A Novel Method for Manifold Construction

Wei-Chen Cheng and Cheng-Yuan Liou*
Department of Computer Science and Information Engineering
National Taiwan University
Republic of China
*cyliou@csie.ntu.edu.tw

25th, Nov., 2008
15:50-17:30
Auckland

Related Works

Year	People	
1992	Erwin, Obermayer, Schulten	Approx. energy function of SOM
1991	Luttrell	Least distortion measure in SOM
1992		
1997	Kohonen	Average expected distortion measure
1997	Pedrycz, Waletzky	Anisotropic mapping
2000	Liou, Tai	Conformality, angle invariance in SOM
2000	Liou, Chen, Huang	Separating two classes and clustering the same class in SOM
2007	Liou, Cheng	Distance invariance

Energy function for SOM

- E. Erwin, K. Obermayer, K. Schulten, (1992)

$$\varepsilon_i \equiv \varepsilon \int \hat{H}\left(H, W\right) \frac{1}{2} \left(x - w_i\right)^2 P\left(x\right) d^p x + Z_i$$

- Approximated form

Least distortion measure

- Luttrell (1991, 1992)

$$d_i = \sum_j h_{ij} \left\| x - m_j \right\|^2$$

Trying to comprehend SOM

Average expected distortion measure

- T. Kohonen (1997)

$$E = \sum_i \int_{x \in X_i} \sum_k h_{ik} \left\| x - m_k \right\|^2 p(x) dx$$

$$E = \int ep(x) dx = \int \sum_{i \in L} h_{ci} d(x, m_i) p(x) dx$$

Anisotropy mapping for SOM

- W. Pedrycz, J. Waletzky (1997)
- The anisotropy of this metric means that one can find such pairs of patterns x_1, x_2 and x_3, x_4 such that $\|\mathbf{x}_1 - \mathbf{x}_2\|$ and $\|\mathbf{x}_3 - \mathbf{x}_4\|$ are equal yet the values $\|NN(\mathbf{x}_1) - NN(\mathbf{x}_2)\|$ and $\|NN(\mathbf{x}_3) - NN(\mathbf{x}_4)\|$ can differ quite substantially.

$$Q = \left[\text{target} - \|NN(\mathbf{x}_1) - NN(\mathbf{x}_2)\|^2\right]^2$$

Conformality

- Liou, Dai (2000)
- Differential geometry
- Angle preservation in SOM
- Patterns may not have a fixed space structure, such as tree or chain with flexible joints.

SIR

Internal representations for SOM

- Liou, Chen, Huang (2000)

$$E^{rep} = \frac{1}{2}\sum_{p_1}^{P}\sum_{p_2}^{P}\left(d\left(y^{(p_1)} - y^{(p_2)}\right)\right)^2 = \sum_{p_1}^{P}\sum_{p_2}^{P} E_{p_1 p_2}$$

- Separating two classes and clustering the same class in a hidden layer
- $d\left(y^{(p_1)} - y^{(p_2)}\right)$ flexible for various designs
- Flexible design of output space y, tree
- Relative distances only
- Capable of anisotropy

Distance invariance

- Liou, Cheng (2007)

$$E(r)=\frac{1}{4}\sum_{p}\sum_{\mathbf{x}^q\in U(p,r)}\left(\left\|\mathbf{y}^p-\mathbf{y}^q\right\|^2-\left\|\mathbf{x}^p-\mathbf{x}^q\right\|^2\right)^2=\frac{1}{4}\sum_{p}\sum_{\mathbf{x}^q\in U(p,r)}\left(d\left(y^p-y^q\right)\right)^2$$

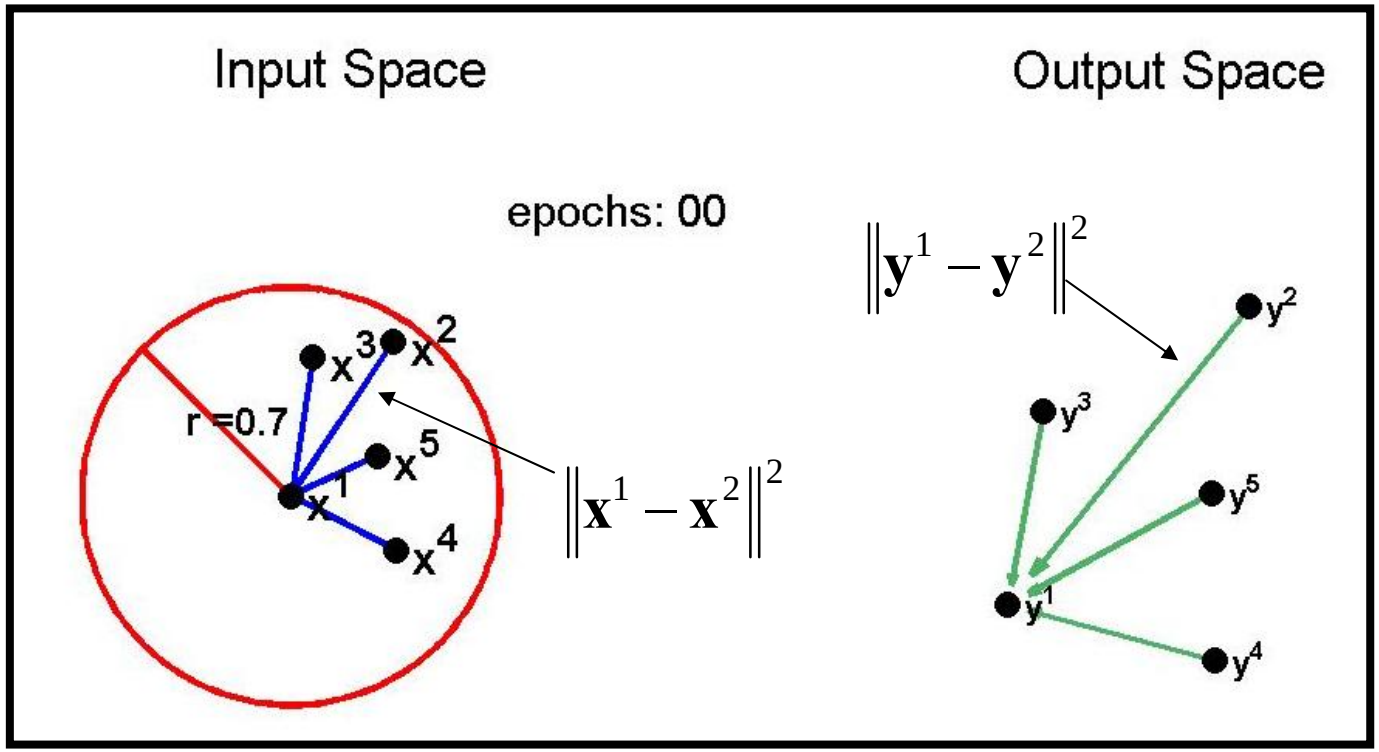
- Relative distances only
- Flexible design of output space
  - 2D, tree, or chain
- Solving S-shape problem in LLE, Isomap (2007)
- Maximally resemble the pattern distortions
  - Visualization of physical meaning among patterns
- Perfect energy function
  - N = number of data, save $p(x)$ for $p(y)$
- Isotropy

