$$D_X = \delta e^T + e \delta^T - 2XX^T$$

and

$$HD_X H = -2X^c X^{cT}$$

Classical MDS aims at minimizing

$$\min_{Y \in \mathbb{R}^{n \times q}} \|HD_X H - HD_Y H\|^2 = \|X^c X^{cT} - Y^c Y^{cT}\|^2$$

Y^c is the eigen matrix of $X^c X^{cT}$ that is the result of the SVD of X^c

Weighted MDS: Sammon's projection

Reinforce closed neighbors (... and penalized distant ones)

$$\min_{Y \in \mathbb{R}^{n \times q}} \sum_{i=1}^n \sum_{j=1}^{i-1} w_{i,j} \left(d_X(i,j) - \underbrace{\|y_i - y_j\|}_{d_Y(i,j)} \right)^2$$

$$w_{i,j} = \frac{1}{\|x_i - x_j\|}$$

Optimization via an iterative descent algorithm (slow) Quasi Newton (L-BFGS).

Sammon's projection Quasi Newton

$$J_S(Y) = \sum_{i=1}^n \sum_{j=1}^n w_{i,j} (d_X(i,j) - d_Y(i,j))^2 \quad \text{with} \quad d_Y(i,j) = \|\mathbf{y}_i - \mathbf{y}_j\|$$

$$\begin{aligned} \nabla_{\mathbf{y}_i} J_S(Y) &= -2 \sum_{j=1, j \neq i}^n w_{i,j} (d_X(i,j) - d_Y(i,j)) \nabla_{\mathbf{y}_i} d_Y(i,j) \\ &= -2 \sum_{j=1, j \neq i}^n \frac{d_X(i,j) - d_Y(i,j)}{d_X(i,j) d_Y(i,j)} (\mathbf{y}_i - \mathbf{y}_j) \end{aligned}$$

2 issues with Sammon's projection

- $d = 0 \rightarrow w = \infty$ because

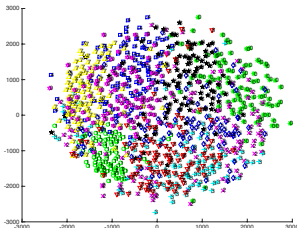
$$w_{i,j} = \frac{1}{\|x_i - x_j\|}$$

the criterion preserve all small distances.

Alternatives:

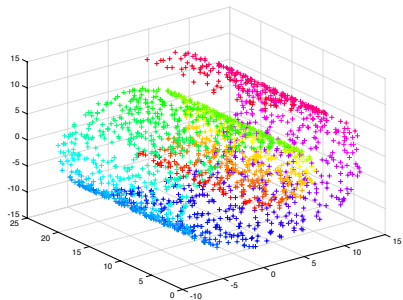
$$w_{i,j} = \begin{cases} 1 & \text{si } \|x_i - x_j\| \leq \varepsilon \\ 0 & \text{sinon} \end{cases}$$

- Y are uniformly distributed in a circle

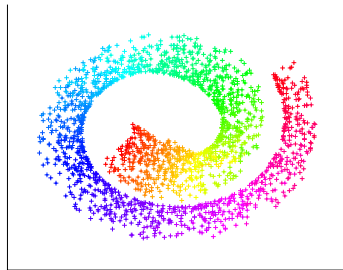


an example of MDS limitation

X



Y

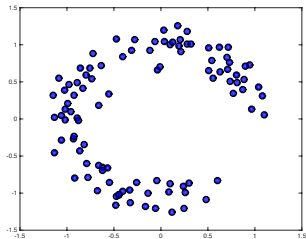


Solution: *metric learning* of $d(i, j)$

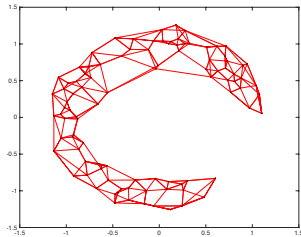
$$d_X(i, j) \text{ euclidean} \quad \rightarrow \quad d_g(i, j) \text{ geodesic}$$

Example of geodesic distance

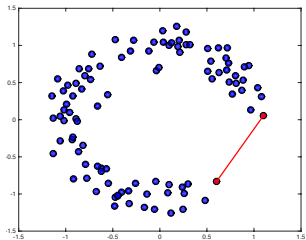
cloud of points



proximity graph

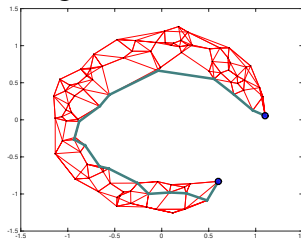


euclidean distance



small $d_X(i, j)$

geodesic distance



large $d_g(i, j)$

Isometric Feature Mapping (ISOMAP)

