The intrusion-extrusion compromise for the projection and visualization of high-dimensional data

John A. Lee and Michel Verleysen

Machine Learning Group

Université catholique de Louvain, Belgium

{john.lee, michel.verleysen}@uclouvain.be

Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

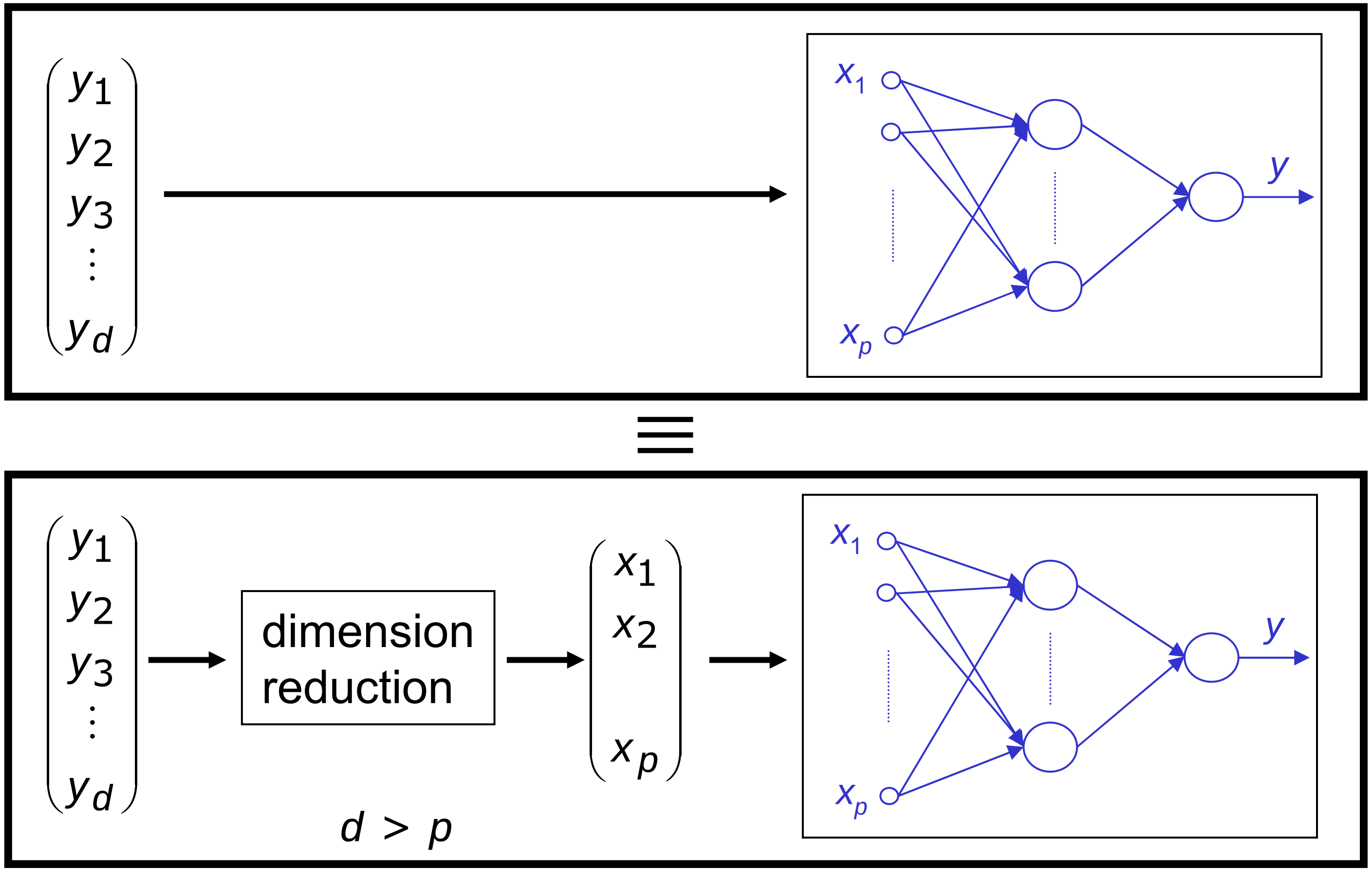
Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

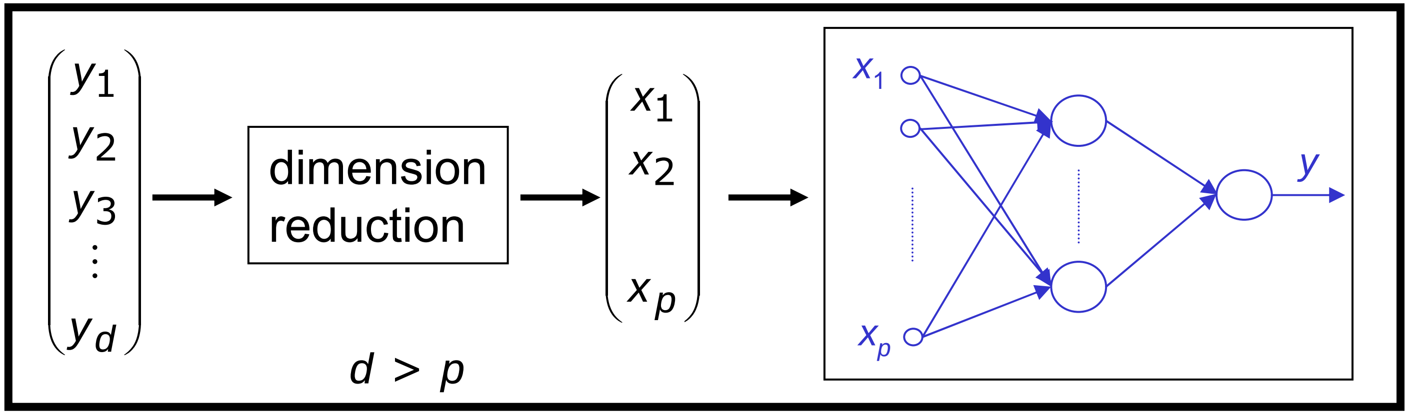
Motivation

- High-dimensional data are
  - difficult to represent
  - difficult to understand
  - difficult to analyze
- Example: nonlinear models such as MLP (Multi-Layer Perceptron) or RBFN (Radial-Basis Function Network) with many inputs: difficult convergence, local minima, etc.
- Need to reduce the dimension of data while keeping information content!

Reducing (the curse of) dimensionality



Reducing (the curse of) dimensionality



- Reducing the dimensionality
  - reduces the curse if dimensionality
  - makes models easier to learn
    - Local minima
    - Redundancy between inputs (non-idenfiability)
    - “Fills” the space

Visualization

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

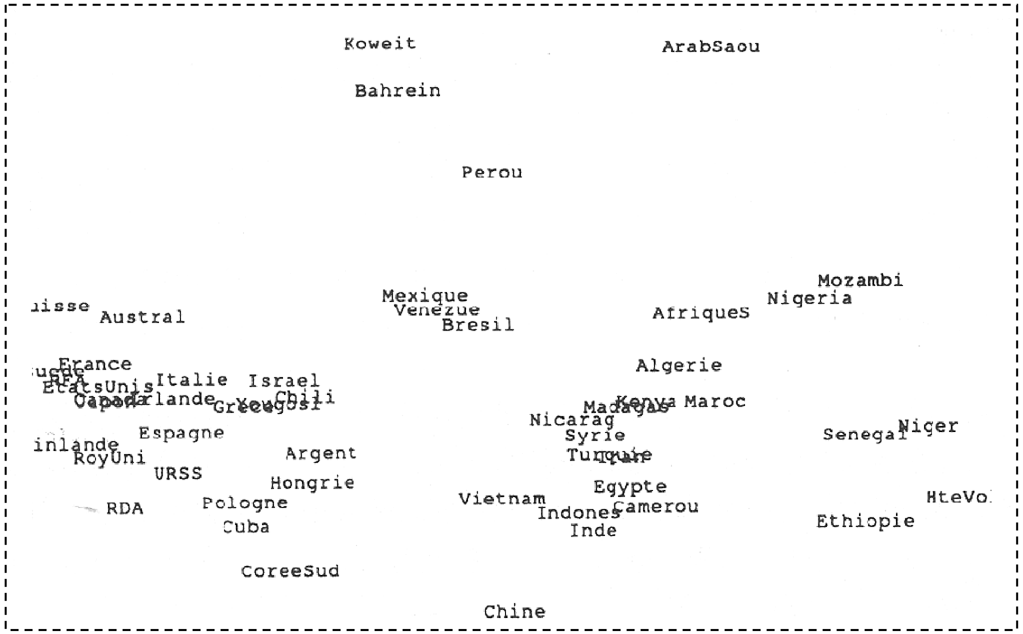
annual increase (%), infant mortality (‰), illiteracy ratio (%), school attendance (%), GIP, annual GIP increase (%)

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

[From Samos-Matisse, Univ. Paris 1]

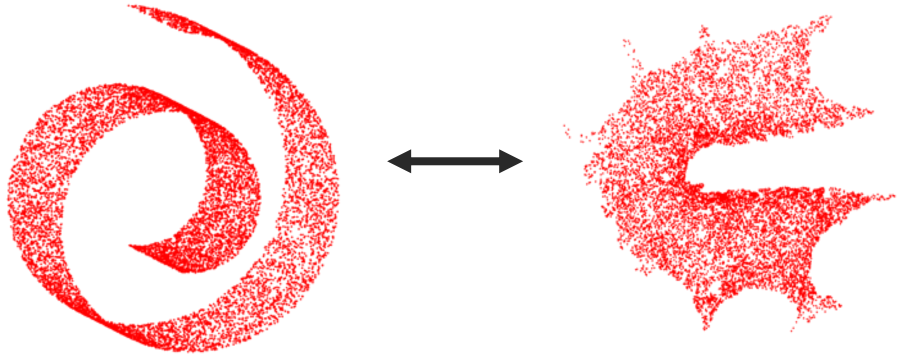
Visualization

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



What is a “perfect” method ?