Flexible design

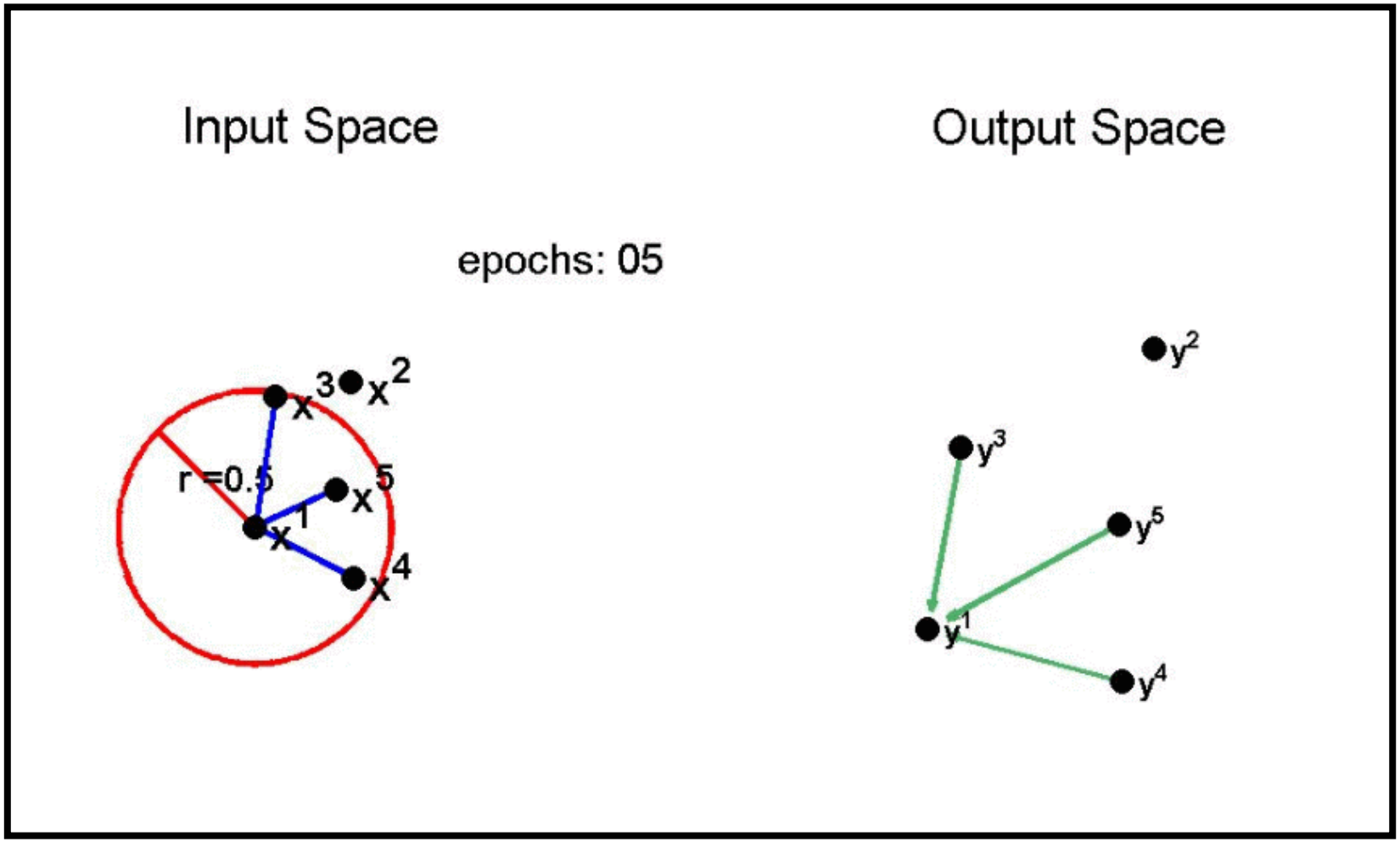
Density problem in LLE

Objective Function

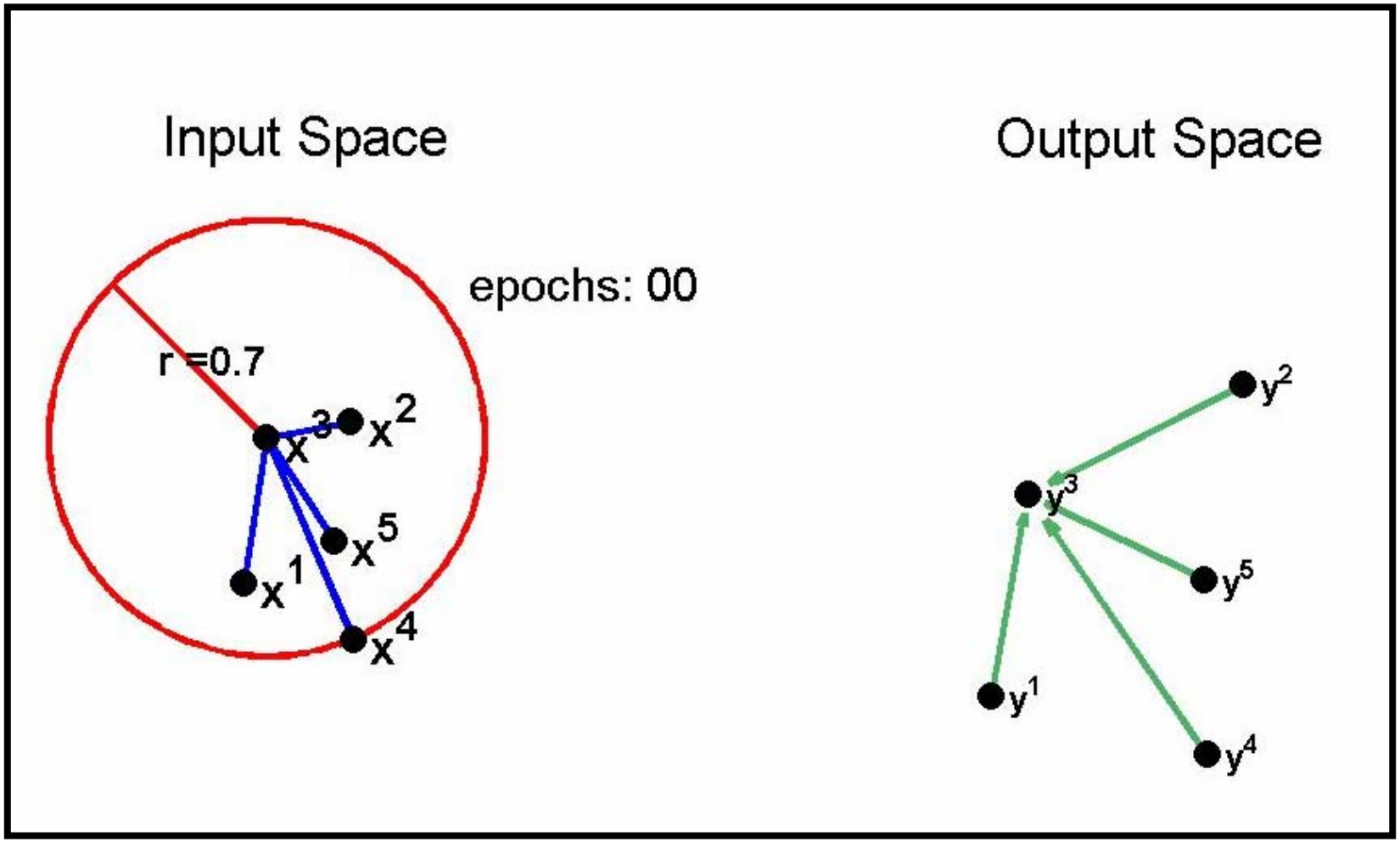
$$E(r)=\frac{1}{4}\sum_{p}\sum_{\mathbf{x}^q\in U(p,r)}\left(\left\|\mathbf{y}^p-\mathbf{y}^q\right\|^2-\left\|\mathbf{x}^p-\mathbf{x}^q\right\|^2\right)^2$$