- 1 build a neighbor graph V
 - ▶ create the graph of the k nearest neighbors for each data point x_i
 - ▶ connect x_i with x_j if $\|x_i - x_j\| \leq \varepsilon$
- 2 find the shortest path in the graph (Dijkstra : $\mathcal{O}(kn^2 \log n)$)

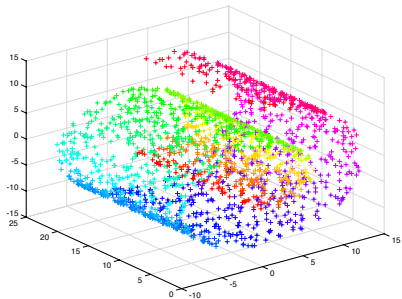
$$d_g(i, j) = \begin{cases} \sum_{\ell=1}^{n_{ij}} d_X(x_{\phi(\ell)}, x_{\phi(\ell+1)}) & \text{with } \phi \text{ the shortest path} \\ \infty & \text{else} \end{cases} \text{ connecting on } V, x_i \text{ to } x_j$$

- 3 compute Y with MDS using d_G instead of d_X ,

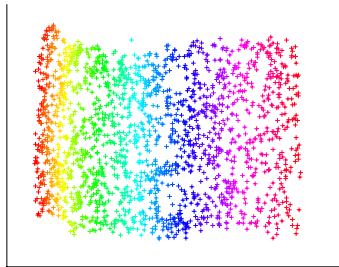
$$\min_{Y \in \mathbb{R}^{n \times q}} \sum_{i=1}^n \sum_{j=1}^{i-1} w_{i,j} (d_g(i, j) - d_Y(i, j))^2$$

Isometric Feature Mapping (ISOMAP)

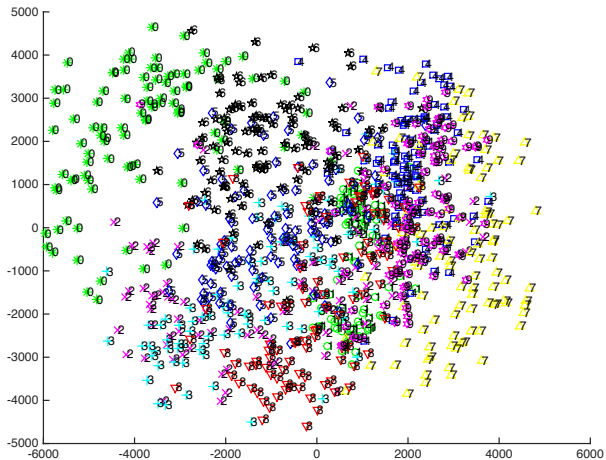
X



Y

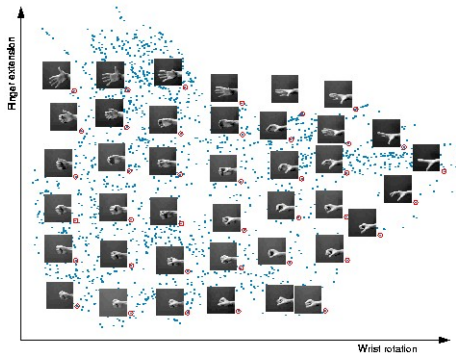


ISOMAP on MNIST



Problem with ISOMAP

- its a global non sparse method
- doesn't scale $\mathcal{O}(n^3)$
- given a new Y the decoding function is not known X .
The pre image problem

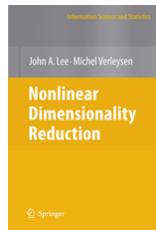


Isomap ($k = 6$) applied to $n = 2000$ images (64 pixels by 64 pixels) of a hand in different configurations. The images were generated by making a series of opening and closing movements of the hand at different wrist orientations, designed to give rise to a two-dimensional manifold.

