One the Role and Impact of the Metaparameters in t-distributed Stochastic Neighbor Embedding

John A. Lee and Michel Verleysen
Machine Learning Group
Université catholique de Louvain
Louvain-la-Neuve, Belgium
michel.verleysen@uclouvain.be

Motivation for nonlinear dimensionality reduction

- High-dimensional data are
  - difficult to represent
  - difficult to understand
  - difficult to analyze

- Motivation #1:
  - To visualize data living in a d-dimensional space ($d > 3$)

- Motivation #2:
  - Models (regression, classification, clustering) based on high-dimensional data suffer from the curse of dimensionality
  - Need to reduce the dimension of data while keeping information content!

Visualization

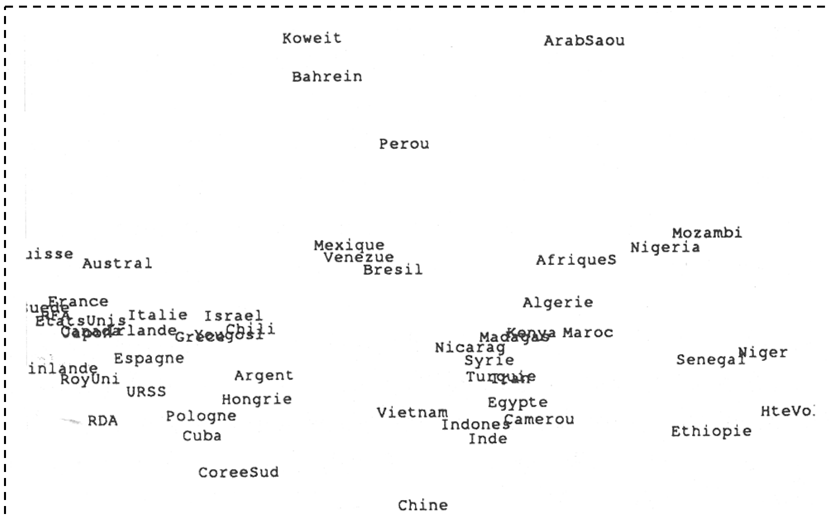
- These are data
- It is difficult to see something…

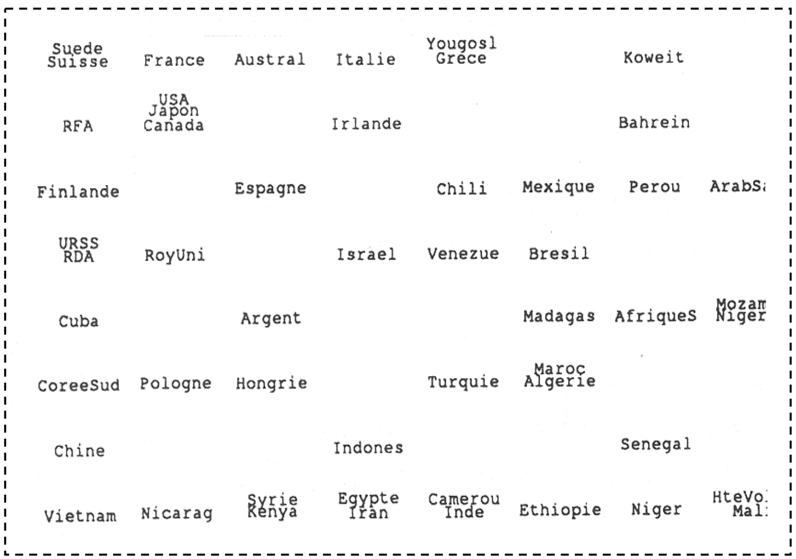
annual increase (%), infant mortality (‰), illiteracy ratio (%), school attendance (%), GIP, annual GIP increase (%)

Afrique du sud	2.9	89.0	50.0	19.0	2680.0	-2.9
Algerie	2.9	114.0	58.5	47.9	2266.0	0.1
Arabie Saoudite	4.2	111.0	75.4	39.7	10827.0	-10.8
Argentine	1.2	44.0	5.3	69.5	2264.0	2.0
Australie	1.3	10.4	0.0	86.0	9938.0	-1.2
Bahrein	3.8	57.0	20.9	76.3	8960.0	-10.1
Bresil	2.2	75.0	23.9	62.3	1853.0	-3.9
Cameroun	2.4	106.0	55.1	44.5	939.0	6.5
Canada	1.0	10.0	0.9	93.0	9857.0	3.0
Chili	1.7	42.0	7.7	85.2	1853.0	-0.5
Chine	1.4	71.0	31.0	44.0	231.0	10.0
Coree du Sud	1.6	33.0	8.3	82.1	1716.0	9.3
Cuba	0.7	16.8	8.9	78.7	2046.0	5.2
Egypte	2.7	74.0	58.1	45.8	626.0	6.0
Espagne	0.9	9.6	6.8	88.0	5316.0	2.3
Etats Unis	1.0	11.2	0.8	91.0	11732.0	3.3
Ethiopie	2.7	145.0	85.0	23.1	140.0	7.4
Finlande	0.6	6.5	0.6	98.0	10286.0	5.1
France	0.4	9.1	1.2	86.0	11326.0	0.5
Grece	1.1	15.1	11.7	81.0	4060.0	0.3
Haute Volta	1.7	208.0	88.6	7.6	240.0	3.6
Hongrie	0.0	20.0	0.9	42.0	1963.0	0.9
Inde	1.8	121.0	57.6	71.7	260.0	6.5
Indonesie	1.7	99.0	32.3	41.3	488.0	5.0
Iran	2.7	105.0	57.2	57.9	2346.0	5.2
Irlande	1.2	11.0	1.0	93.0	4813.0	0.5
Israel	2.2	15.0	6.7	74.0	4531.0	1.1
Italie	0.4	13.0	4.6	73.0	6869.0	-1.2
Japon	0.9	6.6	0.8	92.0	9704.0	3.0
Kenya	4.0	85.0	52.9	59.3	376.0	3.6
Kowait	6.5	33.0	35.9	73.0	20900.0	-0.5
Madagascar	2.7	69.0	38.8	30.4	259.0	0.9
Maroc	2.5	104.0	65.0	34.9	864.0	0.6
Mali	2.8	152.0	86.5	16.7	190.0	1.5
Mexique	2.6	54.0	17.3	70.1	1900.0	-4.6
Mozambique	2.7	150.0	66.8	16.1	155.0	-6.9
Nicaragua	4.4	88.0	10.0	52.5	760.0	5.1
Niger	3.0	143.0	90.2	9.2	330.0	2.5
Nigeria	3.3	133.0	66.0	29.3	807.0	-4.0
Perou	2.8	85.0	19.3	72.0	997.0	-12.0
Pologne	0.9	24.6	0.6	77.0	2545.0	4.5
RDA	-0.2	11.4	0.5	89.0	5103.0	4.2
RFA	-0.1	12.0	0.7	87.0	12176.0	1.0
Royaume Uni	-0.1	10.1	0.8	83.0	8655.0	3.5
Sénégal	2.6	152.0	77.5	19.2	430.0	2.3
Suède	0.1	7.0	0.6	85.0	13920.0	1.8
Suisse	0.6	8.0	0.9	88.0	15522.0	-0.1
Syrie	3.8	60.0	46.3	50.7	1717.0	5.8
Turquie	2.1	119.0	31.2	42.0	1491.0	3.0
URSS	0.9	28.8	0.8	96.0	4562.0	4.0
Venezuela	3.0	40.0	19.0	57.7	3823.0	-2.0
Vietnam	2.3	97.0	13.0	59.5	220.0	5.2
Yougoslavie	0.9	31.0	13.2	83.0	2067.0	-1.3

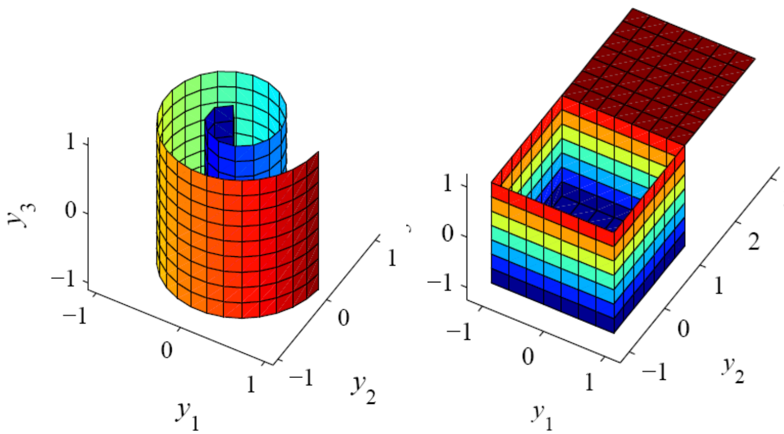
Visualization

- These are the same data
- under different visualization paradigms
- possible to see groups, relations, outliers, ...