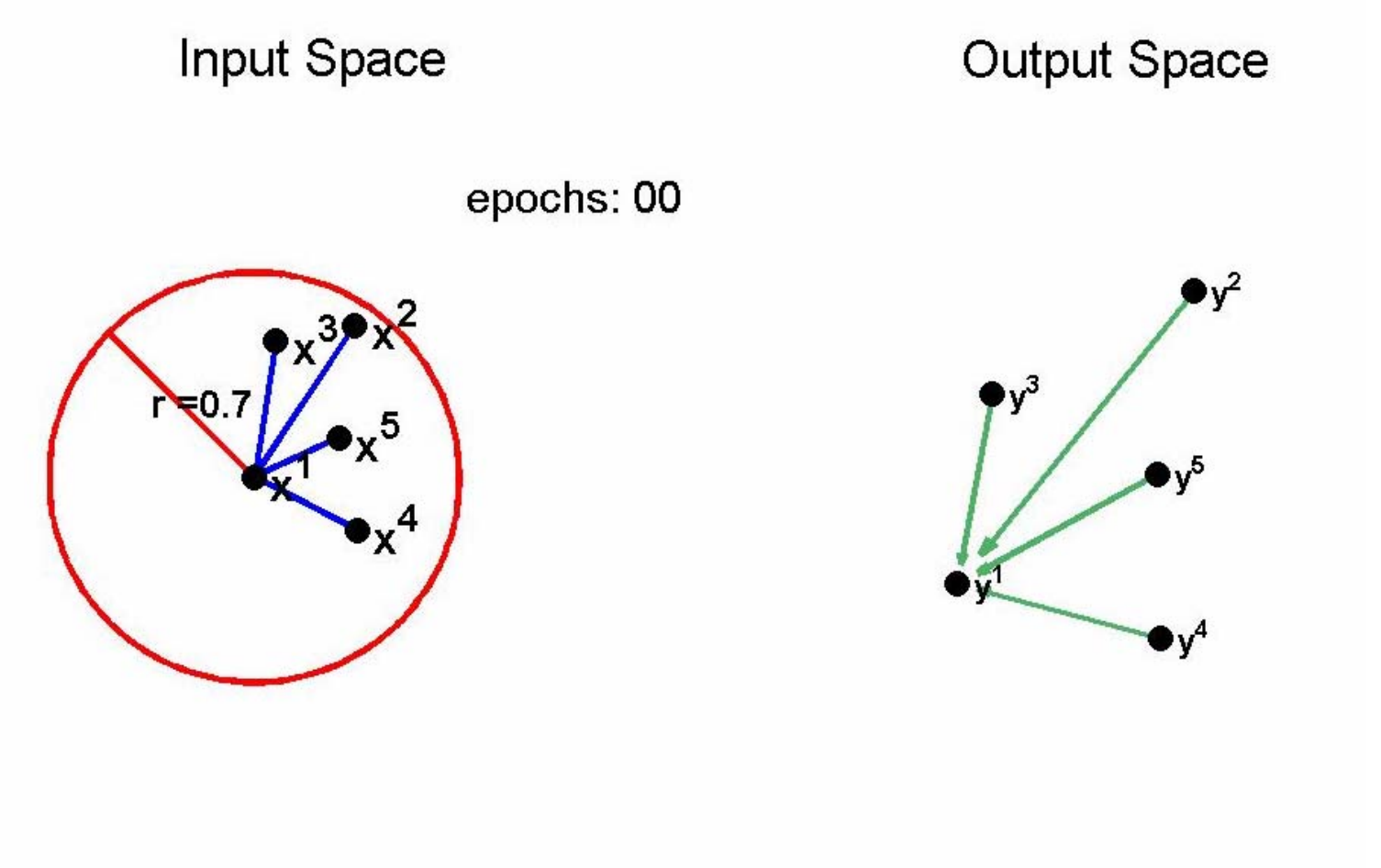
Each epoch


Input Space
Output Space
epochs: 05
r =0.5
x3
x2
x5
x1
x4
y2
y3
y5
y1
y4

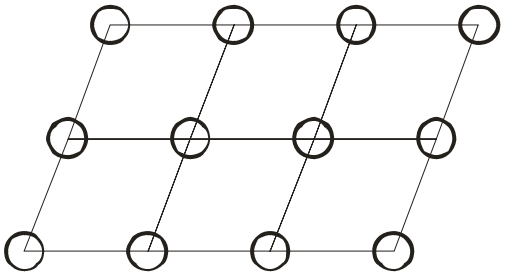
Each pattern



Ten epochs and all patterns

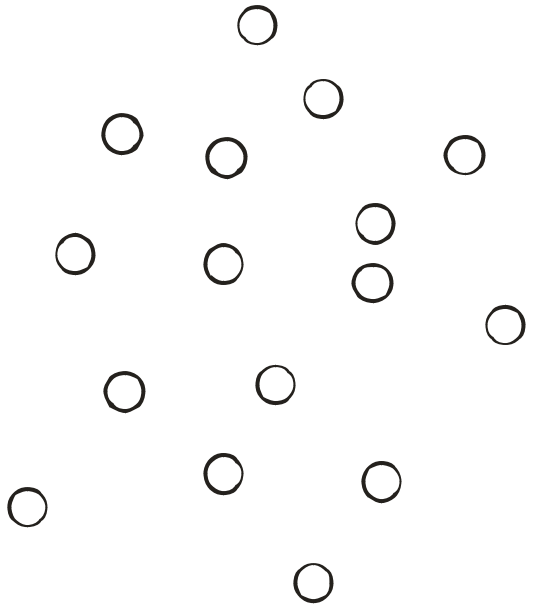


Self-organizing map



N << number of data
Regular cell position
Fixed cell positions

LDP irregular cells

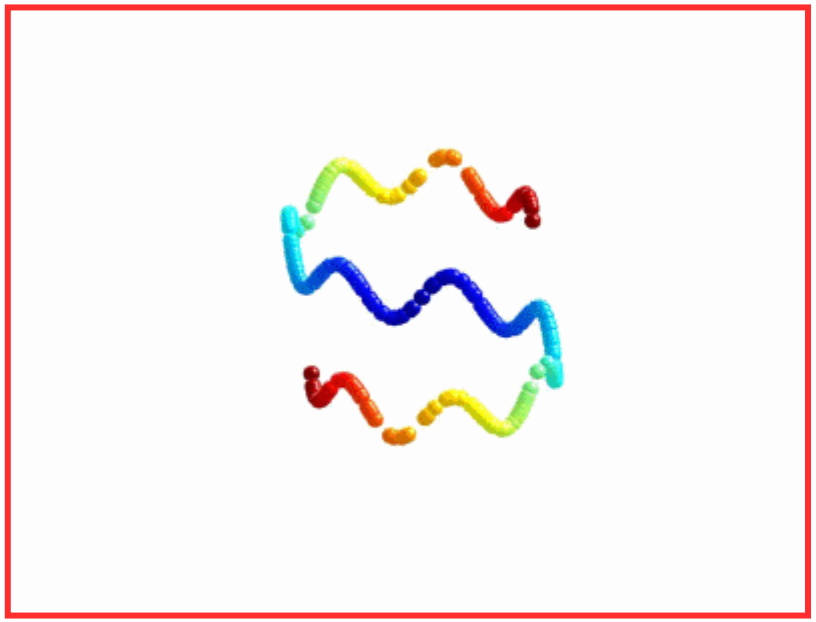


N = number of data
Save $p(x)$ for $p(y)$
Irregular and unfixed cell positions

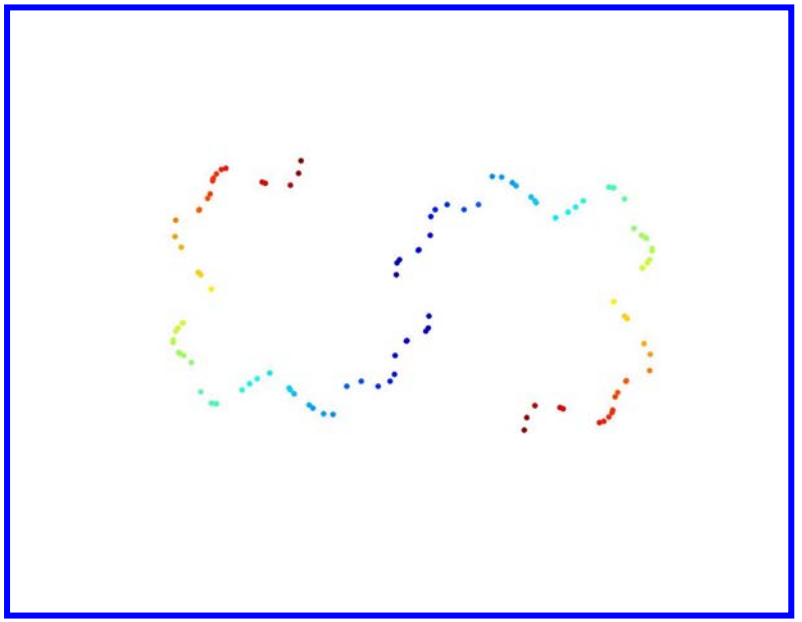
Artificial Data

- There are 280 input patterns in 3D input space.
- Initial the output cells by MDS