1. A bijective mapping ?



2. A “nice” mapping ?

3. A mapping that preserves distances ?

4. A mapping that preserves topology (neighbors) ?

- Importance (and difficulty) to evaluate projections

Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Nonlinear projections: the paradigms

- Distance preservation
  - Distances between pairs of points in the original space, should match distances in the projection space
- Topology preservation
  - Neighbors in in the original space, should match neighbors in the projection space
- Information preservation
  - Forget the topology and distances, but pay attention to the reconstruction error

Nonlinear projections: the paradigms

- Distance preservation
  - Distances between pairs of points in the original space, should match distances in the projection space
- Topology preservation
  - Neighbors in in the original space, should match neighbors in the projection space
  - Few algorithms, beside SOM !
- Information preservation
  - Forget the topology and distances, but pay attention to the reconstruction error
  - No geometry, not quite adapted to visualization !

Nonlinear projections: the paradigms

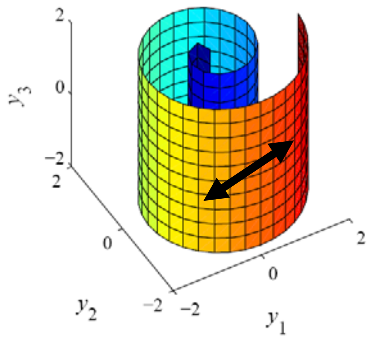
- Distance preservation
  - Distances between pairs of points in the original space, should match distances in the projection space
- Two main research directions:
  - Algebraic (spectral) methods
    - Linear models (possibly with nonlinear distances)
    - + easy calculations
      - often not adapted
  - Nonlinear objective criteria
    - Nonlinear models, more general
    - + more powerful, close to objectives
      - optimization more difficult

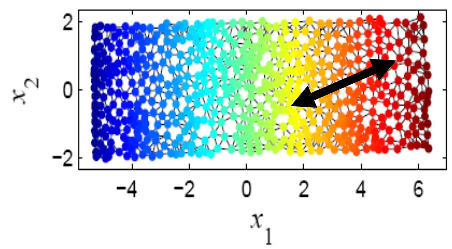
Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Distance preservation

- Many variations around the same theme





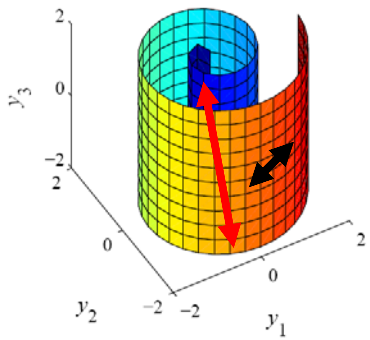
$d_y(i, j) = d(y(i), y(j)),\ \ y(i), y(j) \in \Re^d$ $\qquad$ $d_x(i, j) = d(x(i), x(j)),\ \ x(i), x(j) \in \Re^p$

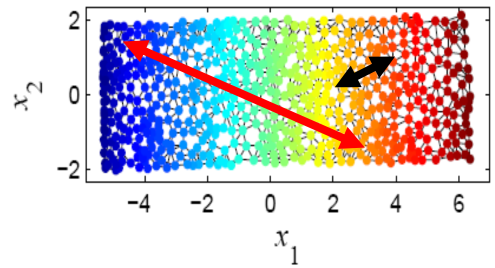
- The parameters of the method are the locations $x(i)$
- The objective (or cost, error, stress function) is some measure of discrepancy between $d_y(i,j)$ and $d_x(i,j)$

Metric multi-dimensional scaling (MDS)

- Metric MDS is roughly equivalent to minimizing

$$E = \sum_{i,j=1}^{N} \left(d_Y(i,j) - d_X(i,j) \right)^2$$



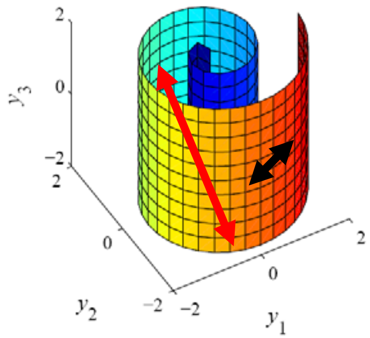


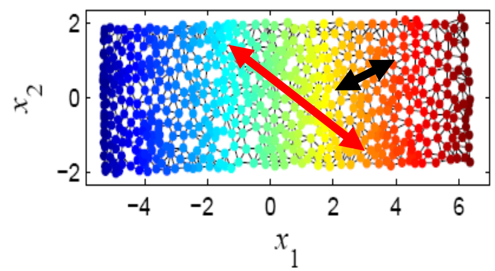
- Problem:
  - large distances contribute more (squared criterion), and
  - large distances are those that need to be enlarged (see )

Sammon's nonlinear mapping (NLM)

$$E_{NLM} = \sum_{\substack{i=1 \\ i<j}}^{N} \frac{\left(d_y(i,j) - d_x(i,j)\right)^2}{d_y(i,j)}$$

- Idea: to give more weight to the short distances

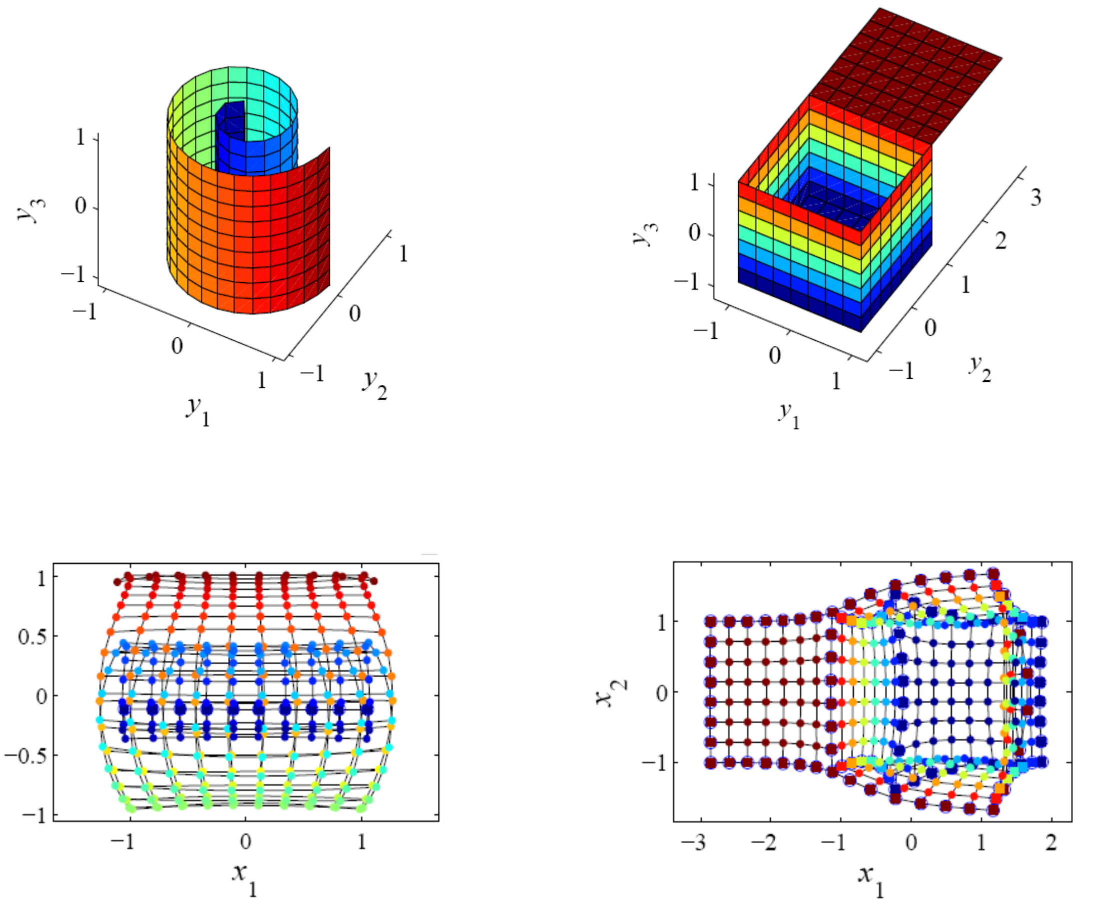




- Intuitively, (short black arrow) can be (approximately) preserved, while (long red arrow) will necessarily be enlarged

Sammon's nonlinear mapping (NLM)

- Examples



Curvilinear component analysis (CCA)