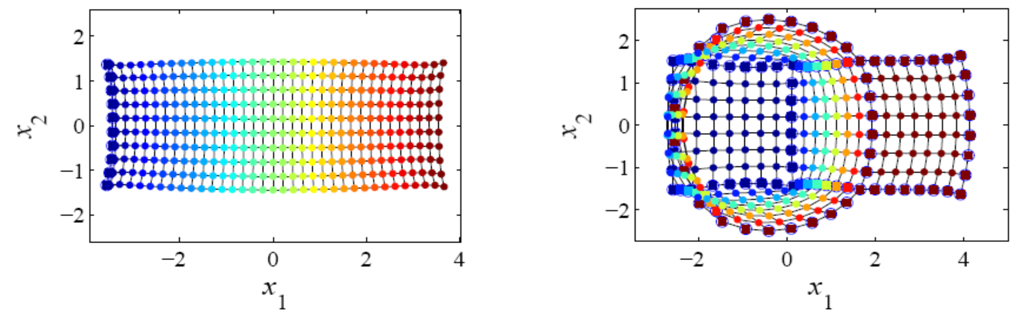




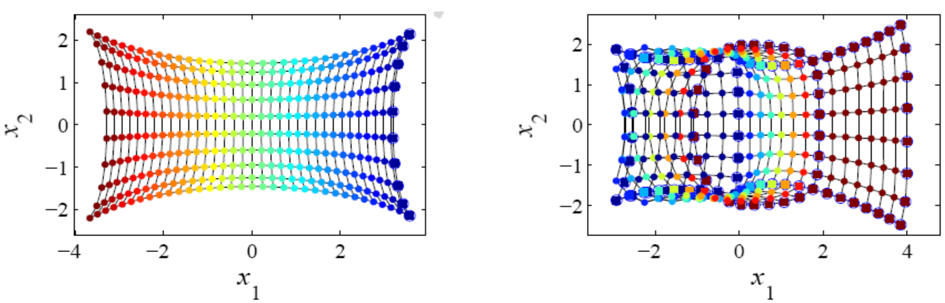
Not all NLDR methods perform equally !