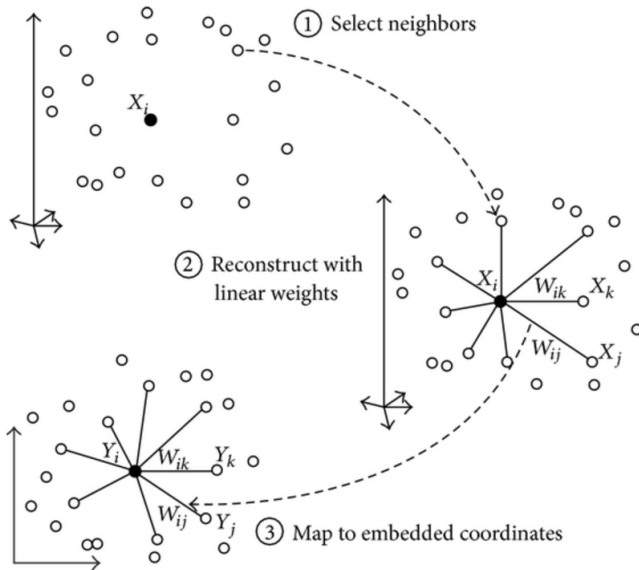
<http://web.mit.edu/cocosci/isomap/handfig.html>

Lecture road map

- 1 Introduction to hidden variables
- 2 PCA: principal component analysis
- 3 Distance preservation approaches (global methods)
- 4 Local approaches



Locally Linear Embedding (LLE)



Locally Linear Embedding (LLE)

Topology conservation: define a local metric

- 1 $V_{i,j} = 0$ if i and j are not among the $k < p$ nearest neighbors
- 2 compute the non zero weight: only for $V_{i,j} \neq 0$

$$\min_{V \in \mathbb{R}^{n \times n}} \sum_{i=1}^n \|x_i - \sum_{j=1}^n v_{i,j} x_j\|^2 \quad \text{avec} \quad \sum_{j=1}^n v_{i,j} = 1, \quad i = 1 : n$$

→ n least square problems

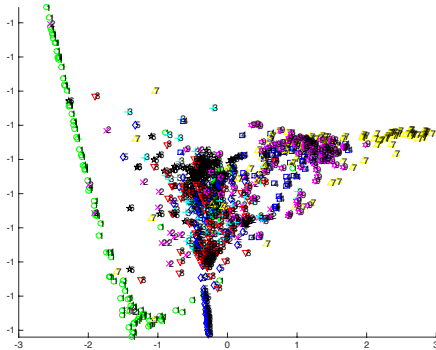
- 3 how to get Y

$$\min_{Y \in \mathbb{R}^{n \times q}} \sum_{i=1}^n \|y_i - \sum_{j=1}^n v_{i,j} y_j\|^2 = \|Y - VY\|_F^2$$

→ SVD($I - V$), the 2 smallest non zero singular values.

Problems with LLE

- Most of the data points concentrate at the center
- A few points are far from the center to satisfy the unit variance constraint.



Stochastic Neighbor Embedding (SNE)

Model the conditional probability of a point x conditionally to our position in x_i

$$\mathbb{P}_X(x|x_i) = \frac{1}{Z_x} \exp \frac{-\|x-x_i\|^2}{2\sigma_i^2} \quad \mathbb{P}_Y(y|y_i) = \frac{1}{Z_y} \exp^{-\|y-y_i\|^2}$$

- 1 tune σ_i so that each point x_i have k neighbors
 - ▶ or to have the same perplexity p at each point

$$p = \text{entropy}(P_i) = \log(k)$$

- 2 Minimize the Kullback-Leibler divergence between both distributions

$$\min_Y \sum_{i=1}^n KL(\mathbb{P}_X(i)||\mathbb{P}_Y(i)) = \sum_{i=1}^n \sum_{j=1}^n \mathbb{P}_X(j|i) \log \frac{\mathbb{P}_X(j|i)}{\mathbb{P}_Y(j|i)}$$

SNE optimization

Symmetric case: build \mathbb{P}_X so that $\mathbb{P}_X(j|i) = \mathbb{P}_X(i|j)$

$$\frac{\mathbb{P}_X(j|i) + \mathbb{P}_X(i|j)}{2}$$

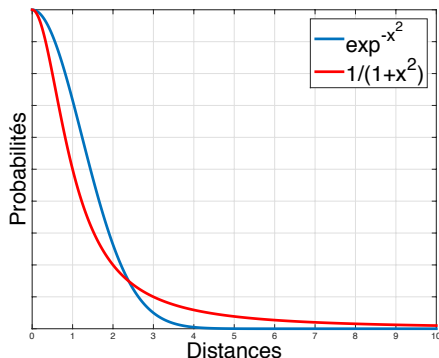
in that case:

$$\nabla_{Y(i)} KL(\mathbb{P}_X || \mathbb{P}_Y) = 2 \sum_{j=1}^n \underbrace{(\mathbf{y}_i - \mathbf{y}_j)}_{\text{similarity}} \underbrace{(\mathbb{P}_X(j|i) - \mathbb{P}_Y(j|i))}_{\text{rigidity}}$$