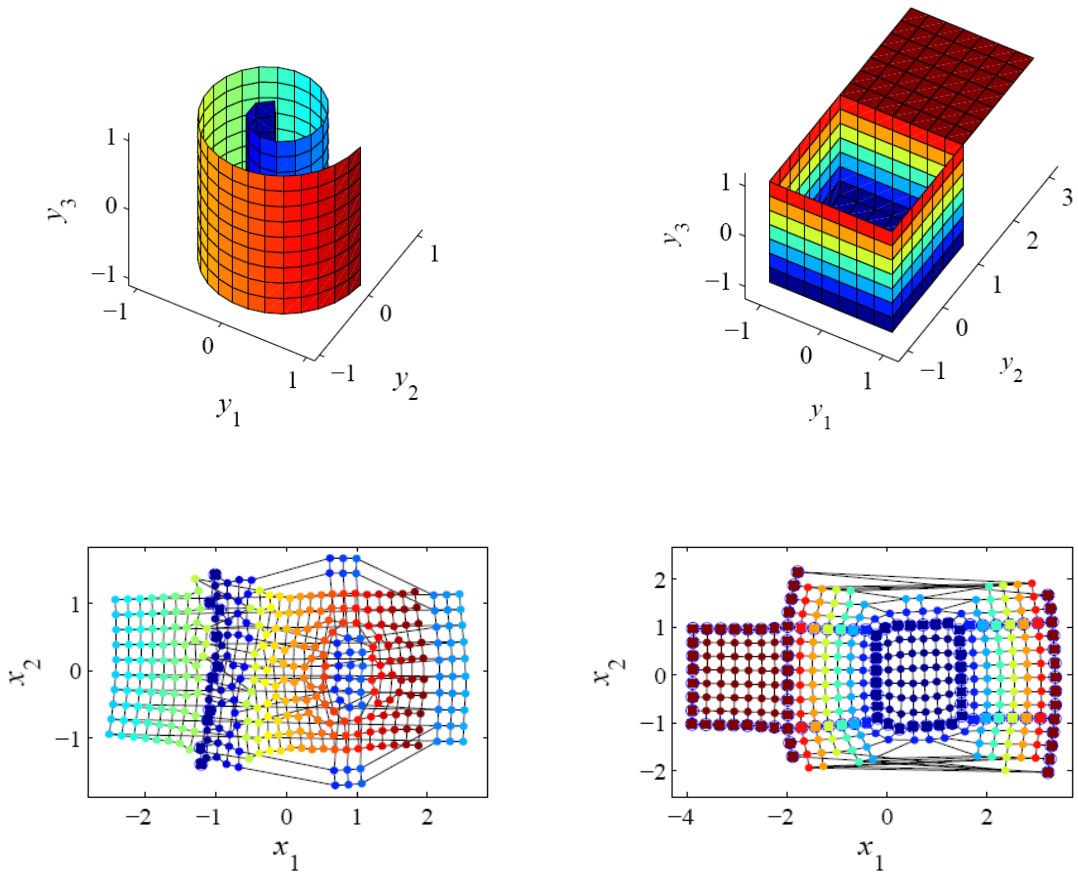
$$E_{CCA} = \sum_{\substack{i=1 \\ i<j}}^{N} \left(d_y(i,j) - d_x(i,j) \right)^2 F_\lambda\left(d_x(i,j) \right)$$

where F_λ is a monotonically decreasing function

- Idea: to give more weight to the short distances
- But: to short distances in the projection space (d_x, not d_y !)
  - This makes the differences for cuts: small d_y, large d_x is now possible!

Curvilinear component analysis (CCA)

- Examples

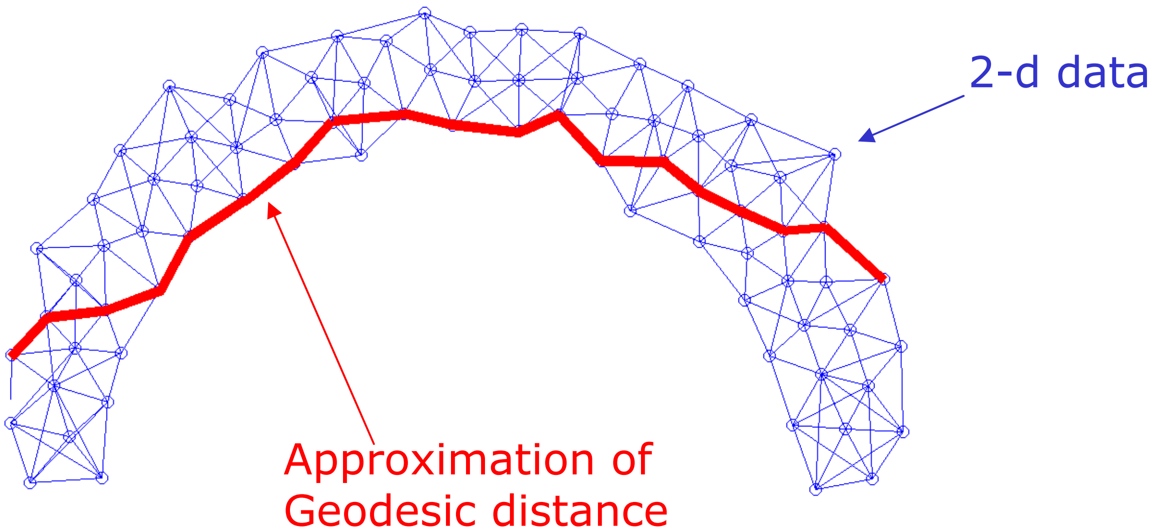


[Demartines – Hérault 92]

Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Geodesic distances



- How to build the graph from the data?
  - Connect each data to its *k* nearest neighbors, or
  - Connect each data to all other ones in a ε-ball
  - Ensure connectivity of the graph

Distance preservation: summary

	Euclidean distance	*Geodesic distance*
No weight	Metric MDS	Isomap
Weights on distances in y space	Sammon's mapping	Geodesic NLM
Weights on distances in x space	Curvilinear component analysis (CCA)	Curvilinear distance analysis (CDA)

Outline

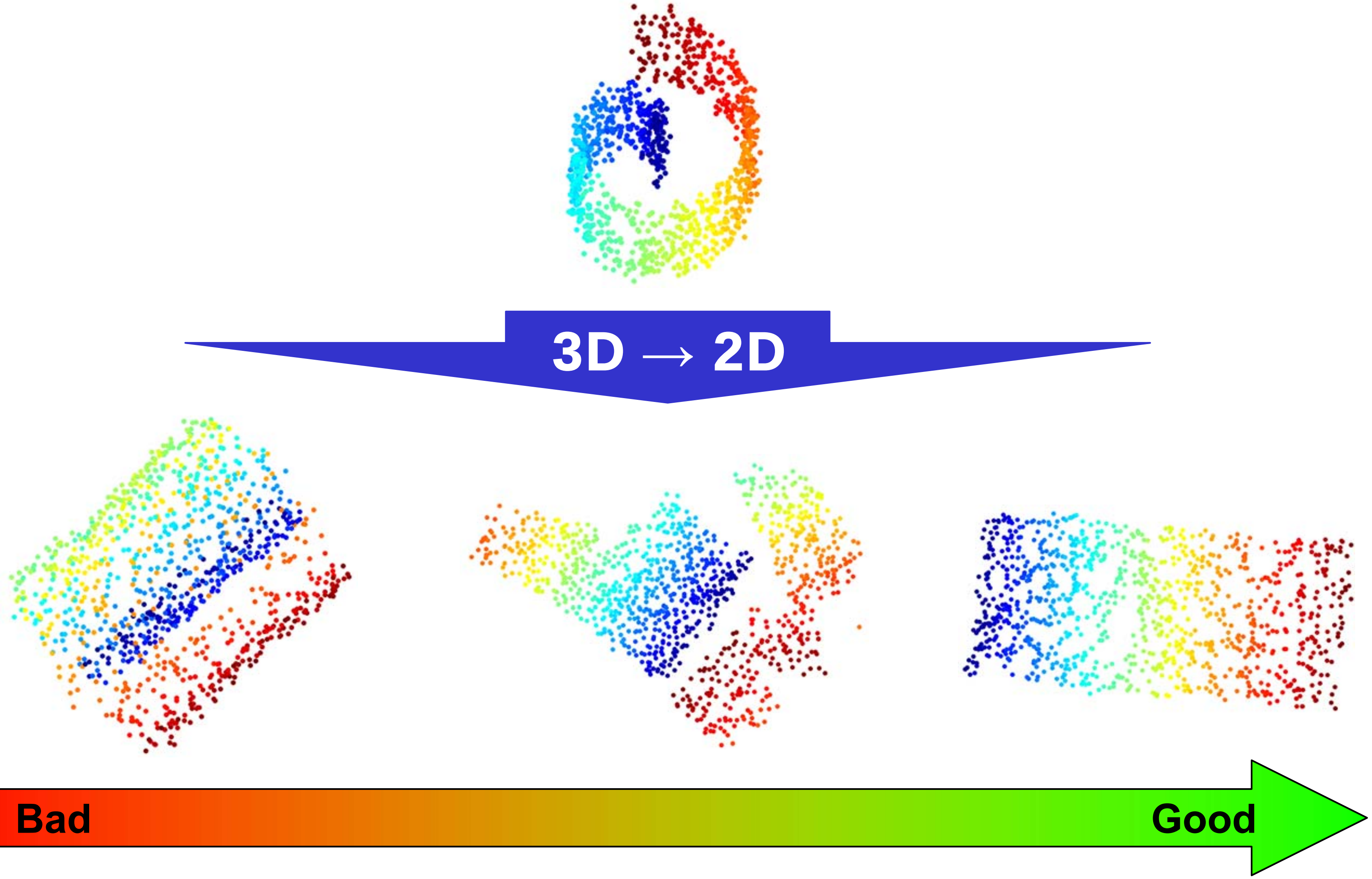
- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Performance evaluation

- The key question (in this talk ☺):

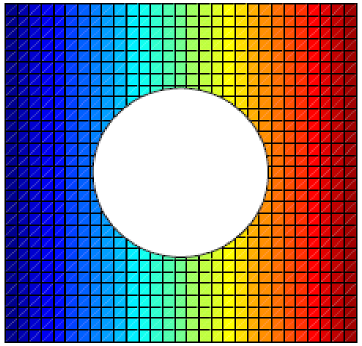
How to evaluate the performances of these methods?