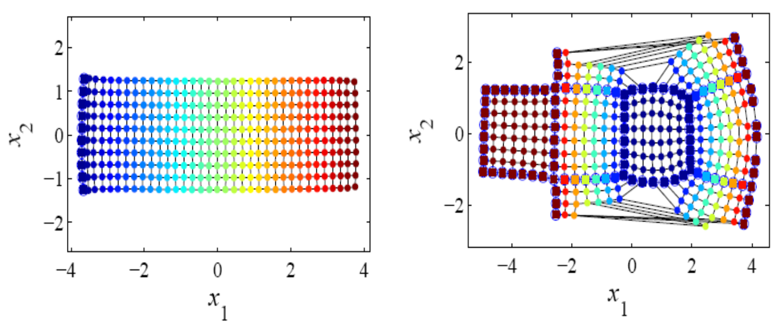
Geodesic NLM



Isomap

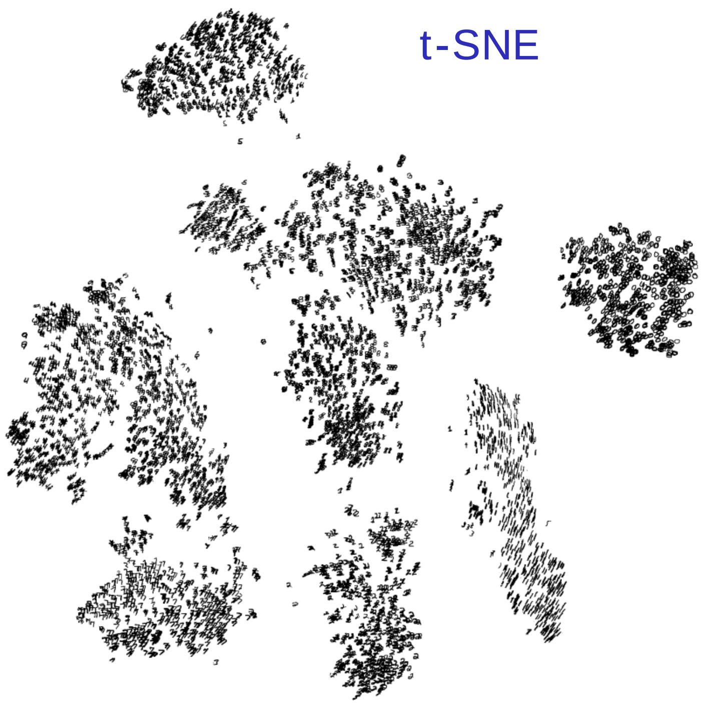


CDA

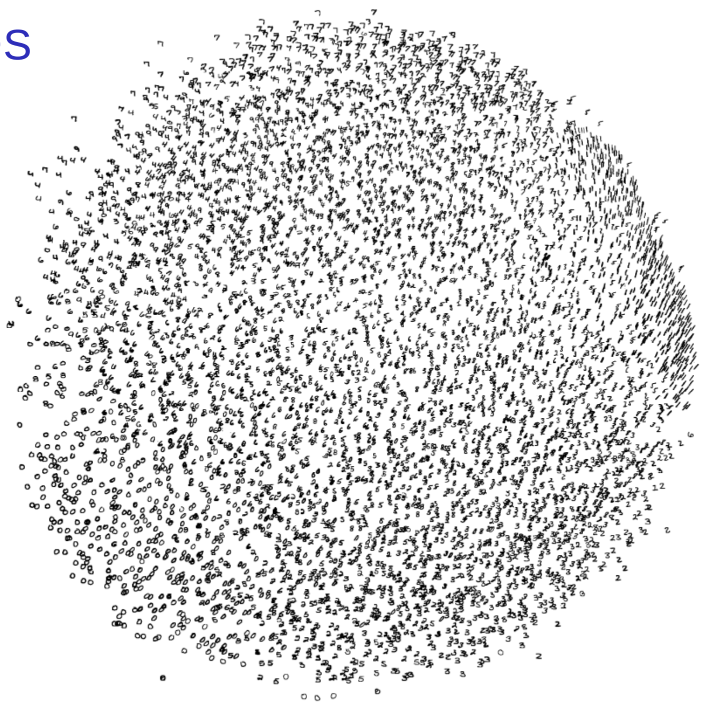


Stochastic Neighbor Embedding

- SNE and t-SNE are nowadays considered as ‘good’ methods for NDLR
- Examples



t-SNE

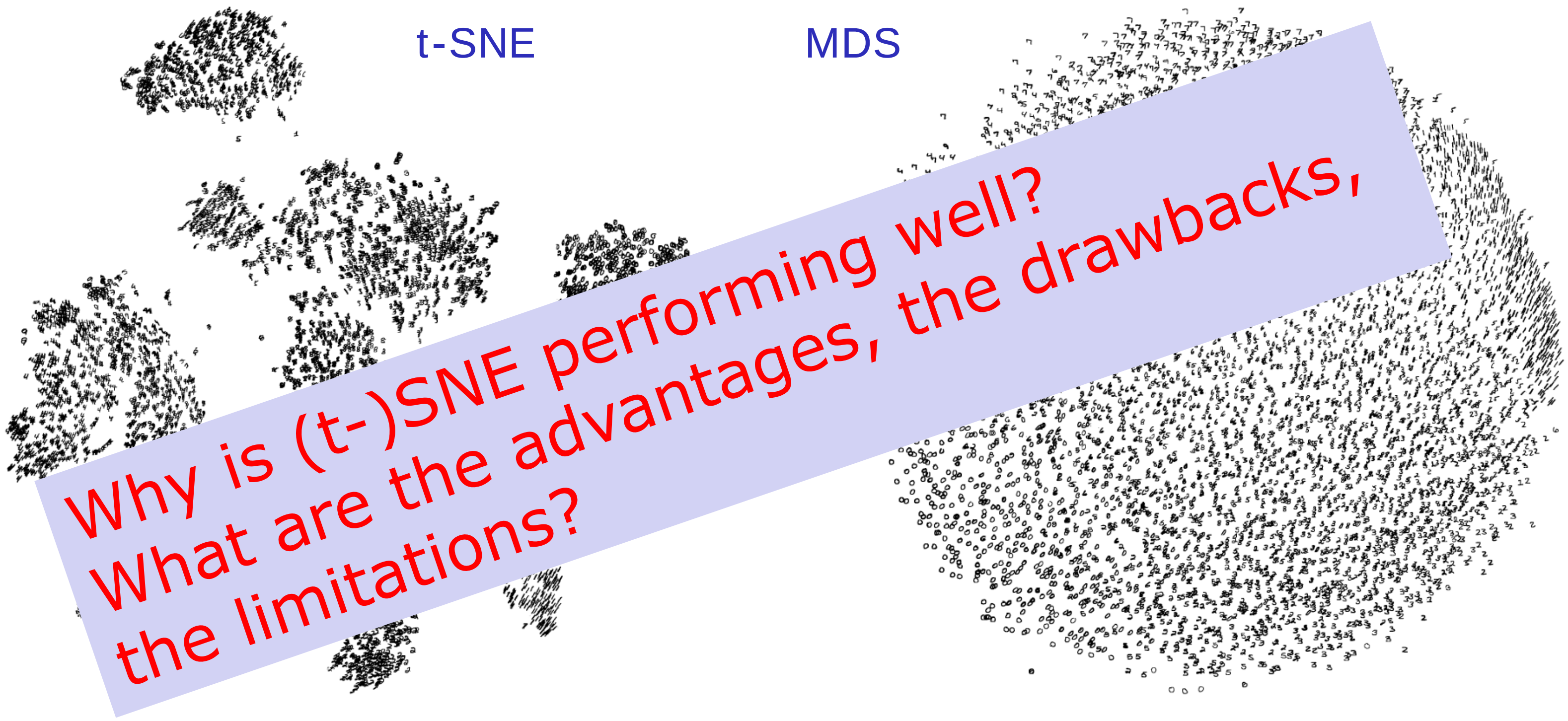


MDS

From: L. Van der Maaten & G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579-2605

Stochastic Neighbor Embedding

- SNE and t-SNE are nowadays considered as ‘good’ methods for NDLR
- Examples



From: L. Van der Maaten & G. Hinton, Visualizing Data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579-2605

Outline

- NDLR: a historical perspective
  - stress function
  - intrusion and extrusions
  - geodesic distances
- SNE and t-SNE
  - algorithm
  - gradient
  - transformed distances
- Experiments
  - with Euclidean distances
  - with geodesic distances
- Conclusions

From MDS to more general cost functions

- MDS follows the idea of

$$\min_X \sum_{i<j} \left(\delta_{ij}^2 - d_{ij}^2\right)^2 \qquad \text{where} \quad \begin{aligned} \delta_{ij} &= \left\| y_i - y_j \right\| \\ d_{ij} &= \left\| x_i - x_j \right\| \end{aligned}$$

- Extension:

$$\min_X \sum_{i<j} w_{ij} \left(\delta_{ij}^2 - d_{ij}^2\right)^2$$

to give more importance to
  - small distances
  - close data
  - ...

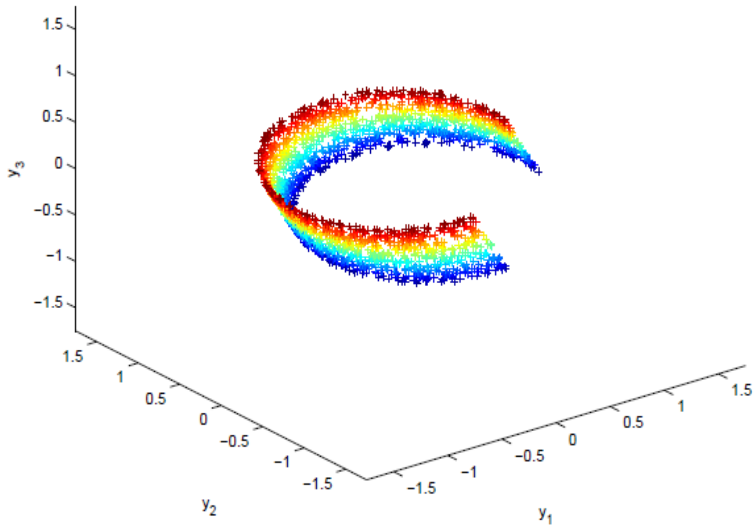
Breakthrough #1

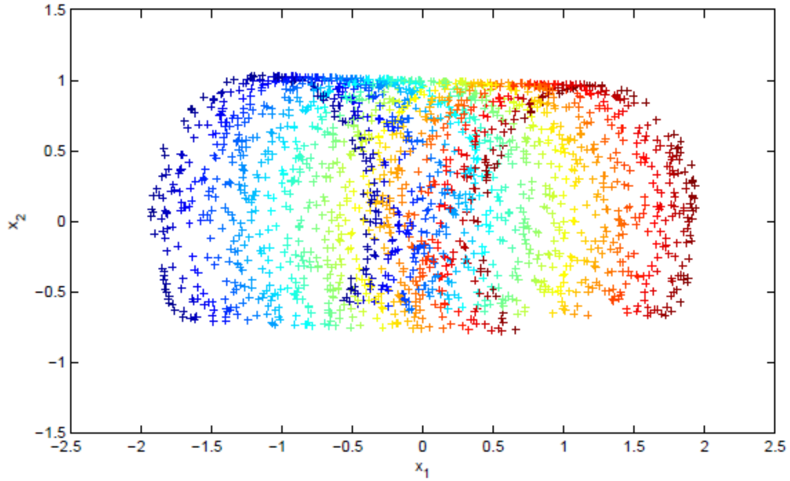
Traditional « stress » function:

$$\min_X \sum_{i<j} w_{ij} \left(\delta_{ij} - d_{ij}\right)^2$$

Limitations of linear projections

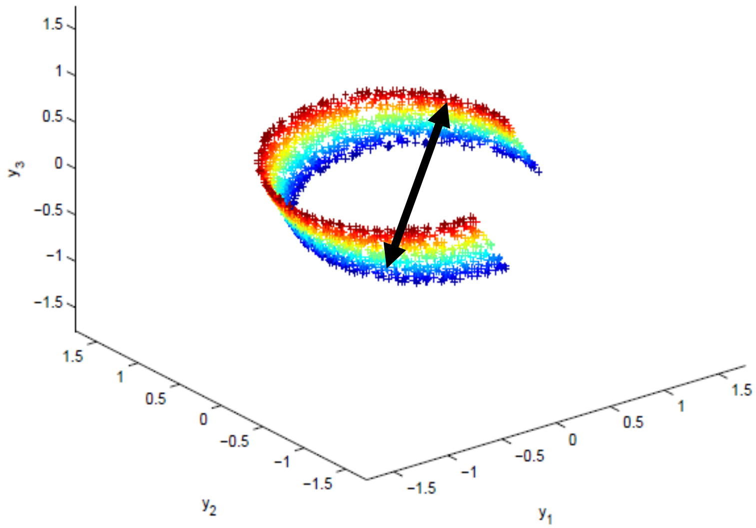
- Even *simple* manifolds can be poorly projected