Possible acceleration thanks to the Barnes-Hut-SNE $\mathcal{O}(n \log n)$

t-Stochastic Neighbor Embedding (t-SNE)

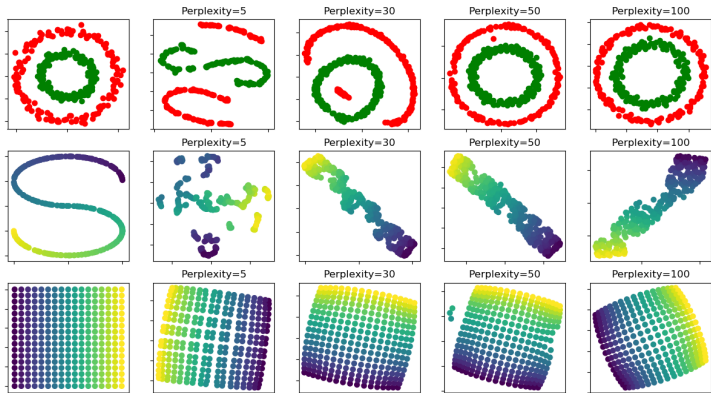
$$\mathbb{P}_Y(y_j|y_i) = \frac{1}{Z} \frac{1}{1 + \|y - y_i\|^2}$$



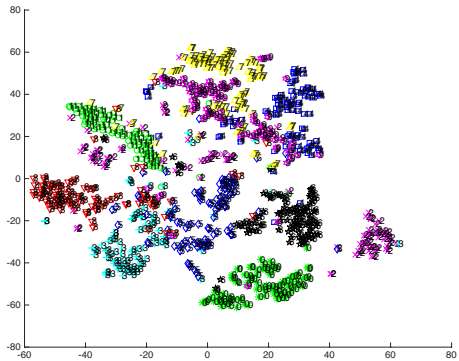
- $\mathbb{P}_X = \mathbb{P}_Y$ large $\Rightarrow d_Y < d_X$ (attraction)
- $\mathbb{P}_X = \mathbb{P}_Y$ small $\Rightarrow d_Y > d_X$ (repulsion)

t-SNE: influence of the perplexity

"The perplexity can be interpreted as a smooth measure of the effective number of neighbors"

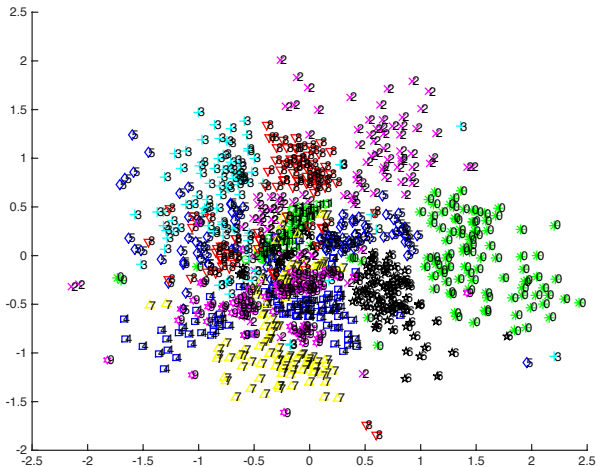


t-SNE on MNIST



<https://lvdmaaten.github.io/tsne/>

Multi-scale similarities in SNE




John A. Lee, Diego H. Peluffo, Michel Verleysen Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure *Neurocomputing* 2015, 169:246-261.