Quality Assessment: Intuition

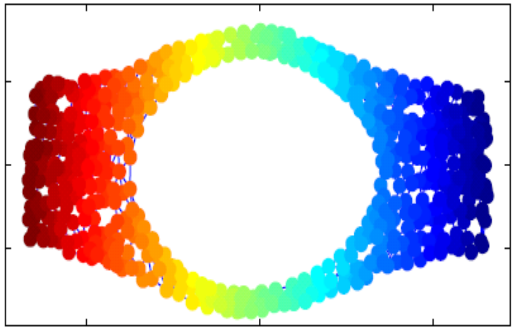


Quality Assessment: difficulty

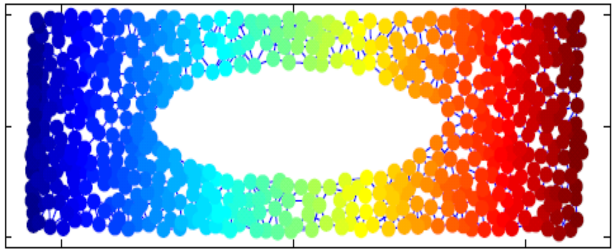
- A less intuitive assesment. When projecting



is this better



or that ?



Objective Quality Assessment

- We have:
  - An NLDR method to assess
- Some ideas:
  - Use its objective function
  - Quantify the distance preservation
  - Quantify the ‘topology’ preservation

Objective Quality Assessment

- We have:
  - An NLDR method to assess
- Some ideas:
  - Use its objective function ☹
  - Quantify the distance preservation ☹
  - Quantify the 'topology' preservation ☺

Objective Quality Assessment

- We have:
  - An NLDR method to assess
- Some ideas:
  - Use its objective function ☹
  - Quantify the distance preservation ☹
  - Quantify the 'topology' preservation ☺
- Topology in practice:
  - *K*-ary neighborhoods
  - Neighborhood ranks
- Literature:
  - 2001, Venna & Kaski: trustworthiness & continuity T&C
  - 2006, Chen & Buja: local continuity meta criterion LCMC
  - 2007, Lee & Verleysen: mean relative rank errors MRREs

Outline

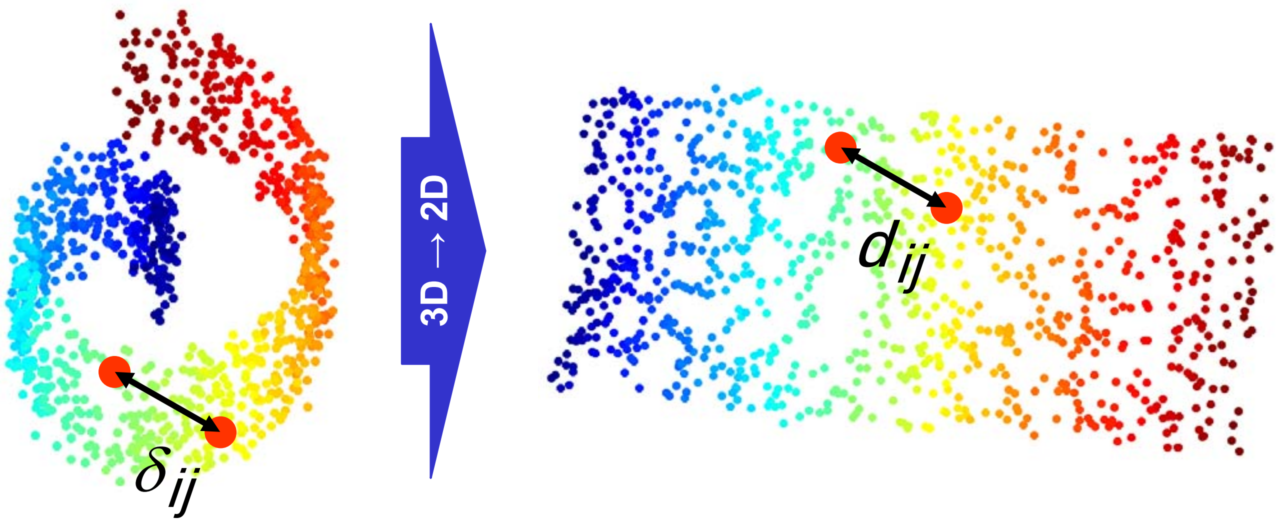
- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Distances, Ranks, and Neighbourhoods

- Distances: δ_{ij} denotes the distance from y_i to y_j
  d_{ij} denotes the distance from x_i to x_j

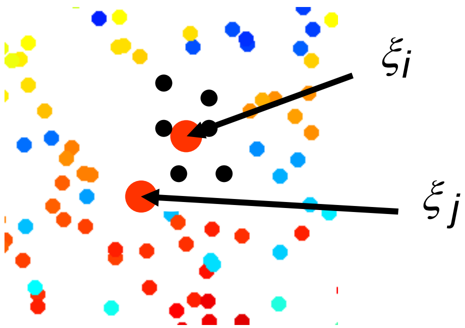
$$\boldsymbol{Y} = [y_i]_{1 \le i \le N}$$

$$\boldsymbol{X} = [x_i]_{1 \le i \le N}$$

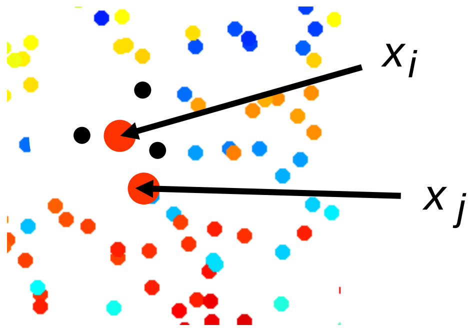


Distances, Ranks, and Neighbourhoods

- Ranks: $\rho_{ij} = \left|\left\{k : \delta_{ik} < \delta_{ij}\right\}\right|$ $\qquad r_{ij} = \left|\left\{k : d_{ik} < d_{ij}\right\}\right|$



$\rho_{ij} = 6$



$r_{ij} = 4$

- Neighborhoods: sets of indexes of black points (up to neighbor K)

$$\nu_i^K = \left|\left\{j : 1 \le \rho_{ij} < K\right\}\right|$$

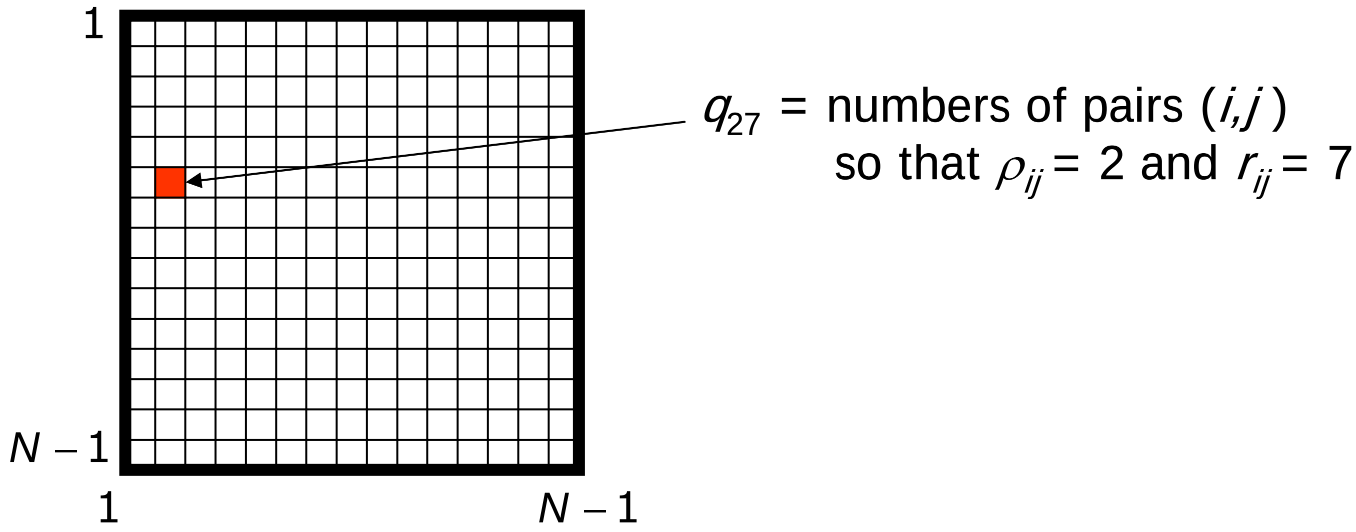
$$n_i^K = \left|\left\{j : 1 \le r_{ij} < K\right\}\right|$$