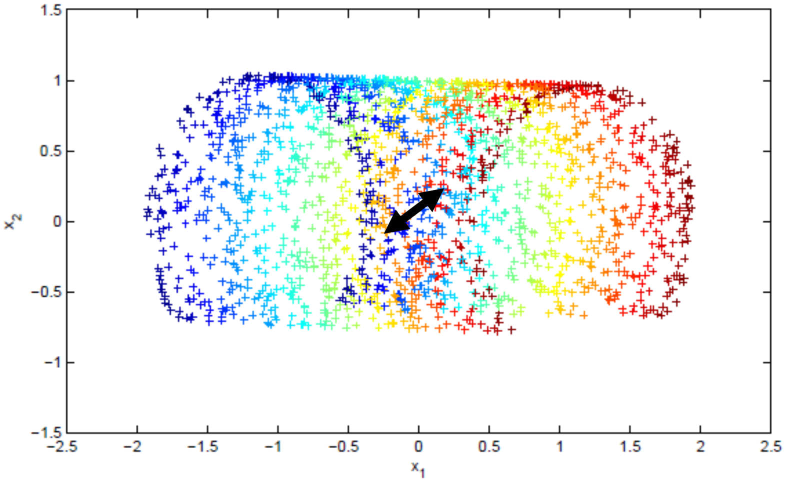




Limitations of linear projections

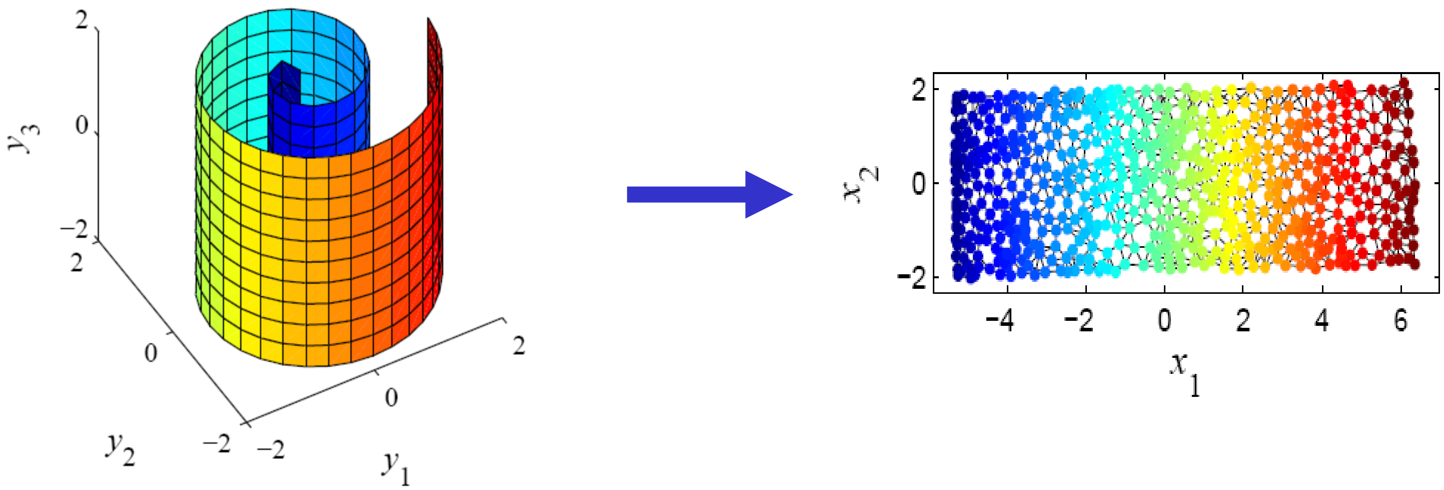
- Even *simple* manifolds can be poorly projected
- Points originally far from eachother are projected close: this is an intrusion





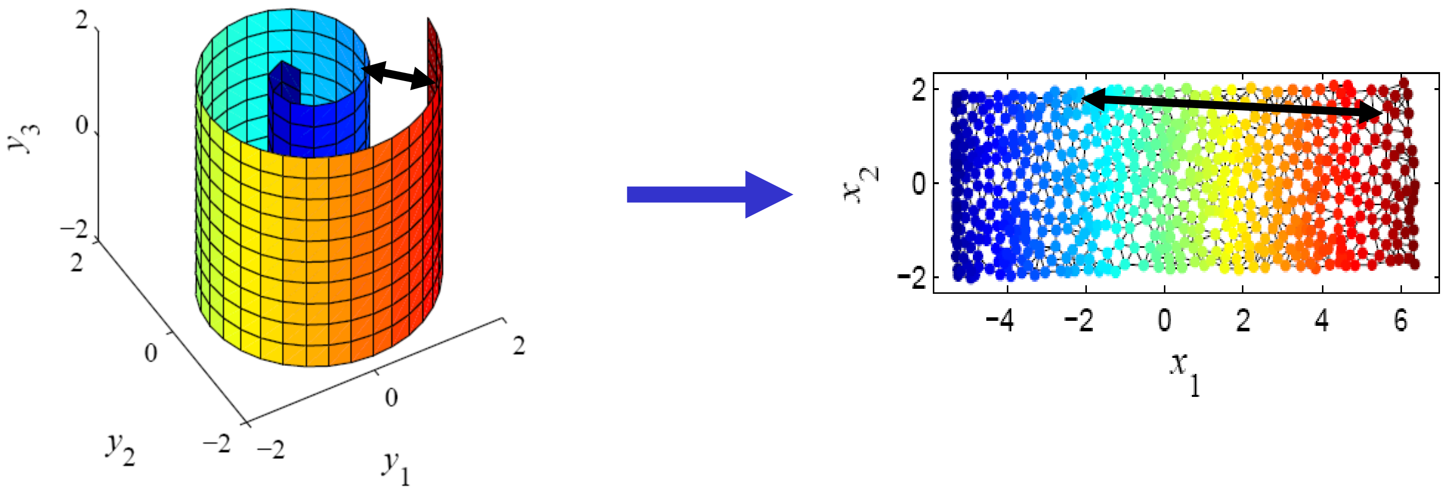
Nonlinear projections

- Goal: to unfold, rather than to project (linearly)



Nonlinear projections

- Goal: to unfold, rather than to project (linearly)
- Intrusions can be hopefully decreased, but extrusions could appear



The user's point of view

- Favouring intrusions or extrusions is related to the application (user's point of view)

- General way of handling the compromise:

$$w_{ij} = \underbrace{\lambda f\left(\frac{d_{ij}}{\sigma}\right)}_{\text{allows intrusions}} + \underbrace{(1-\lambda) f\left(\frac{\delta_{ij}}{\sigma}\right)}_{\text{allows extrusions}}$$

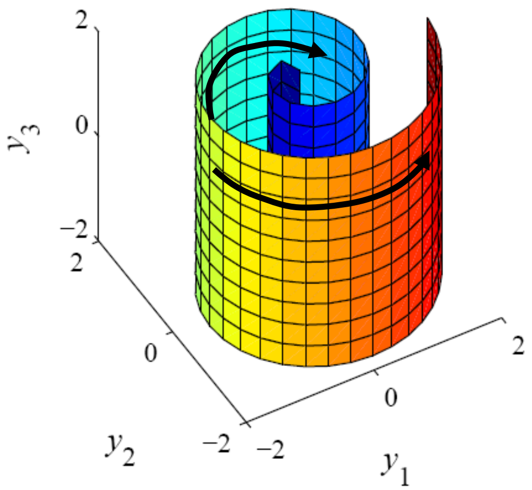
Breakthrough #2

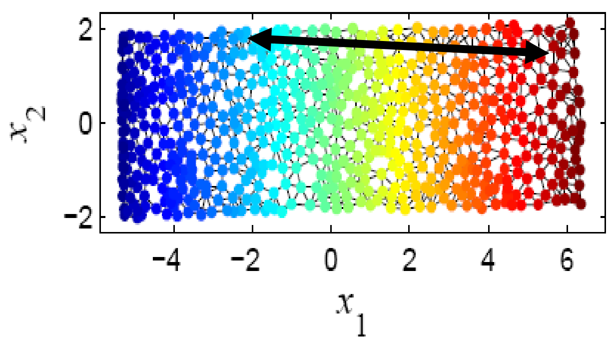
- Nowadays, few methods acknowledge this need for a trade-off !