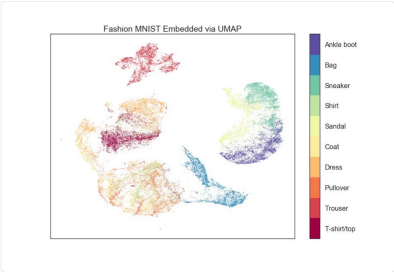
<http://dx.doi.org/10.1016/j.neucom.2014.12.095>

Uniform manifold approx. & projection (UMAP)

tweets **524** following **42** followers **566** likes **673**

 **Leland McInnes** @leland_mcinnes · 10 Nov 2017

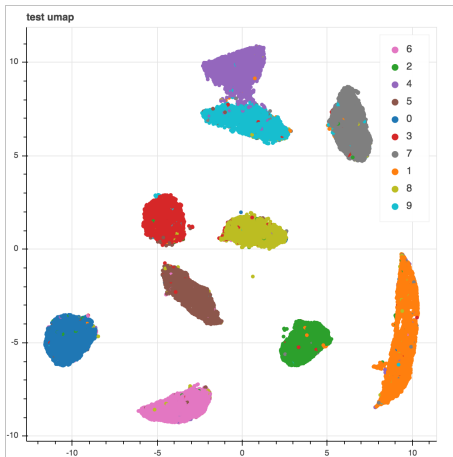
The new numba based version of UMAP is out. Now faster than ever, it takes only 2.5 minutes to embed the full 70000 points of the 784-dimensional "Fashion MNIST" dataset. github.com/lmcinnes/umap



The figure is a scatter plot titled "Fashion MNIST Embedded via UMAP". It shows a 2D visualization of 70,000 data points from the Fashion MNIST dataset, where each point is colored according to its class. The classes are: Ankle boot (dark purple), Bag (blue), Sneaker (teal), Shirt (light green), Sandal (yellow), Coat (orange), Dress (red-orange), Pullover (red), Trouser (dark red), and T-shirt/top (dark purple). The plot shows distinct clusters for each class, indicating that UMAP has successfully separated the different categories in the high-dimensional space. A legend on the right side of the plot lists the classes with their corresponding colors.

11 199 513

Uniform manifold approx. & projection (UMAP)



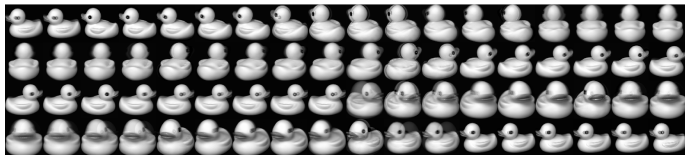
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction Leland McInnes, John Healy (Submitted on 9 Feb 2018)

<https://github.com/lmcinnes/umap>

COIL 20

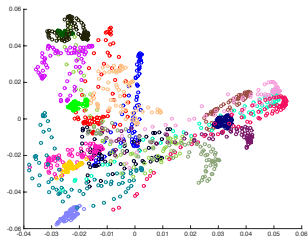


- Columbia University Image Library
- $n = 20 \times 72 = 1440$ images
- $p = 128 \times 128 = 16384$ pixels
- images for all of the objects in which the background has been discarded

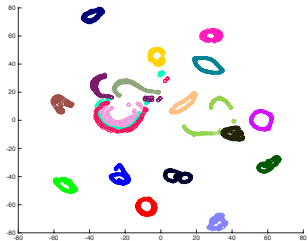
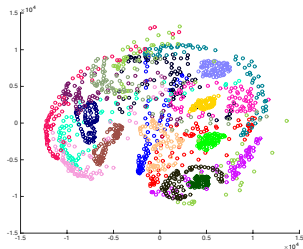


COIL 20

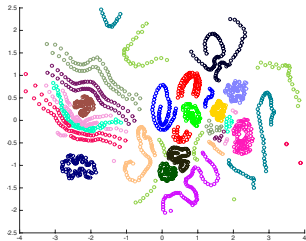
PCA



Sammon's projection (MDS)



t-SNE



Multi-scale similarities SNE

Conclusions

- 1 And the winner is UMAP ... and t-SNE
- 2 other approaches...
 - ▶ Self organizing feature maps - SOM (Kohonen, 1974)
 - ▶ Curvilinear component analysis (CCA, Demartines & Héroult, 1995)
 - ▶ Kernel PCA
 - ▶ Laplacian Eigenmaps
 - ▶ Curvilinear distance analysis (CDA, Lee et al. 2004)
 - ▶ Semidefinite Embedding (SDE, Weinberger and Saul 2004)
 - ▶ Réseaux de neurones de type Autoencoder.
 - ▶ Maximum Variance Unfolding (MVU)
 - ▶ Weighted t-SNE research.cs.aalto.fi/pml/software/ne/
 - ▶ Metric learning
 - ▶ ...
- 3 to play with: doc.gold.ac.uk/~lfedd001/three/demo.html