Distances, Ranks, and Neighbourhoods

- Co-ranking matrix:

$$\boldsymbol{Q} = [q_{kl}]_{1 \leq k,l \leq N-1}$$

with $q_{kl} = \left| \{ (i,j) : \rho_{ij} = k \text{ and } r_{ij} = l \} \right|$



(***Q*** is a sum of *N* permutation matrices of size *N*-1)

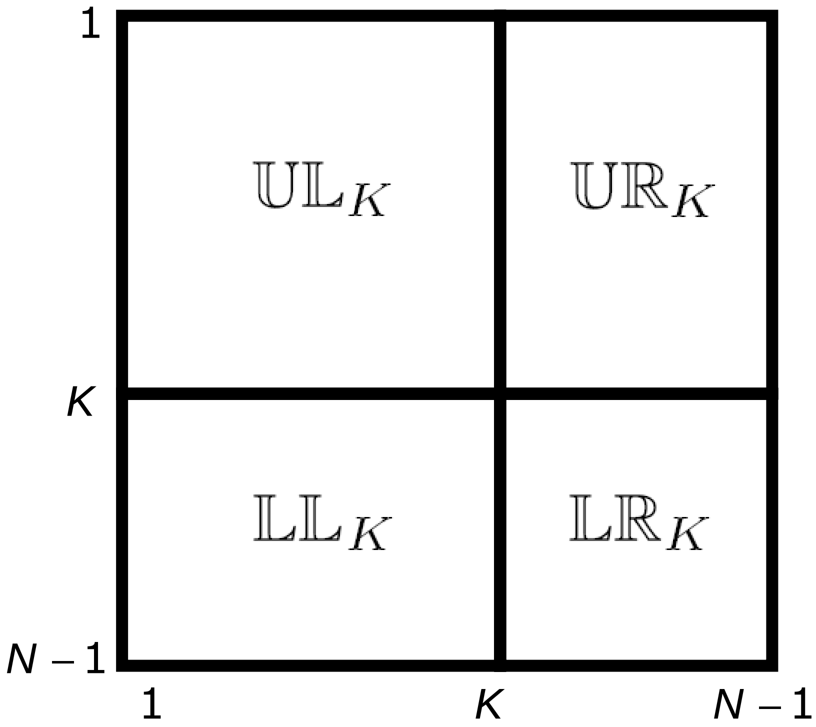
Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Co-ranking Matrix: Blocks

- *K*-ary neighbourhoods

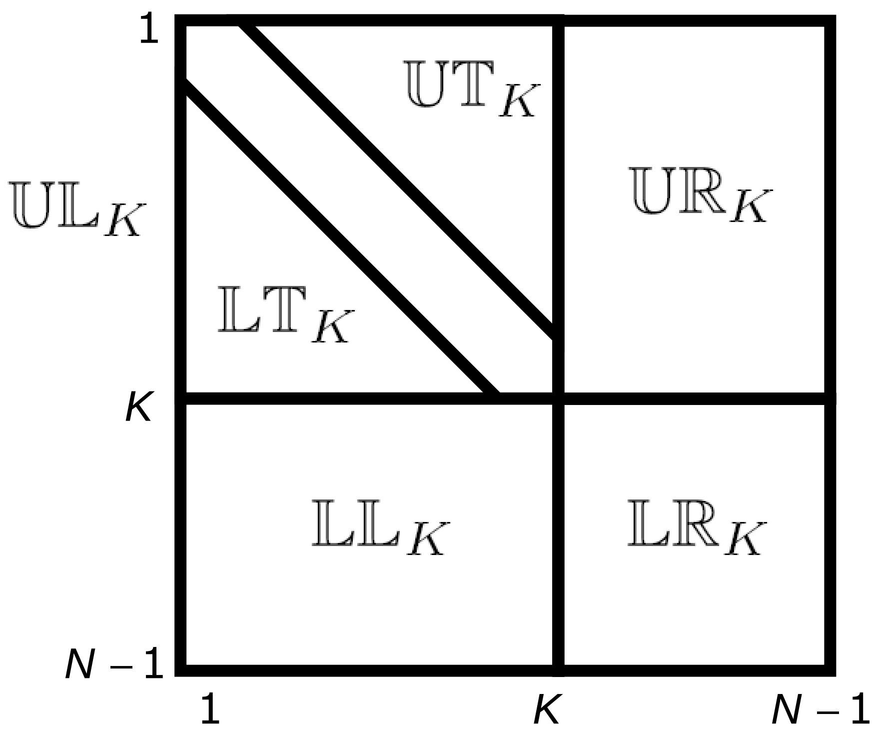
$$\mathbf{Q} = [q_{kl}]_{1 \leq k, l \leq N-1}$$



Co-ranking Matrix: Blocks

- *K*-ary neighbourhoods

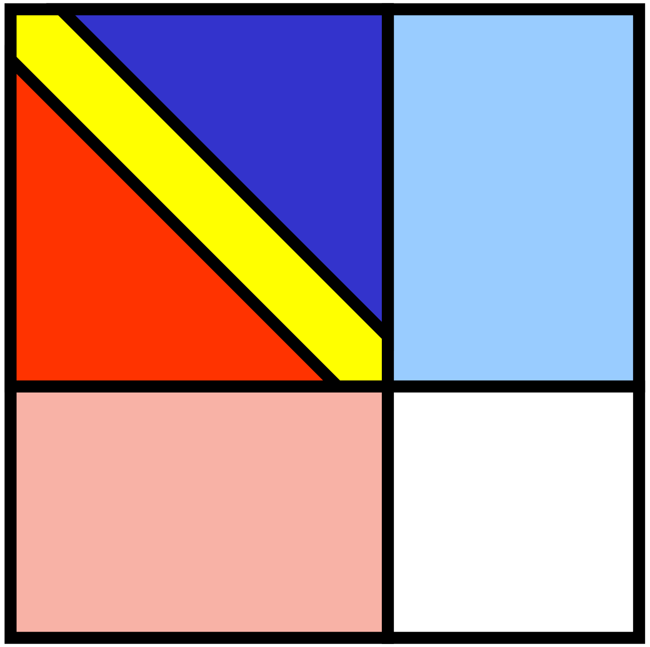
$$\mathbf{Q} = [q_{kl}]_{1 \leq k, l \leq N-1}$$



Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Intrusions and extrusions



- mild intrusions
- hard intrusions
- mild extrusions
- hard extrusions
- same rank

Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Existing criteria

- Thrustworthiness and Continuity
  (Venna and Kaski)

- Mean Relative Rank Errors
  (Lee and Verleysen)

- Local Continuity Meta Criterion
  (Chen & Buja)

Trustworthiness & Continuity

- Formulas:
  - trustworthiness

$$W_T(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in n_i^K \setminus \nu_i^K} \left(\rho_{ij} - K\right)$$

hard intrusions

  - continuity

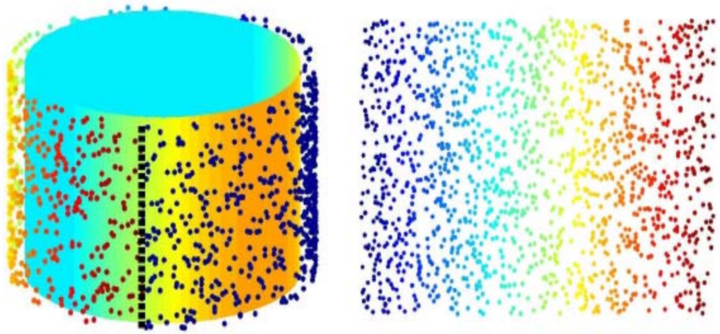
$$W_C(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K \setminus n_i^K} \left(r_{ij} - K\right)$$

hard extrusions

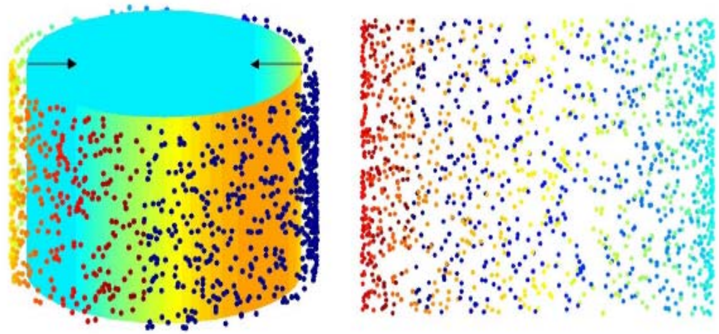
with $G_K = N \min\{K(2N - 3K - 1), (N - K)(N - K - 1)\}$

Why two criteria ?