Geodesic distances

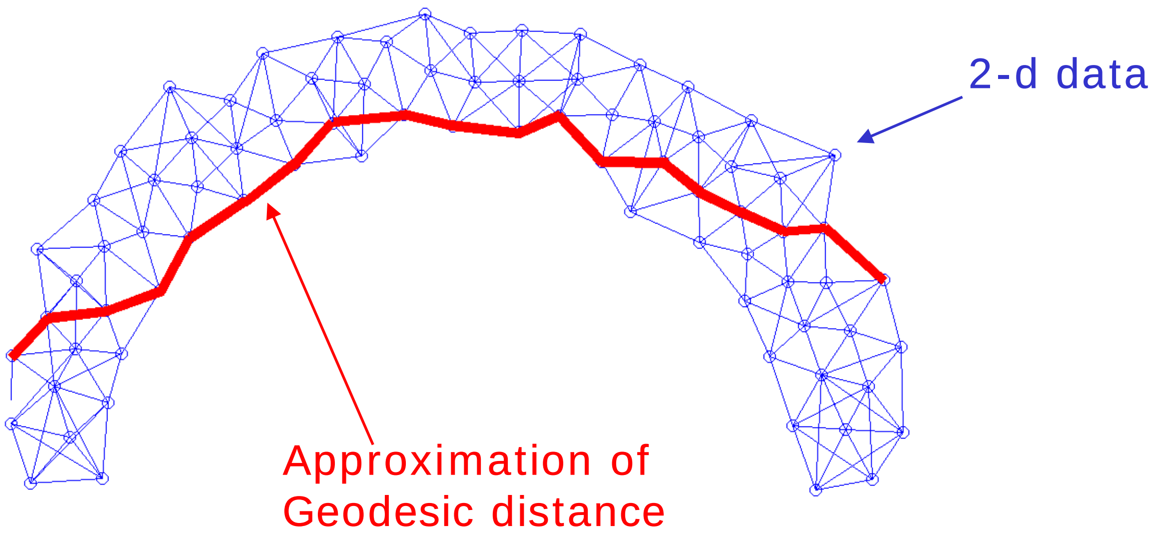
- Goal: to measure distances along the manifold
- Such distances are more easily preserved

Breakthrough #3





Geodesic and graph distances



- Geodesic distances: finding the shortest way between data along the manifold
- Problem: the manifold is unknown → approximate it by a graph
- It exists efficient algorithms for finding shortest paths
- The graph can be built by connecting data in a k-neighborhood, or in a ε-ball

Distance preservation methods

		Euclidean distances in HD space	Geodesic distances in HD space
$E = \sum_{i,j=1}^{N} \left(d_y(i,j) - d_x(i,j)\right)^2$		Metric MDS	Isomap
$E_{NLM} = \sum_{\substack{i=1 \\ i<j}}^{N} \frac{\left(d_y(i,j) - d_x(i,j)\right)^2}{d_y(i,j)}$	Favors intrusions	Sammon NLM	Geodesic NLM
$E_{CCA} = \sum_{\substack{i=1 \\ i<j}}^{N} \left(d_y(i,j) - d_x(i,j)\right)^2 F_\lambda\left(d_x(i,j)\right)$	Favors extrusions	CCA	CDA

Distance preservation methods

		Euclidean distances in HD space	**Geodesic distances in HD space**
$E = \sum_{i,j=1}^{N} \left(d_y(i,j) - d_x(i,j)\right)^2$		Metric MDS	Isomap
$E_{NLM} = \sum_{\substack{i=1 \\ i<j}}^{N} \frac{\left(d_y(i,j) - d_x(i,j)\right)^2}{d_y(i,j)}$	Favors intrusions	Sammon NLM	Geodesic NLM
$E_{CCA} = \sum_{\substack{i=1 \\ i<j}}^{N} \left(d_y(i,j) - d_x(i,j)\right)^2 F_\lambda(d_x(i,j))$	Fav… extrusions	CCA	CDA

Computational load ↓
Performances ↓

Computational load ↑
Performances ↑

Outline

- NDLR: a historical perspective
  - stress function
  - intrusion and extrusions
  - geodesic distances
- SNE and t-SNE
  - algorithm
  - gradient
  - transformed distances
- Experiments
  - with Euclidean distances
  - with geodesic distances
- Conclusions

SNE and t-SNE

- In the original space, the similarity between y_i and y_j is defined as

$$p_{j|i}(\lambda_i) = \begin{cases} 0 & \text{if } i = j \\ \dfrac{g(\delta_{ij}/\lambda_i)}{\sum_{k \neq i} g(\delta_{ik}/\lambda_i)} & \text{otherwise} \end{cases} \qquad \left(g(u) = \exp\left(\frac{-u^2}{2}\right) \right)$$

- Similarities are not symmetric (individual widths) !
- $p_{j|i}$ is the empirical probability of y_j to be a neighbor of y_i

SNE and t-SNE

- In the original space, the similarity between y_i and y_j is defined as

$$p_{j|i}(\lambda_i) = \begin{cases} 0 & \text{if } i = j \\ \dfrac{g(\delta_{ij}/\lambda_i)}{\sum_{k \neq i} g(\delta_{ik}/\lambda_i)} & \text{otherwise} \end{cases} \qquad \left(g(u) = \exp\left(\frac{-u^2}{2} \right) \right)$$

- Similarities are not symmetric (individual widths) !
- $p_{j|i}$ is the empirical probability of y_j to be a neighbor of y_i

- Individuals widths λ_i: set (individually) through a global « perplexity » parameter

$$2^{H(p_{j|i})} = PPXT$$

SNE and t-SNE

- In the embedding space, the similarity between x_i and x_j is defined as

$$q_{ij}(n) = \begin{cases} 0 & \text{if } i = j \\ \dfrac{t(d_{ij}, n)}{\sum_{k \neq l} t(d_{kl}, n)} & \text{otherwise} \end{cases} \qquad \left(t(u, n) = \left(1 + \frac{u^2}{n}\right)^{-\frac{n+1}{2}} \right)$$

- Similarities are symmetric
- $t(u,n)$ is proportional to a Student t with n degrees of freedom (n controls the thickness of the tail)
- SNE: $n \to \infty$ *t-SNE:* $n = 1$

SNE and t-SNE

- Now that similarties are defined in both spaces, how to compare them?

$$E = D_{\mathrm{KL}}\left(p \| q\right)$$

  - This seems to be a major difference with respect to other methods, based on square erros!

- E is minimized by gradient descent, to find locations x_i.

$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} \left(x_i - x_j\right)$$

SNE and t-SNE

- Now that similarties are defined in both spaces, how to compare them?

$$E = D_{\mathrm{KL}}\left(p \| q\right)$$

  - This seems to be a major difference with respect to other methods, based on square erros!
- E is minimized by gradient descent, to find locations x_i.

$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} \left(x_i - x_j\right)$$

x_i moves towards x_j

SNE and t-SNE

- Now that similarties are defined in both spaces, how to compare them?

$$E = D_{\mathrm{KL}}\left(p \| q\right)$$

  - This seems to be a major difference with respect to other methods, based on square erros!
- E is minimized by gradient descent, to find locations x_i.

$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} \left(x_i - x_j\right)$$

x_i moves towards x_j

Similarity error – adjusts amplitude

SNE and t-SNE: gradient

- Now that similarities are defined in both spaces, how to compare them?

$$E = D_{\mathrm{KL}}\left(p \,\|\, q\right)$$

  – This seems to be a major difference with respect to other methods, based on square erros!