Input data

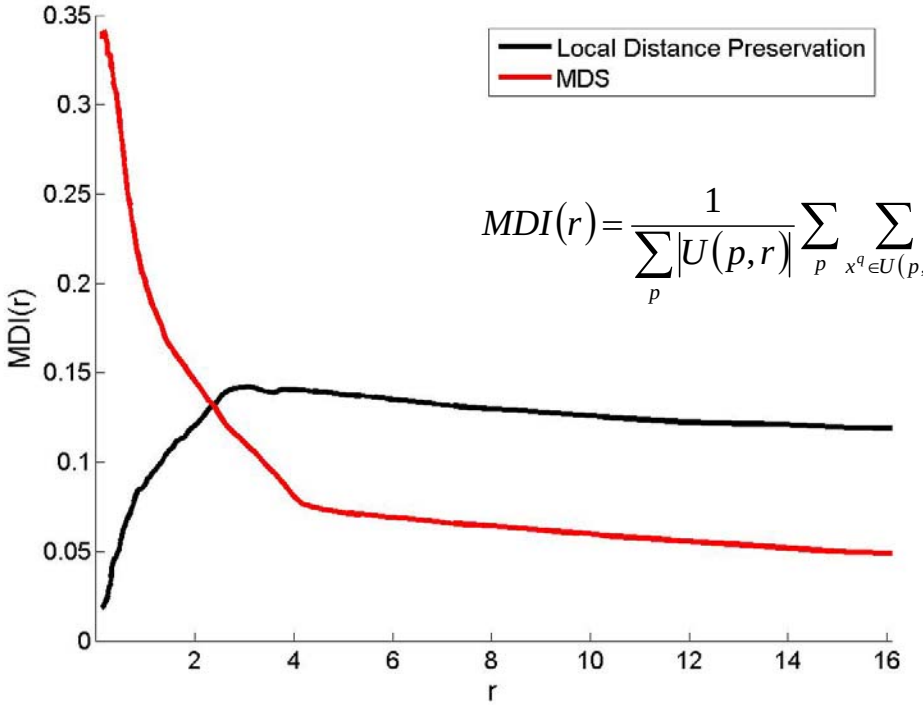


Output data



Artificial Data

- There are 280 input patterns in 3D input space.
- Initial the output cells by MDS
- MDI (measurement of distance invariance)

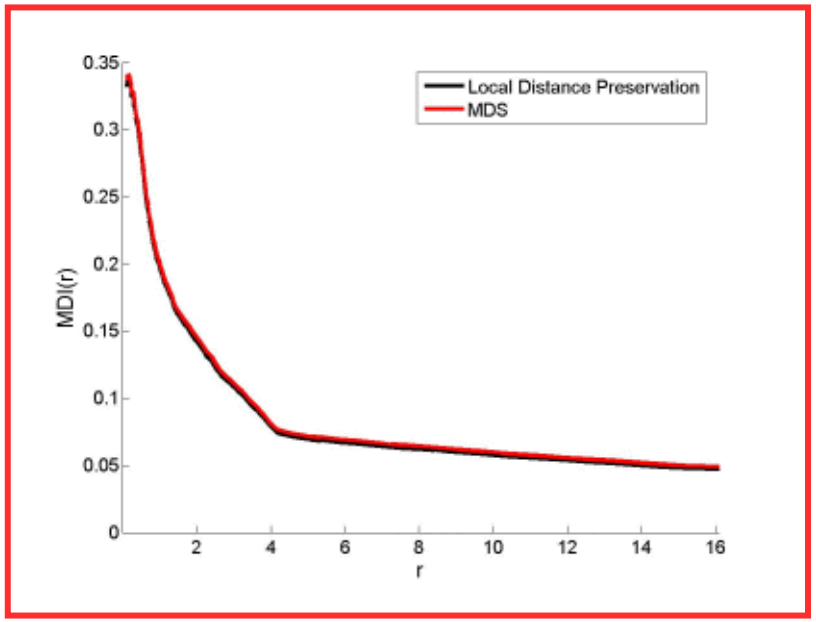


$$MDI(r) = \frac{1}{\sum_p |U(p,r)|} \sum_p \sum_{x^q \in U(p,r)} \frac{\sqrt{\left(\left\| \mathbf{y}^p - \mathbf{y}^q \right\| - \left\| \mathbf{x}^p - \mathbf{x}^q \right\|^2 \right)^2}}{\left\| \mathbf{x}^p - \mathbf{x}^q \right\|}$$

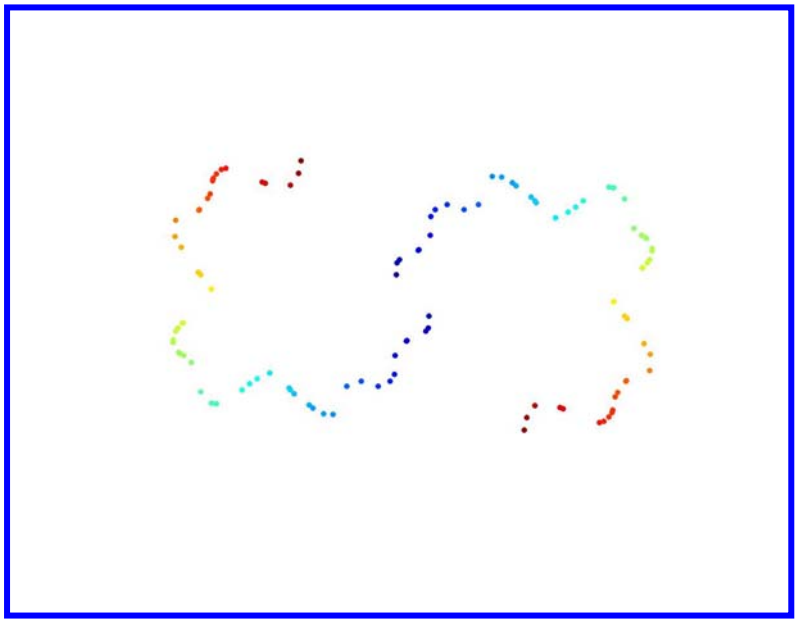
Artificial Data

- There are 280 input patterns in 3D input space.
- Initial the output cells by MDS
- MDI (measurement of distance invariance)