- Because... not obvious to decide if it is better to cut (the projection is not continuous)



or to flatten (the projection is not trusthworthy)



Trustworthiness & Continuity

- Formulas:
  - trustworthiness

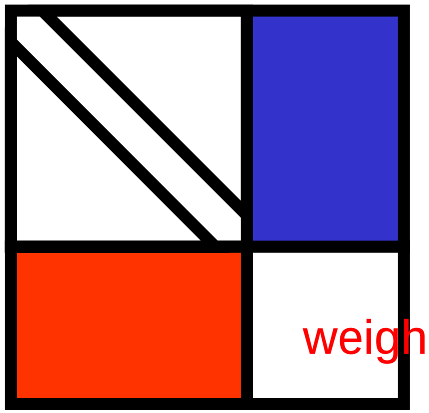
$$W_T(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in n_i^K \setminus \nu_i^K} \left(\rho_{ij} - K\right) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathrm{LL}_K} (k - K) q_{kl}$$

hard intrusions

  - continuity

$$W_C(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K \setminus n_i^K} \left(r_{ij} - K\right) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathrm{UR}_K} (l - K) q_{kl}$$

hard extrusions



weighted q_{kl} used for $W_C(K)$

weighted q_{kl} used for $W_T(K)$

Trustworthiness & Continuity

- Formulas:

$$W_T(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in n_i^K \setminus \nu_i^K} (\rho_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathrm{LL}_K} (k - K) q_{kl}$$

$$W_C(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K \setminus n_i^K} (r_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathrm{UR}_K} (l - K) q_{kl}$$

with $G_K = N \min\{K(2N - 3K - 1), (N - K)(N - K - 1)\}$

- Properties:
  - Distinguish between points that errouneously
    - enter a neighbourhood → trustwortiness
    - quit a neighbourhood → continuity
  - Functions of K (higher is better); range: [0,1] ([0.7,1])
  - Elements q_{kl} are weighted

Mean Relative Rank Errors

- Formulas:

$$E_n(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in n_i^K} \frac{\left|\rho_{ij} - r_{ij}\right|}{\rho_{ij}}$$

K-neighborhood in *X* space

$$E_\nu(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K} \frac{\left|\rho_{ij} - r_{ij}\right|}{r_{ij}}$$

K-neighborhood in *Y* space

with $H_K = N \sum_{k=1}^{K} \frac{|N - 2k|}{k}$

Mean Relative Rank Errors

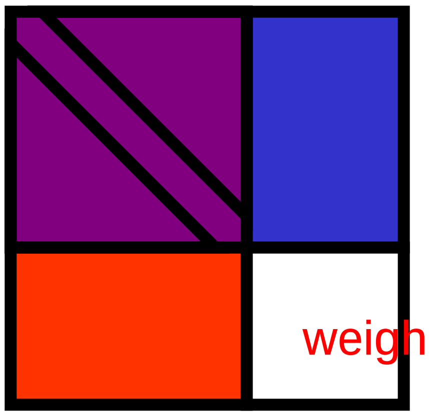
- Formulas:

$$E_n(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in n_i^K} \frac{|\rho_{ij} - r_{ij}|}{\rho_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathrm{UL}_K \cup \mathrm{LL}_K} \frac{|k - l|}{l} q_{kl}$$

K-neighborhood in X space

$$E_\nu(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K} \frac{|\rho_{ij} - r_{ij}|}{r_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathrm{UL}_K \cup UR_K} \frac{|k - l|}{k} q_{kl}$$

K-neighborhood in Y space



Mean Relative Rank Errors

- Formulas:

$$E_n(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in n_i^K} \frac{|\rho_{ij} - r_{ij}|}{\rho_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathrm{UL}_K \cup \mathrm{LL}_K} \frac{|k-l|}{l} q_{kl}$$

$$E_\nu(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K} \frac{|\rho_{ij} - r_{ij}|}{r_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathrm{UL}_K \cup UR_K} \frac{|k-l|}{k} q_{kl}$$

with $H_K = N \sum_{k=1}^{K} \frac{|N-2k|}{k}$

- Properties:
  - Two error types (same idea as in T&C)
  - Functions of K (***lower*** is better); range: [0,1] ([0,*0.3*])
  - Stricter than T&C: ***all*** rank errors are counted
  - Different weighting of q_{kl}

Local Continuity Meta-Criterion

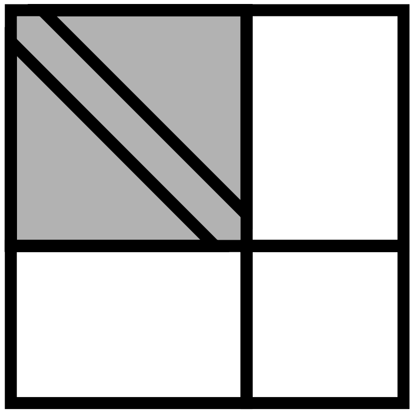
- Formula:

$$U_{LC}(K) = \frac{1}{NK} \sum_{i=1}^{N} \left(\left| n_i^K \cap \nu_i^K \right| - \frac{K^2}{N-1} \right)$$

Local Continuity Meta-Criterion

- Formula:

$$U_{LC}(K) = \frac{1}{NK}\sum_{i=1}^{N}\left(\left|n_i^K \cap \nu_i^K\right| - \frac{K^2}{N-1}\right) = \frac{K}{1-N} + \frac{1}{NK}\sum_{(k,l)\in \mathrm{UL}_K} \mathrm{q}_{kl}$$



unweighted q_{kl} used for $U_{LC}(K)$

Local Continuity Meta-Criterion

- Formula:

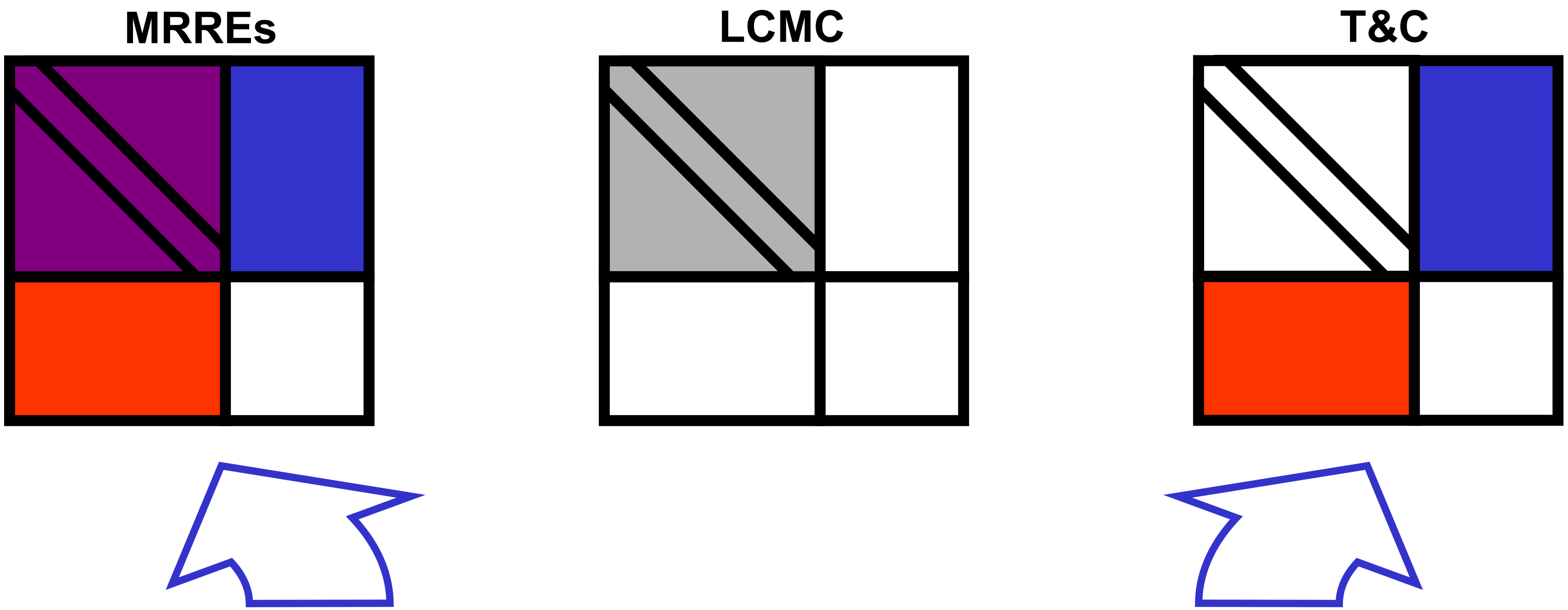
$$U_{LC}(K) = \frac{1}{NK}\sum_{i=1}^{N}\left(\left|n_i^K \cap \nu_i^K\right| - \frac{K^2}{N-1}\right) = \frac{K}{1-N} + \frac{1}{NK}\sum_{(k,l)\in \mathrm{UL}_K} \mathrm{q}_{\mathrm{kl}}$$

- Properties
  - Single measure
  - Function of *K* (higher is better); range: [0,1]
  - *A priori* milder than T&C and MRREs
  - Presence of a baseline term (random neighbourhood overlap)
  - No weighting of q_{kl}

Outline

- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

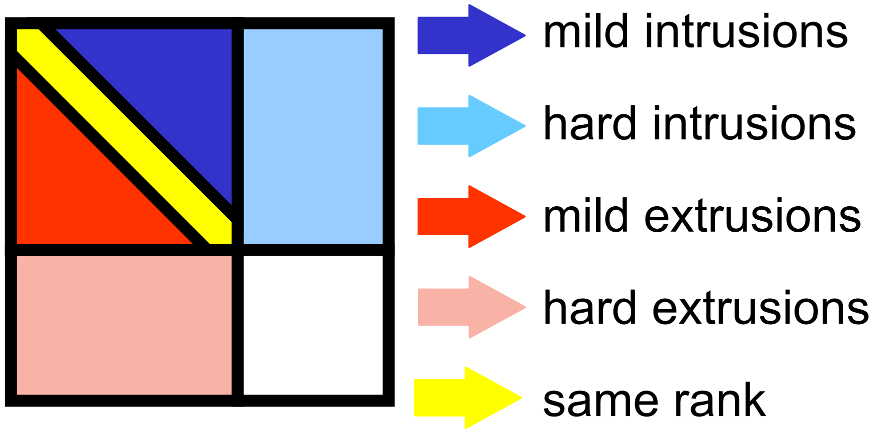
Unifying Framework



$$\sum_{(k,l)\in \mathrm{UL}_K \cup \mathrm{LL}_K} q_{kl} = \sum_{(k,l)\in \mathrm{UL}_K \cup \mathrm{UR}_K} q_{kl} = KN \quad \text{and} \quad \sum_{(k,l)\in \mathrm{LL}_K} q_{kl} = \sum_{(k,l)\in \mathrm{UR}_K} q_{kl}$$

Unweighted case: only the upper left block is important!