- E is minimized by gradient descent, to find locations x_i.

$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} \left(x_i - x_j\right)$$

x_i moves towards x_j

Similarity error – adjusts amplitude

Damping factor

SNE and t-SNE: gradient

$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} \left(x_i - x_j\right)$$

x_i moves towards x_j

Similarity error – adjusts amplitude

Damping factor

- Damping factor is similar to $F_\lambda(d_{ij})$ in CCA and CDA:
  - Large distances are less important
  - Distances in the embedding space are used, to allow tears (favoring extrusions)

SNE and t-SNE: distributions

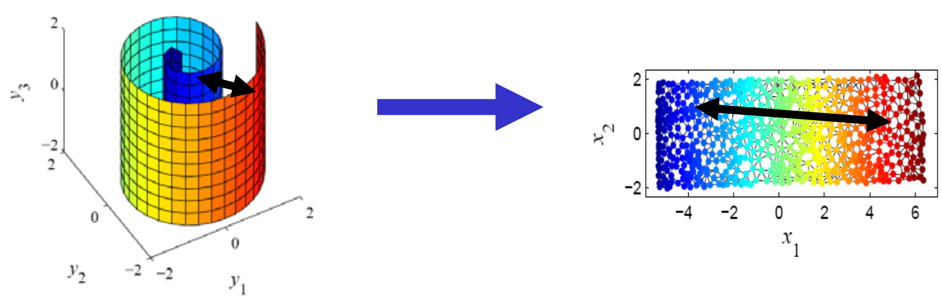
$$\frac{\partial E}{\partial x_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\lambda) - q_{ij}(n)}{1 + d_{ij}^2 / n} (x_i - x_j)$$

x_i moves towards x_j

Similarity error – adjusts amplitude

Damping factor

- Why different distributions for p_{ij} and q_{ij} ?
- Remember that distances have often to be *enlarged*: heavier tails (in the embedding space) help!



SNE and t-SNE: distributions

- Non-trivial solution of min E
- After some (rough) approximations:

$$d_{ij} \approx f(\delta_{ij}) = \sqrt{n \exp\left(\frac{\delta_{ij}^2}{(n+1)\lambda_i^2}\right) - n}$$

- Properties
  - f is monotonically increasing
  - with SNE ($n \rightarrow \infty$): $f(\delta_{ij}) = \delta_{ij} / \lambda_i$
  - if $\delta_{ij} << \lambda_i$, then

$$f(\delta_{ij}) = \delta_{ij} / \left(\lambda_i \sqrt{n+1}\right)$$

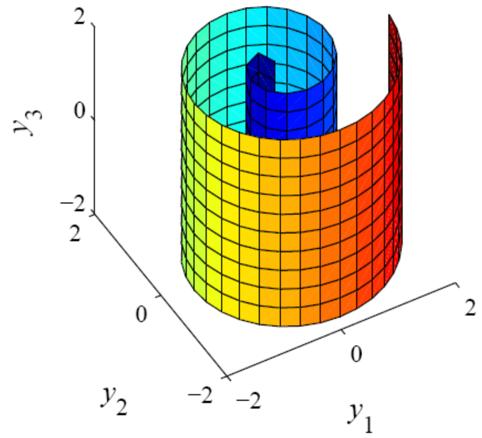
- t-SNE tries to preserved *streched* distances
- SNE distances are scaled by λ_i
- n and λ_i act more or less in the same way

Outline

- NDLR: a historical perspective
  - stress function
  - intrusion and extrusions
  - geodesic distances
- SNE and t-SNE
  - algorithm
  - gradient
  - transformed distances
- Experiments
  - with Euclidean distances
  - with geodesic distances
- Conclusions

Experiments

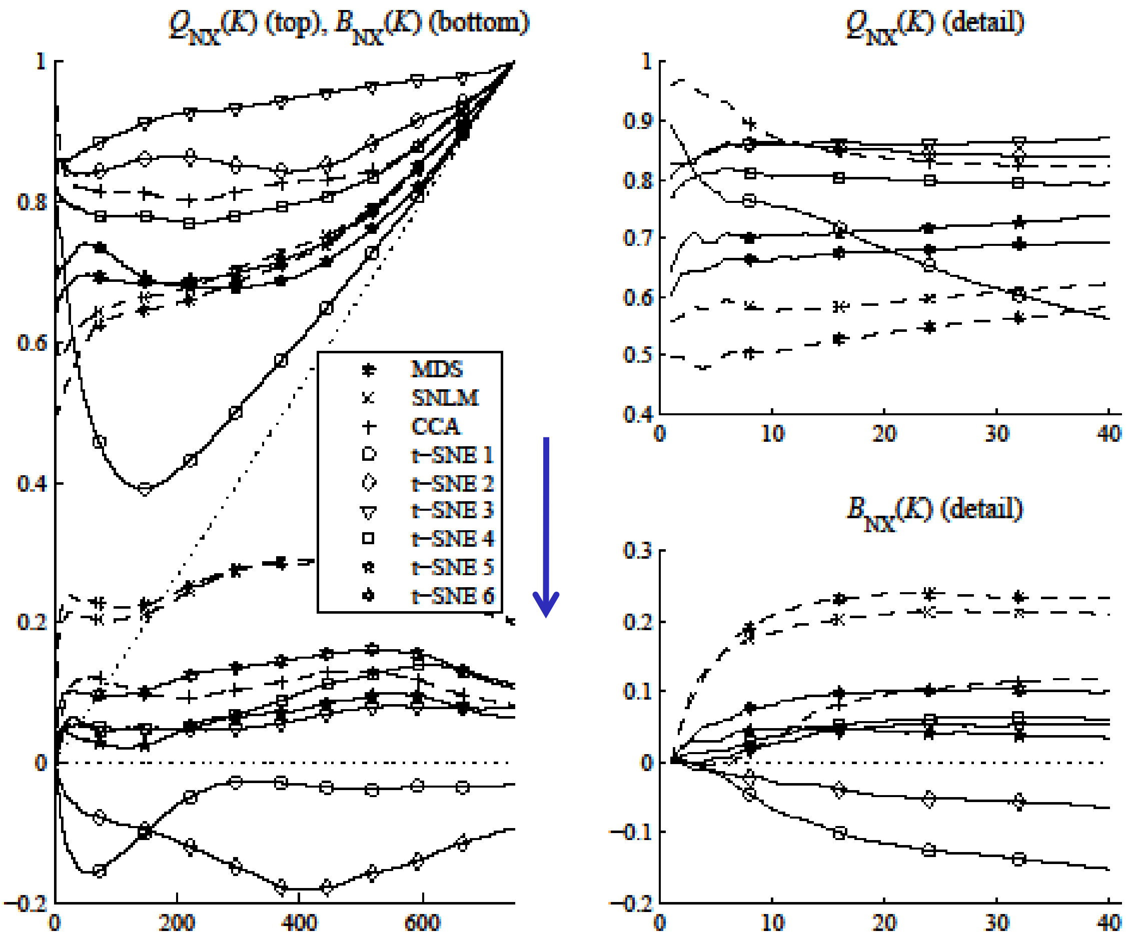
- Data: swiss roll



- Quality measures: in a K-neighborhood, we count the number of intrusions and extrusions. Then
  - $Q_{NX}(K)$ measures the overall number of intrusions and extrusions (higher $Q_{NX}(K)$ means better quality)
  - $B_{NX}(K)$ measures the difference between the number of intrusions and extrusions (positive$B_{NX}(K)$ means intrusive)