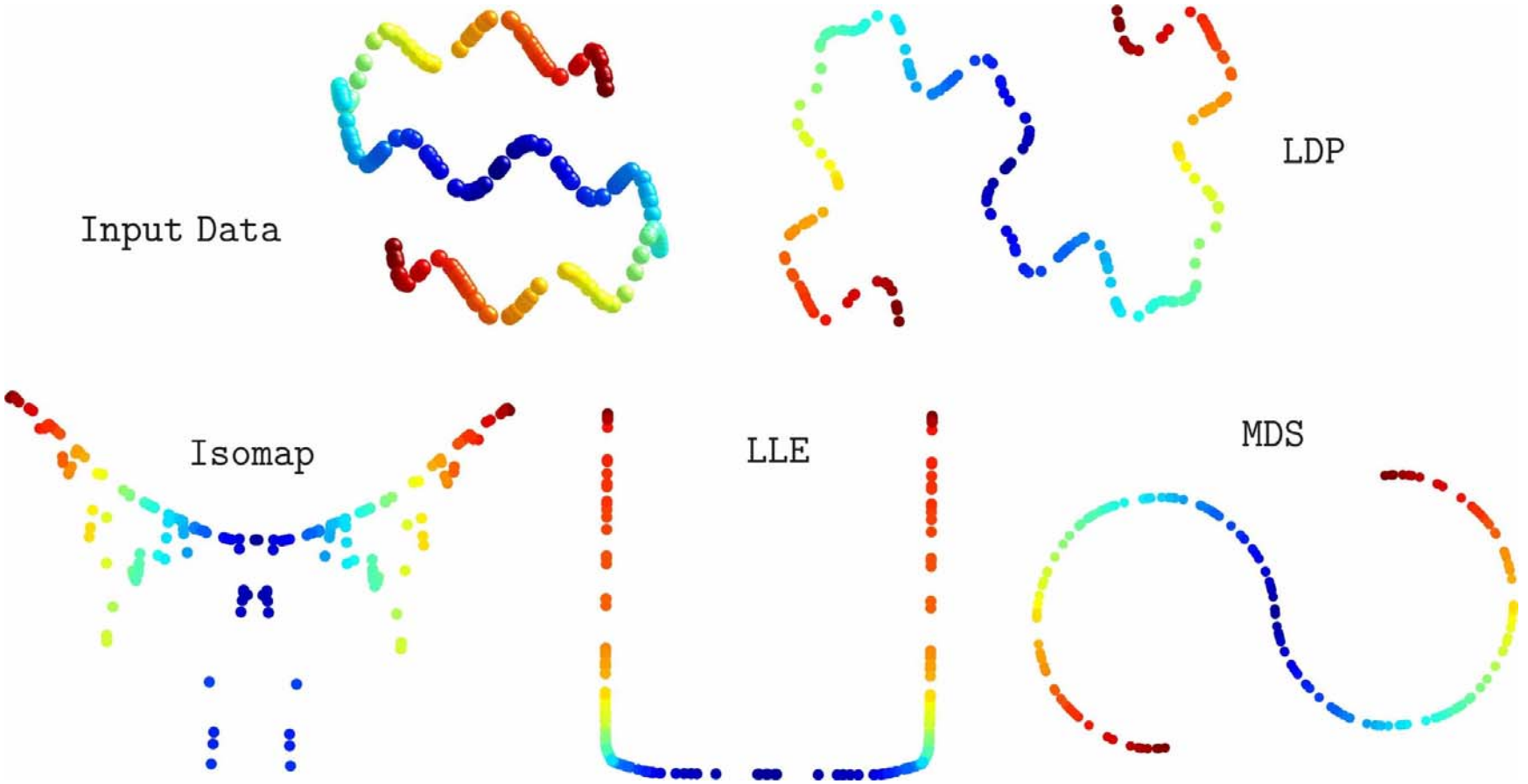
Performance



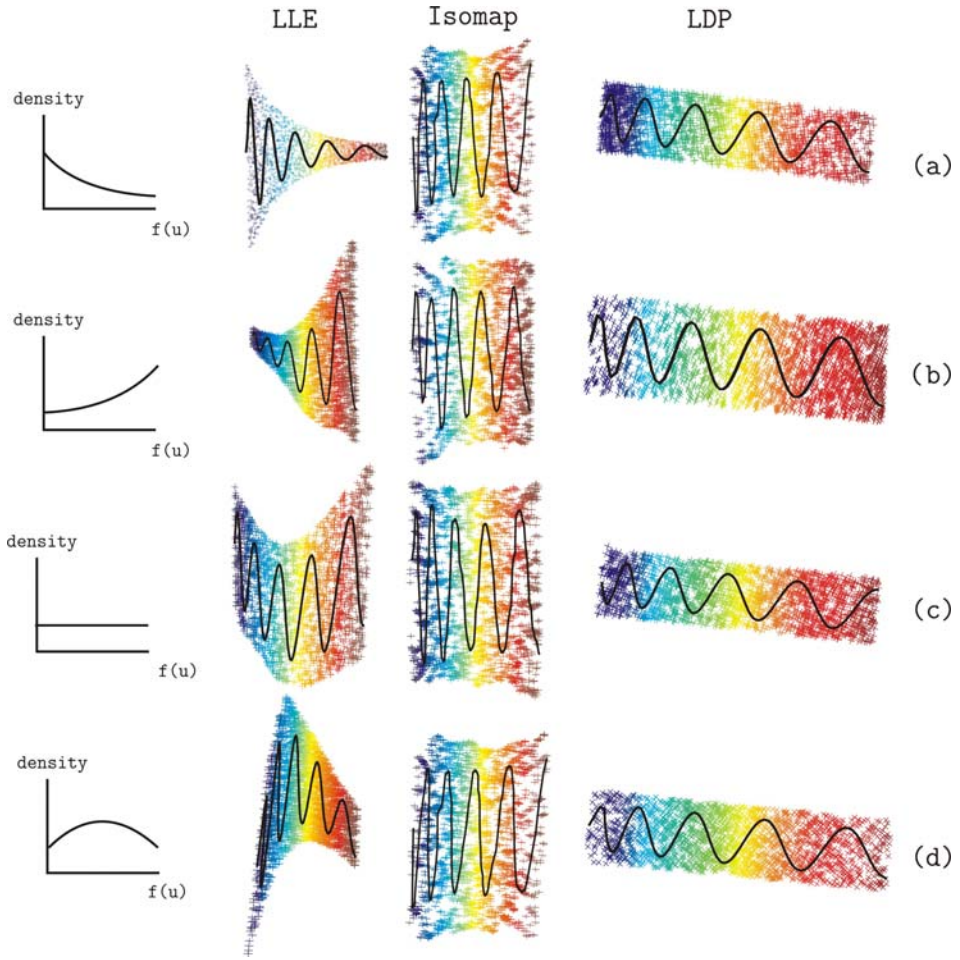
Output data



Comparisons



Sampling densities in swiss roll



Sampling densities in swiss roll

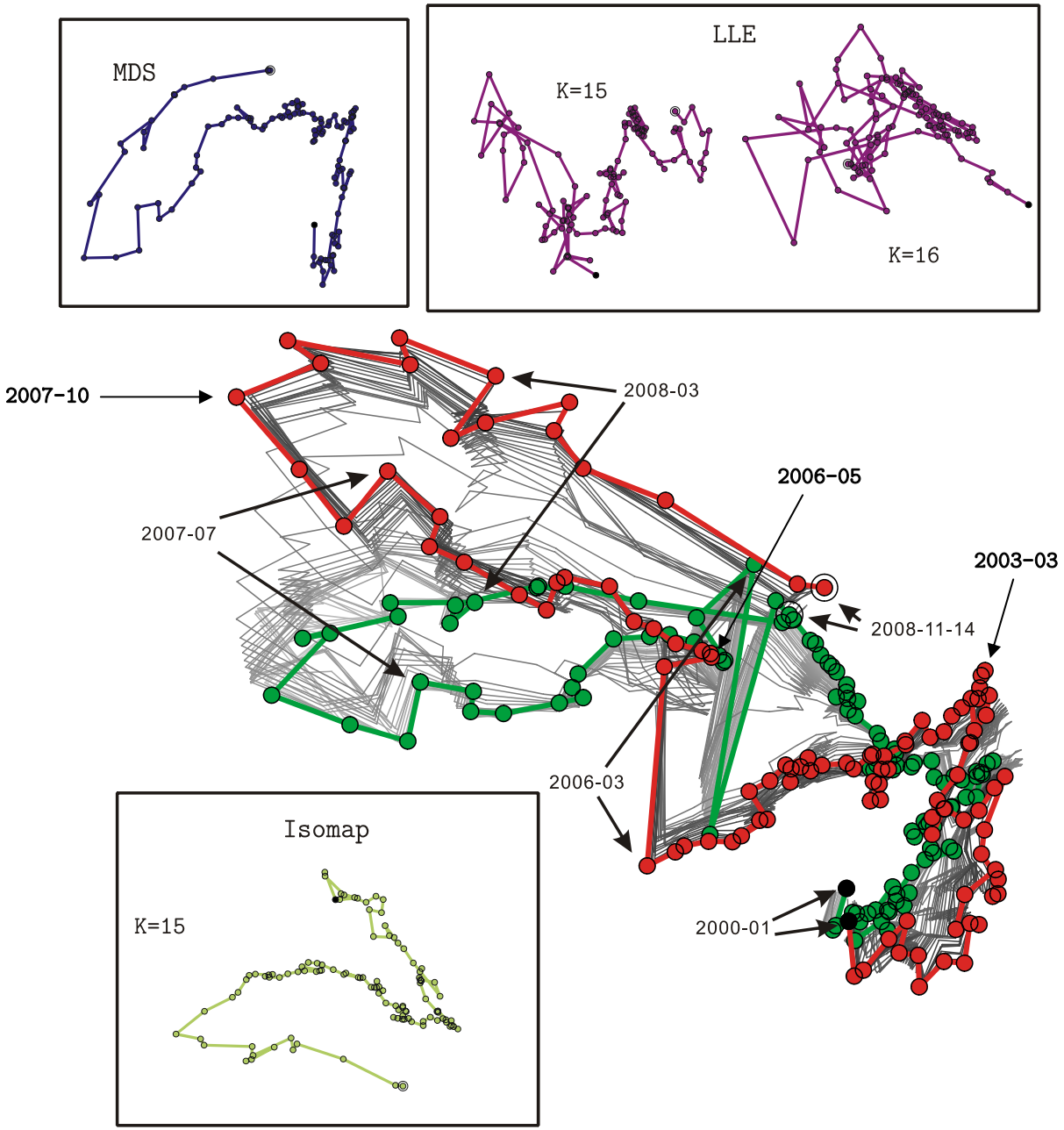
- LDP is not affected by the density distributions.
- This is because every pattern has its correspondent cell in the output space.
- The density of each pattern is equal to the density of its cell.
- The number of cells is equals to the number of patterns.