Unifying criteria



- Count *all* intrusions and extrusions

- Weigh them according to
  1) distance to diagonal
  2) rank

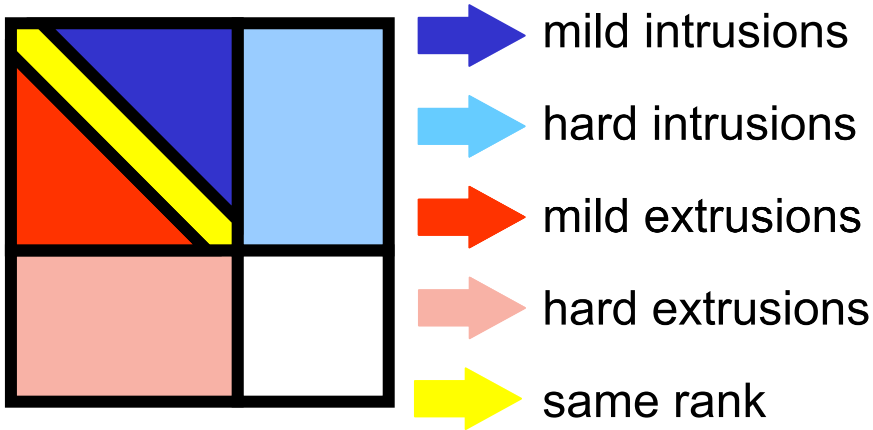
$$W_N^{\nu,w}(K) = \frac{1}{C_K} \sum_{(k,l) \in LT_K \cup LL_K} \frac{(k-l)^{\nu}}{k^{w}} q_{kl}$$

$$W_X^{\nu,w}(K) = \frac{1}{C_K} \sum_{(k,l) \in UT_K \cup UR_K} \frac{(l-k)^{\nu}}{l^{w}} q_{kl}$$

Unifying criteria

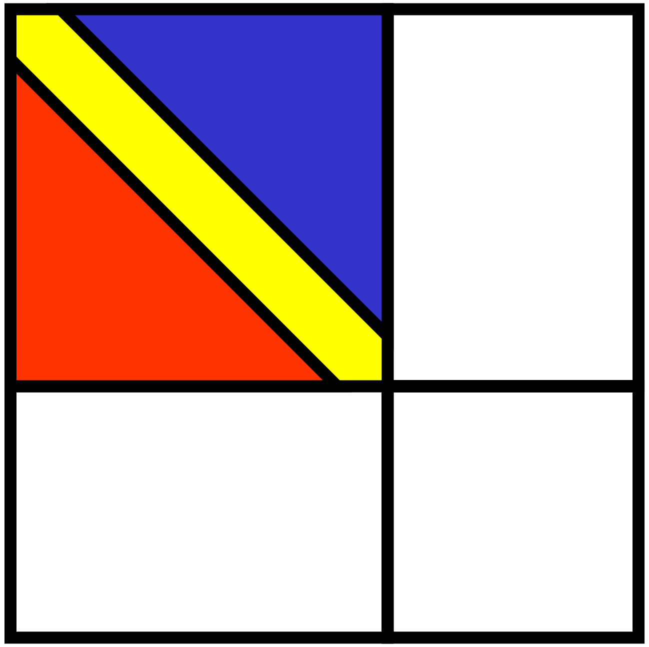
$$W_N^{\nu,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in LT_K \cup LL_K} \frac{(k-l)^{\nu}}{k^{w}} q_{kl}$$

$$W_X^{\nu,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in UT_K \cup UR_K} \frac{(l-k)^{\nu}}{l^{w}} q_{kl}$$



- More or less arbitrary weighting
- But no weighting is useless, because
  # hard K-intrusions = # hard K-extrusions
- ⇒ look inside K-ary neighborhoods

Unifying Framework

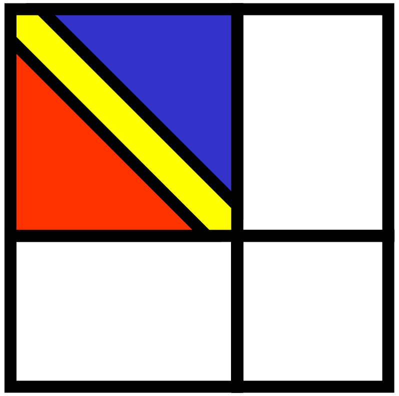


$$U_N(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{UT}_K} q_{kl}$$

$$U_X(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{LT}_K} q_{kl}$$

$$U_P(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{D}_K} q_{kl}$$

Unifying Framework



$$U_N(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{UT}_K} q_{kl}$$

$$U_X(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{LT}_K} q_{kl}$$

$$U_P(K) = \frac{1}{KN} \sum_{(k,l) \in \mathrm{D}_K} q_{kl}$$

- Overall quality of embedding:

$$Q_{NX}(K) = U_P(K) + U_N(K) + U_X(K) = U_{LC}(K) + \frac{K}{N-1}$$

- Overall "behaviour" of embedding

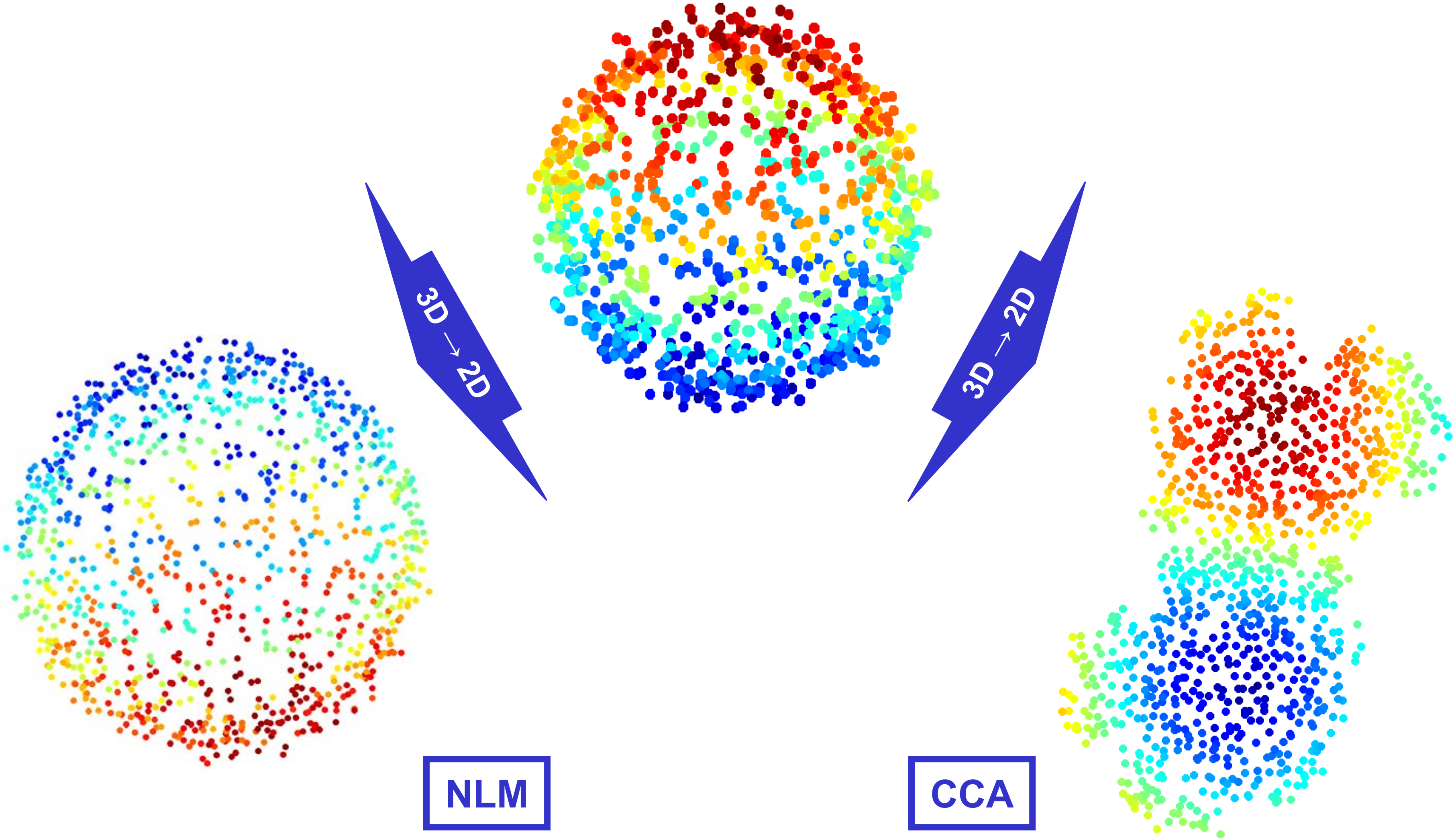
$$B_{NX}(K) = U_N(K) - U_X(K)$$

$B_{NX}(K) > 0$: intrusive
$B_{NX}(K) < 0$: extrusive

Outline

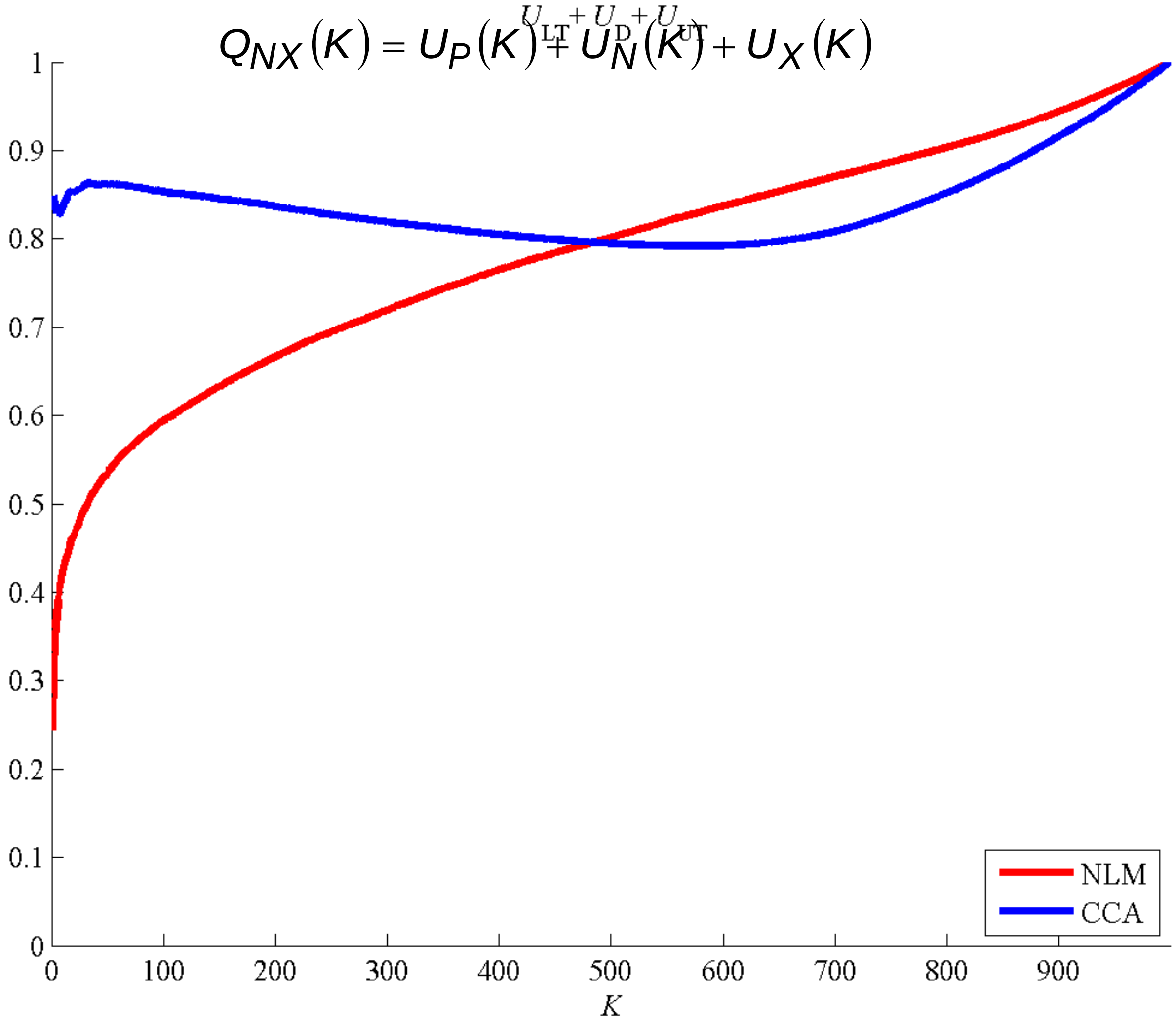
- Motivation: why nonlinear dimensionality reduction?
- Paradigms
- Distance preservation methods
  - Euclidean distances
  - Graph distances
- Quality assessment
  - Distances, Ranks, and Neighbourhoods
  - Co-ranking Matrix
  - Intrusions and extrusions
  - Existing criteria
  - Unifying framework
- Experiments

Experiment: Hollow Sphere

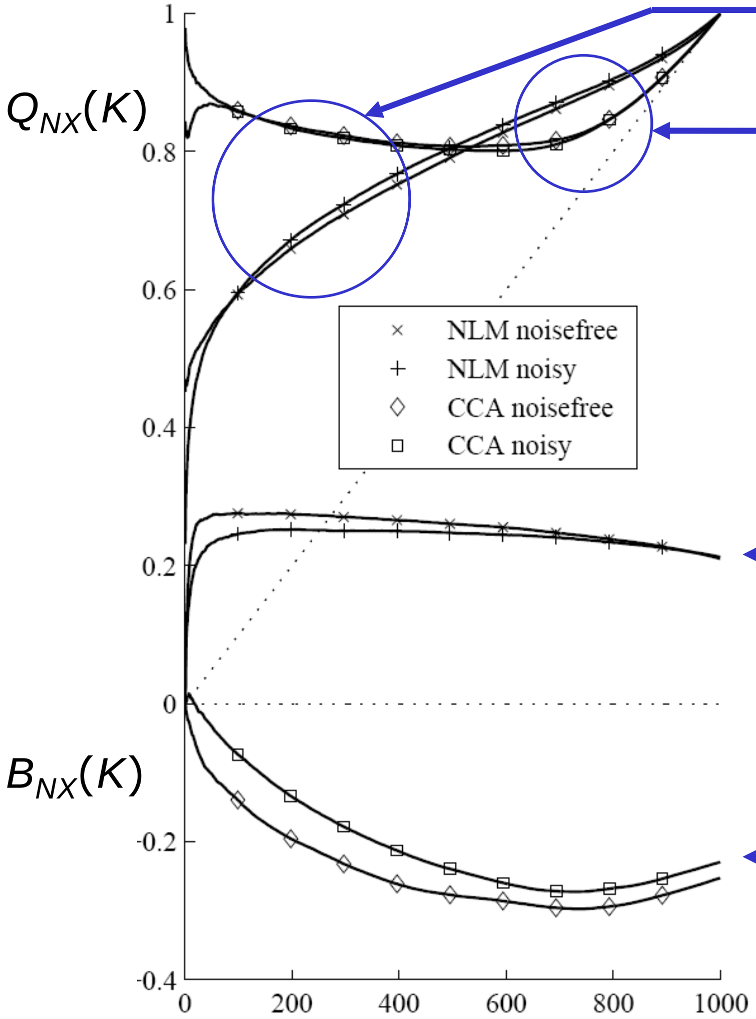


Experiment: Hollow Sphere

$$Q_{NX}(K) = U_P(K) + U_N(K) + U_X(K)$$



Experiment: Hollow Sphere



CCA is locally better than NLM

(except for the global topology)

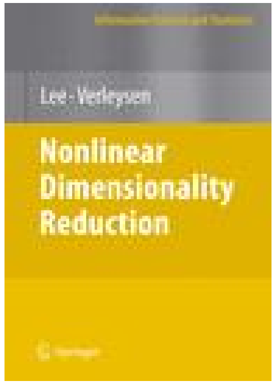
NLM creates intrusions

CCA creates extrusions

Conclusions

- Rank preservation is useful in NLDR QA:
  - More powerful than distance preservation
  - Reflects the appealing idea of 'topology' preservation
- Unifying framework:
  - Relies on the co-ranking matrix
    (≈ Shepard diagram with ranks instead of distances)
  - Involves no (arbitrary) weighting
  - Focuses on the inside of K-ary neighborhoods
    (otherwise a smart weighting is necessary)
  - Defines three errors:
    - A global error (like LCMC)
    - 'Type I and II' errors (like T&C and MRREs)
- Experiments:
  - They confirm the soundness of the approach
- Future prospect:
  - From rank-based NLDR ***QA*** to rank-based NLDR ***methods***

Nonlinear dimensionality reduction: the book



Nonlinear Dimensionality Reduction
Springer, Series: Information Science and Statistics
John A. Lee, Michel Verleysen
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

Thank you for your attention!

If you have any question…

Please visit: http://www.ucl.ac.be/mlg/