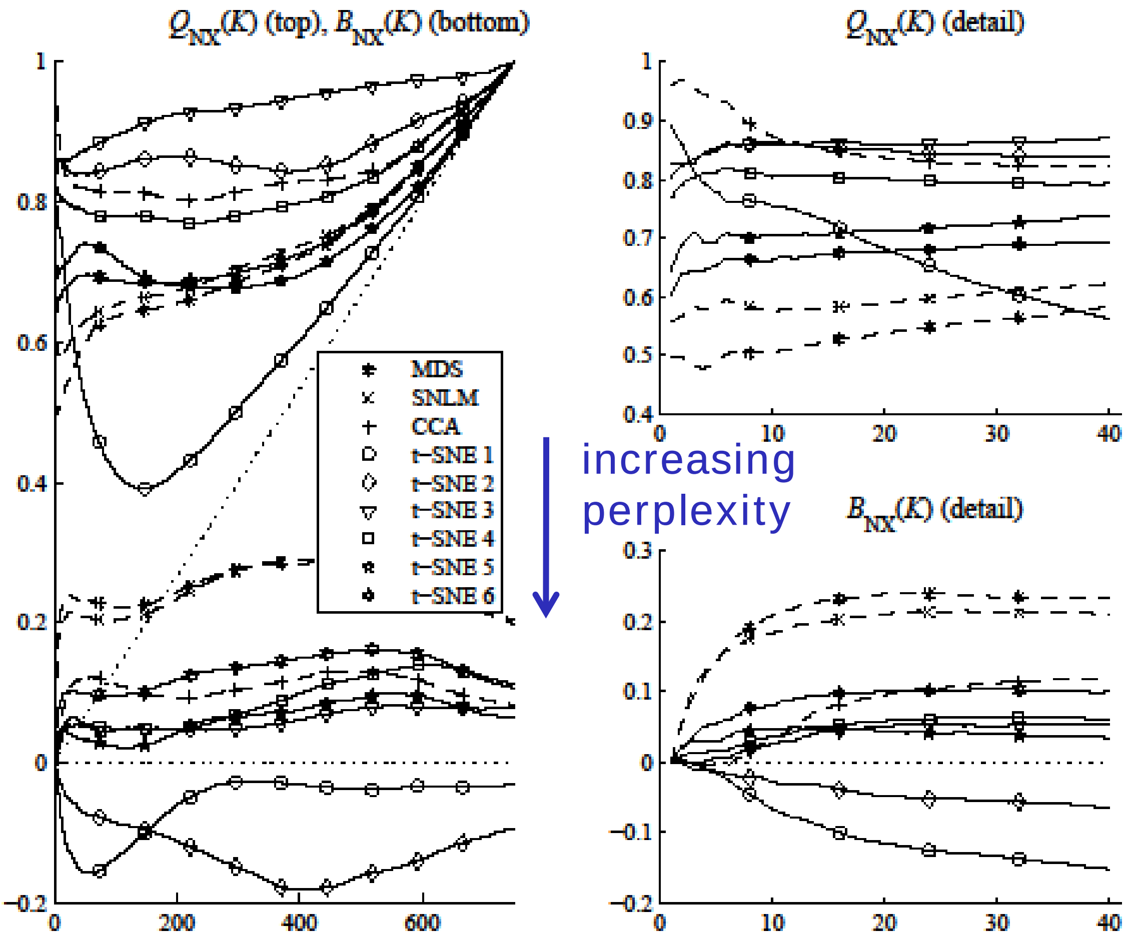
- Use of both Euclidean and geodesic distances