World Stock Indices

Amsterdam	Australia	Bombay	Frankfurt	New York
AEX	ALL ORDS	BSE SENSEX	DAX	DJ-INDUS
London	**Hong Kong**	**Jakarta**	**Kuala**	**Korea**
FTSE100	HANG SENG	JKSE	KLSE	KOSPI
Milan	**Nasdaq**	**Osaka**	**New York**	**OSLO**
MIBTEL	NASDAQ	NIKKEI 225	NYSE COMP	OBX
New York	**Shanghai**	**Swiss**	**Taiwan**	
S&P 500	SSE	SWISS MARKET	TAIEX	

World Stock Indices

- Data: The indices value of 19 country in each month are arranged in the vector form.
- Initial: the initial values of the output cells are from the first two dimension of MDS.


MDS
LLE
K=15
K=16
2007-10
2008-03
2006-05
2007-07
2003-03
2008-11-14
2006-03
2000-01
Isomap
K=15

Flexible design of mapping fctn y=Ax
Edge length estimation of known tree

All input distance pairs

All tree edges

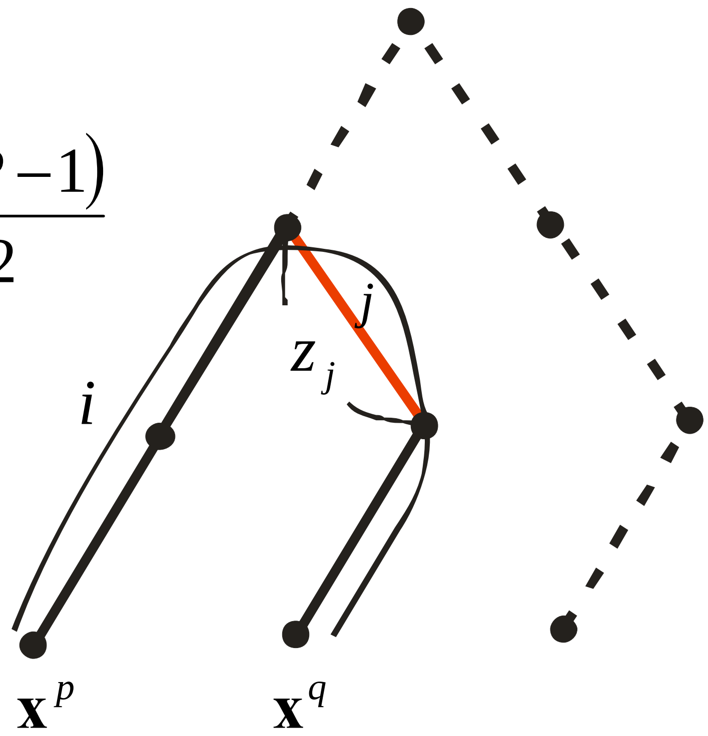
$$\overbrace{\begin{bmatrix} 1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & & \end{bmatrix}}^{2P-2} \begin{bmatrix} z_1 \\ \vdots \\ z_{2P-1} \end{bmatrix} = \left.\begin{bmatrix} \delta_1 \\ \vdots \\ \delta_{\frac{P(P-1)}{2}} \end{bmatrix}\right\} \frac{P(P-1)}{2}$$

(row index i, column index j)

$$A\mathbf{z} \approx \boldsymbol{\delta}$$

subject to $\mathbf{z} \geq 0$

[Sattath, 1977]



$$\delta_i = \left\| \mathbf{x}^p - \mathbf{x}^q \right\|$$

Edge length estimation of known tree

All input distance pairs

All tree edges

j

$2P-2$

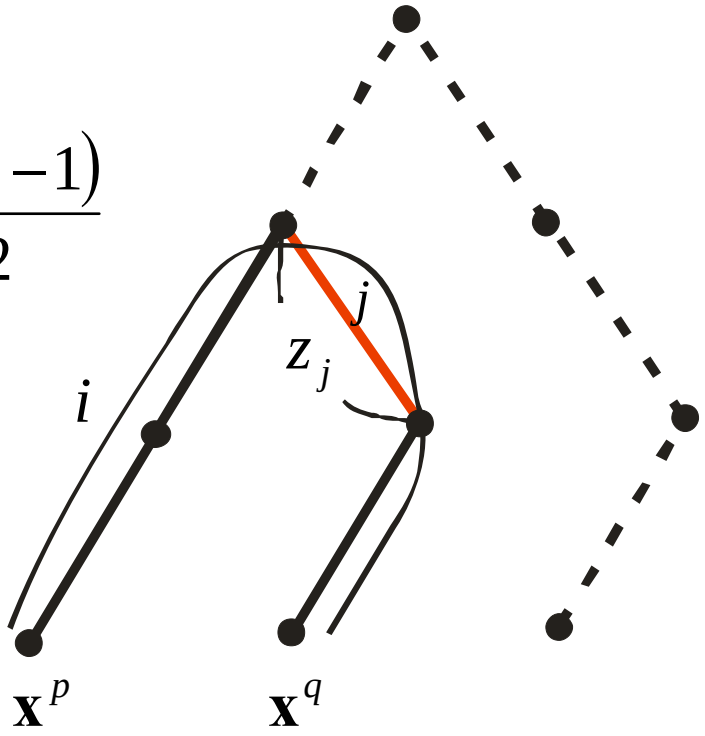
$$i\begin{bmatrix}1 & 0 & \cdots \\ 1 & 1 & \cdots \\ \vdots & & \end{bmatrix}\begin{bmatrix} z_1 \\ \vdots \\ z_{2P-1}\end{bmatrix} = \left.\begin{bmatrix} \delta_1 \\ \vdots \\ \delta_{\frac{P(P-1)}{2}}\end{bmatrix}\right\} \frac{P(P-1)}{2}$$

$$\sum_{j=1}^{2P-1} a_{ij} z_j \approx \delta_i, \text{ if } \delta_i \le r$$

subject to $\mathbf{z} \ge 0$

LDP

(local distance preservation)



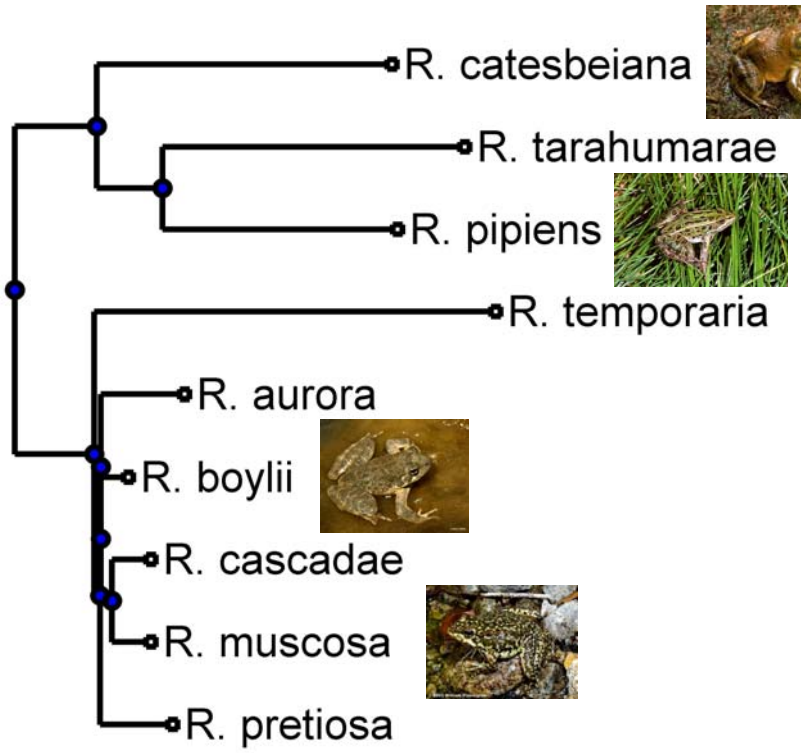
$$\delta_i = \left\| \mathbf{x}^p - \mathbf{x}^q \right\|$$

Phylogenetic Tree

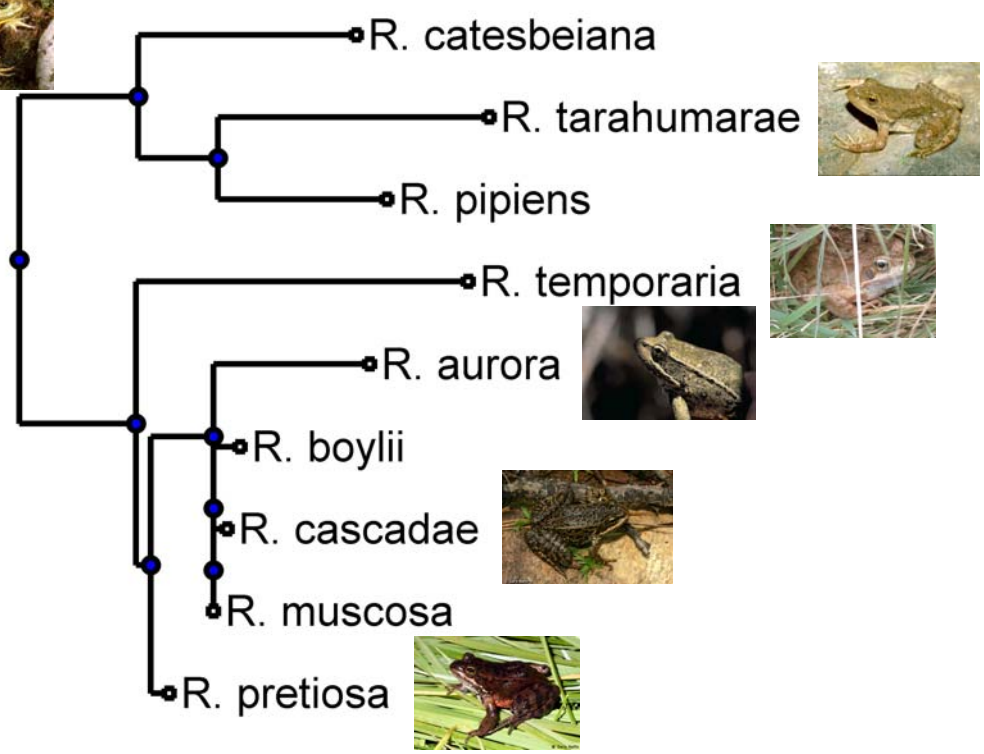
- **Tree:** The tree is constructed by UPGMA (Unweighted Pair Group Method with Arithmetic mean).
- **Path Estimation:** The arc length of the tree is estimated by two different methods for comparison.
  - Non-Negative Least Square Square [Sattath, 1977] Method
  - Local Distance Preservation
- **Data:** Case, S.M.: Biochemical systematics of members of the genus Rana native to western north America. Systematic Zoology 27, 299-311 (1978)

Immunological distance

Local Distance Preservation Method

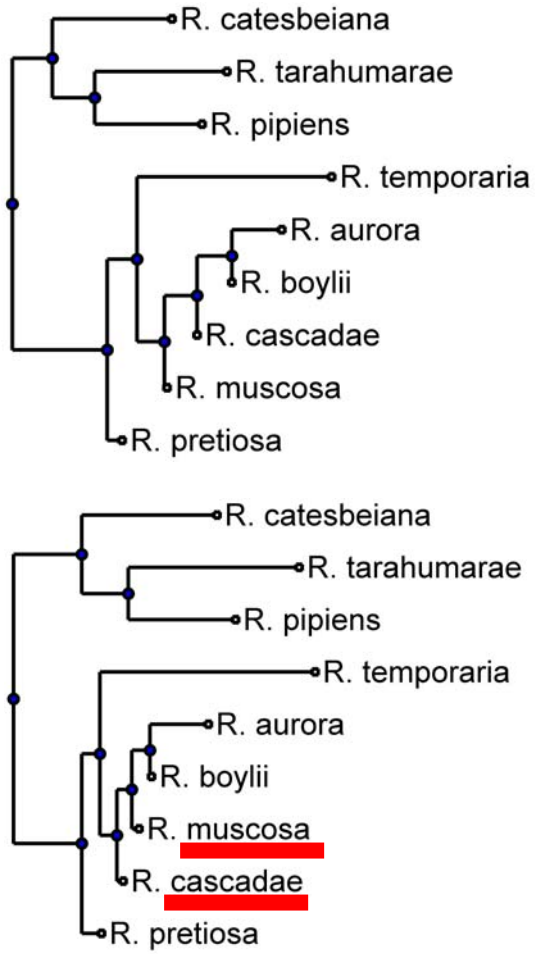


Non-Negative Least Square Square [Sattath, 1977] Method



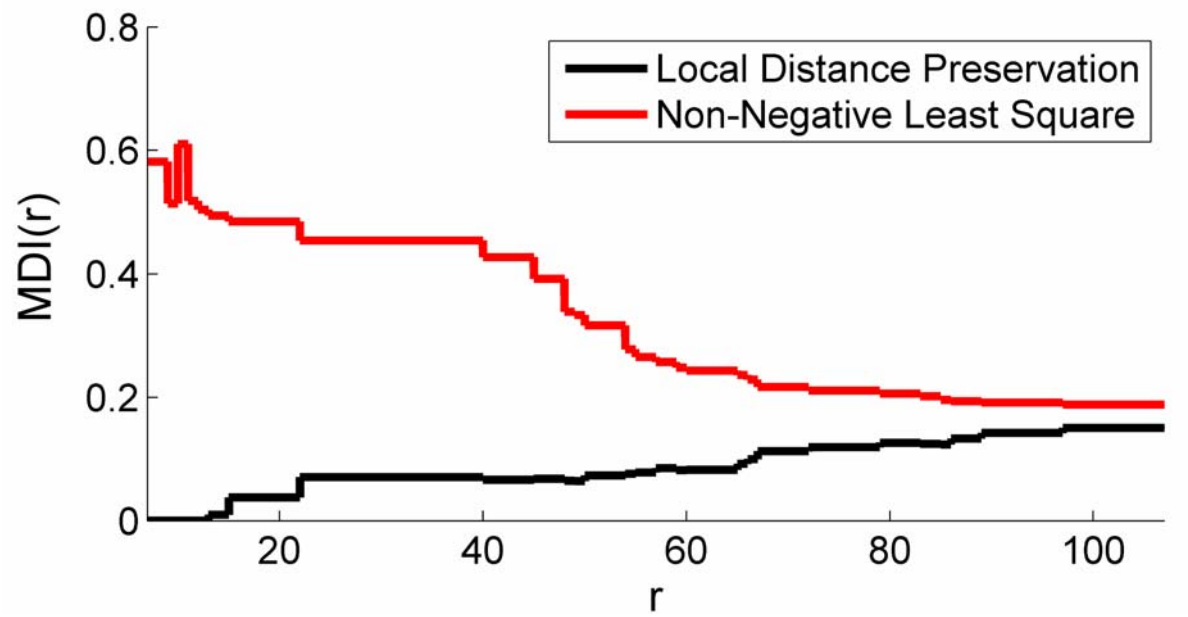
Immunological distance

- Fitch, W.M. 1981, A Non-Sequential Method for Constructing Trees and Hierarchical Classifications, J Mol Evol
- Saitou, N. and Nei, M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Bio and Evol