Results with Euclidean distances



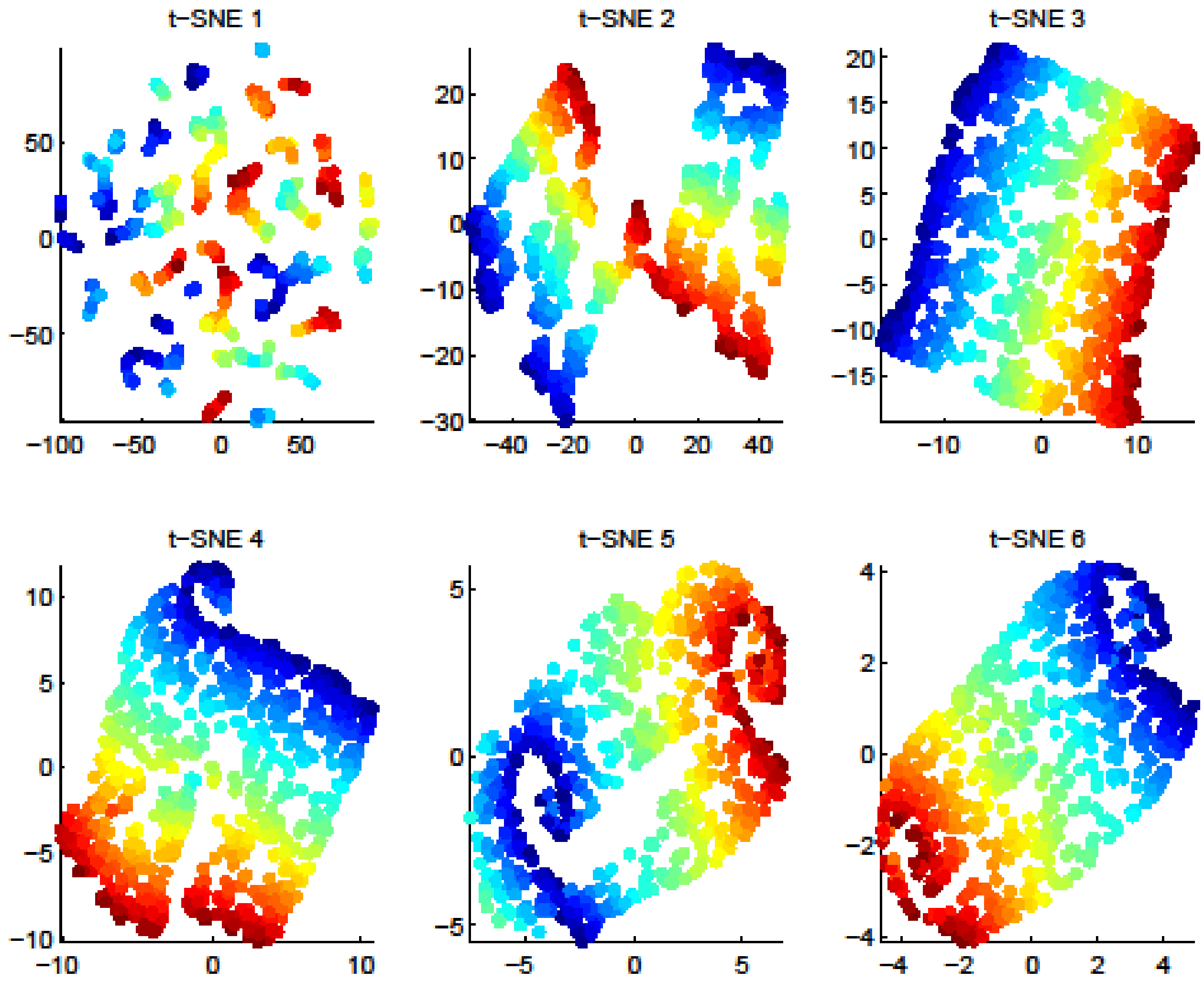
increasing perplexity

Results with Euclidean distances

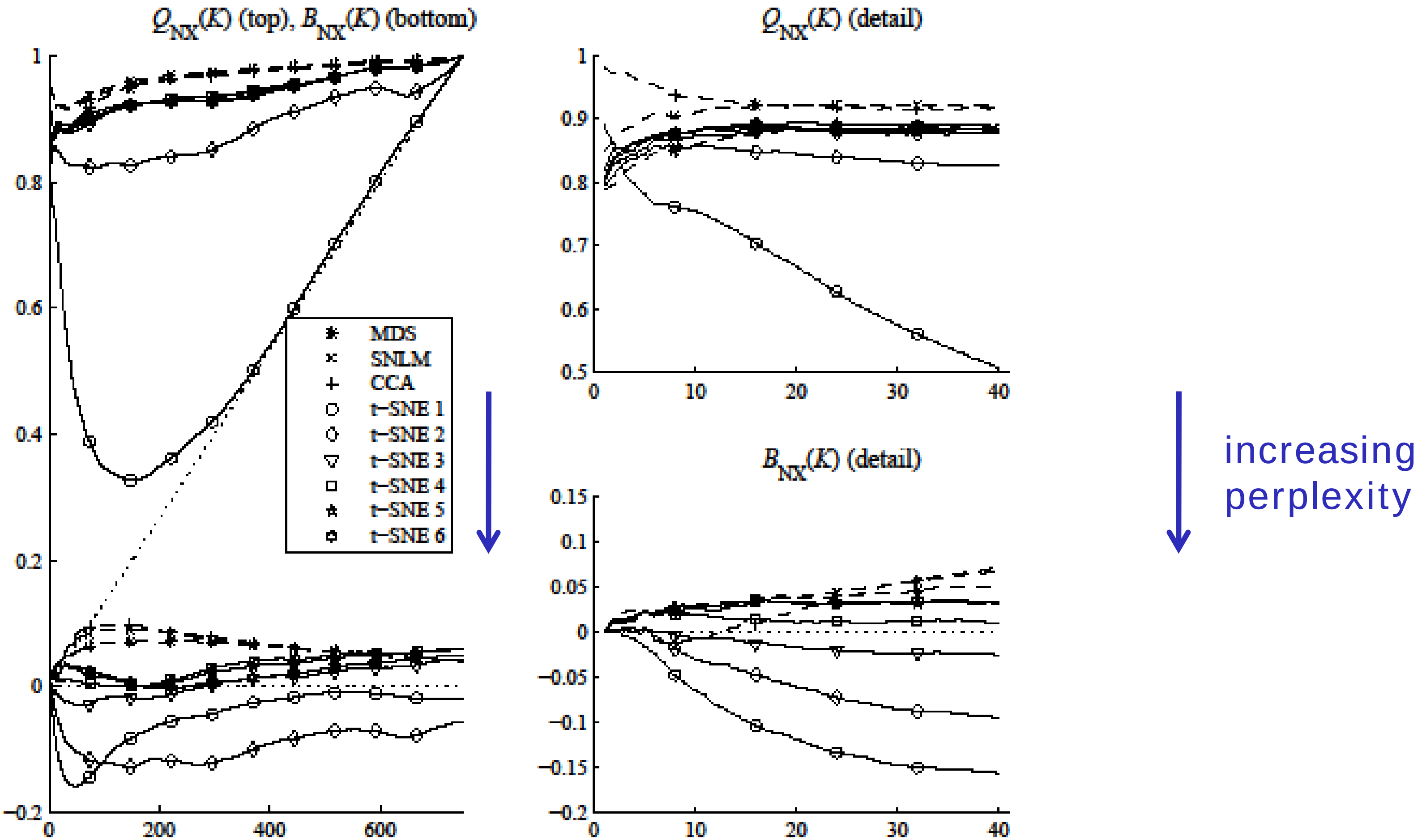


- Difficult problem! (low values of $Q_{NX}(K)$)
- t-SNE largely depends on perplexity

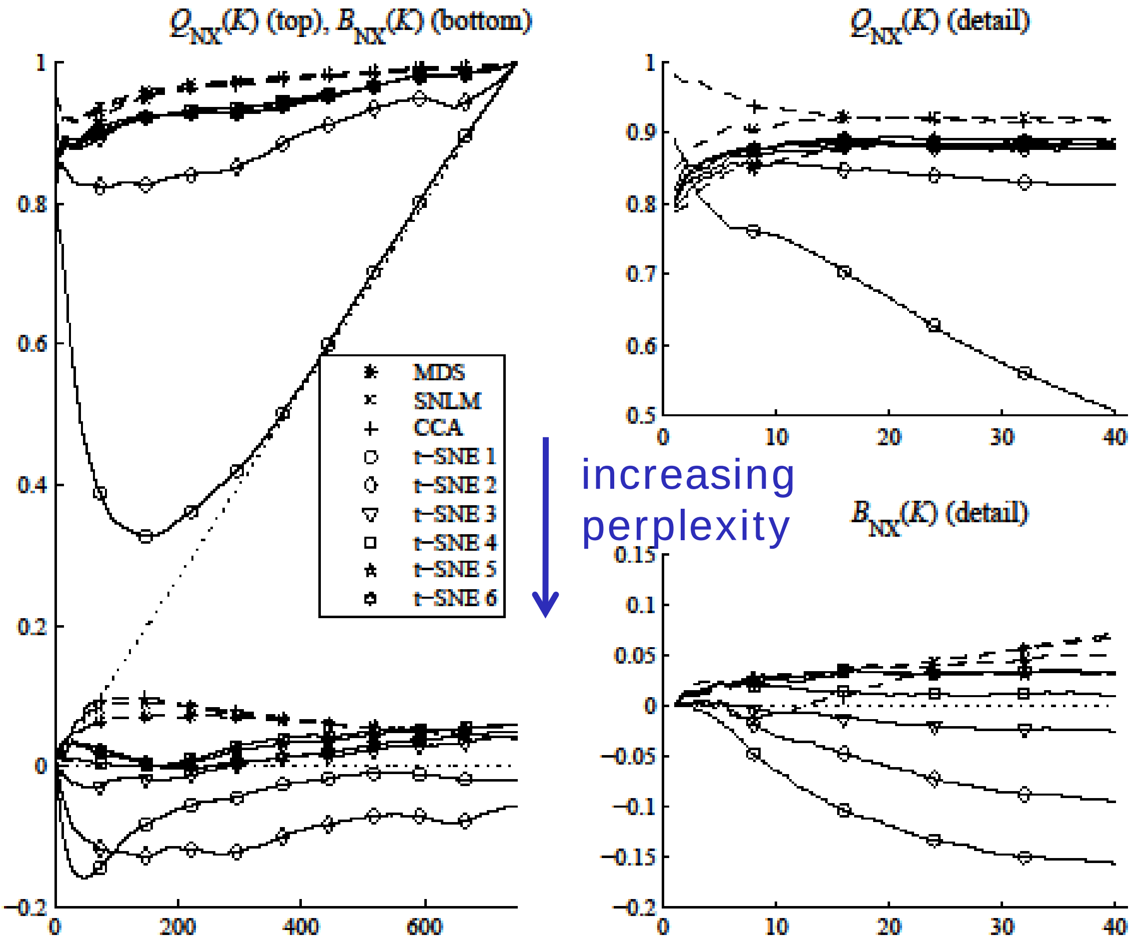
Results with Euclidean distances



Results with geodesic distances



Results with geodesic distances



- Geodesic distances facilitate the task
- CCA performs well!
- t-SNE still depends on perplexity, but large values help

Outline

- NDLR: a historical perspective
  - stress function
  - intrusion and extrusions
  - geodesic distances
- SNE and t-SNE
  - algorithm
  - gradient
  - transformed distances
- Experiments
  - with Euclidean distances
  - with geodesic distances
- Conclusions

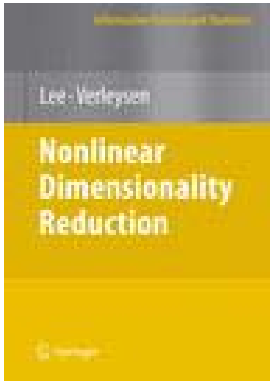
Conclusions

- t-SNE *is* a distance preservation method

- Stretching distances : good idea!
- But transformation in t-SNE not always optimal (not data driven)
- Careful tuning of parameters!

- Damping factor for large distances: good idea
- But this does not solve the issue of non-Euclidean manifolds (ex: hollow sphere)
- Situation is better with clustered data (stretching large distances improves the separation between clusters)

Advertisement



Nonlinear Dimensionality Reduction

Springer, Series: Information Science and Statistics

Lee, John A. - Verleysen, Michel

2007, Approx. 330 p. 8 illus. in color., Hardcover

ISBN: 978-0-387-39350-6

Software available at

http://www.dice.ucl.ac.be/mlg/index.php?page=NLDR