Immunological distance

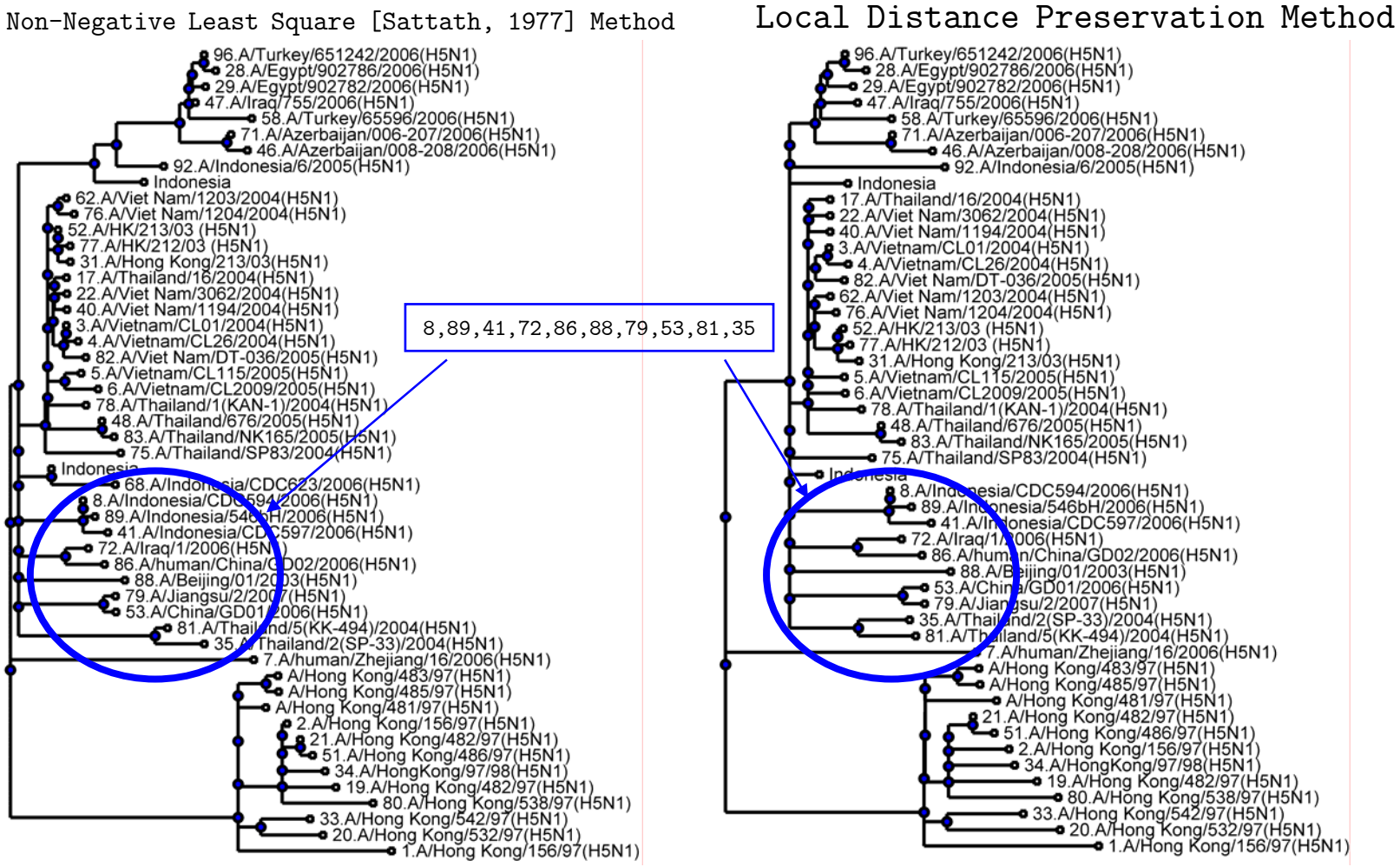
$$MDI(r) = \frac{1}{\sum_p |U(p,r)|} \sum_p \sum_{x^q \in U(p,r)} \frac{\sqrt{\left(t(p,q) - \left\|\mathbf{x}^p - \mathbf{x}^q\right\|\right)^2}}{\left\|\mathbf{x}^p - \mathbf{x}^q\right\|}$$



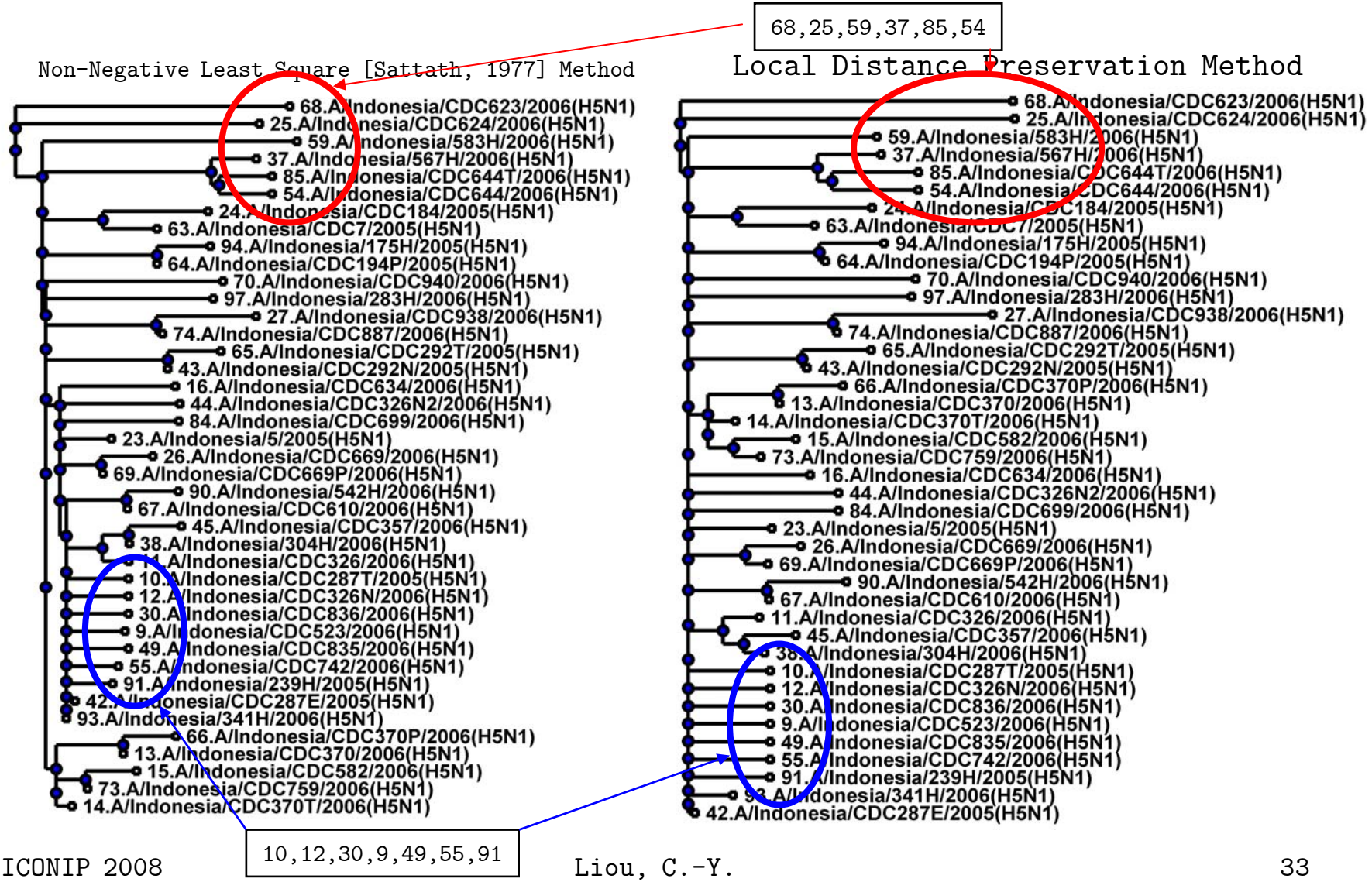
Phylogenetic Tree

- **Tree:** The trees (H5N1, HIV, SARS, Nipah Virus) are constructed by UPGMA (Unweighted Pair Group Method with Arithmetic mean).
- **Path Estimation:** The edge length of the tree is estimated by two different methods for comparison.
  - Non-Negative Least Square Method
  - Local Distance Preservation
- **Assumption:** Local distance is more reliable than global distance.

Phylogenetic Tree



Indonesia subtree



Phylogenetic Tree

- **Distance:** hamming distance after performing multiple alignment
- **Characters:** 20 amino acid
- **# of data:** there are totally 97 H5N1 protein sequences of segment 1 (PB2) in this simulation.
- **Data source:** NCBI Influenza Virus Resource

Summary

- Visualization only, no LVQ
- Flexible design of low dimensional space
  - 2D, tree y=Ax
- Distance invariance
  - Preservation of physical meaning in 2D
  - Perfect energy function
  - Relative distances, no pattern vectors
  - Better edge lengths for short amino acid distances
- number of output cells=P, Save $p(x)$ for $p(y)$
  - no probability manipulation
- Flexible initial setting.
- Parallel and distributed algorithm possible
  - Sequential mode and batch mode

- Let me also emphasize the following facts: 1) The SOM has not been meant for statistical pattern recognition (no probability); It is a clustering, visualization, and abstraction method. Anybody wishing to implement decision and classification processes should use LVQ in stead of SOM. 2)...

Kohonen's book 2001, preface page XI

Thank You

A Novel Method for Manifold Construction

Wei-Chen Cheng and Cheng-Yuan Liou*
Department of Computer Science and Information Engineering
National Taiwan University
Republic of China
*cyliou@csie.ntu.edu.tw

Related Work

Year	People	Contribution
1938	Young, G.	MDS
1989	Kohonen, T.	SOM
2000	Roweis, S.-T.	Locally Linear Embedding
2000	Tenenbaum, J. B.	Isomap