Mathematics of Data
I

姚 远

2011.7.11

Data Science is the science of
Learning from Data.

In mathematics, we look for mappings
from
data domain
to
knowledge domain.

John Dewey

If knowledge comes from the impressions made upon us by natural objects, it is impossible to procure knowledge without the use of objects which impress the mind.

Democracy and Education: an introduction to the philosophy of education, 1916

Statistics has been the most
successful
information science.

Those who ignore Statistics are
condemned
to reinvent it.

---- Bradley Efron

Types of Data

- Data as vectors/matrices/tensors in Euclidean spaces
  - images, videos, speech waves, gene expr., financial data
  - most of statistical methods deal with this type of data
- Data as graphs
  - data as points in metric spaces (molecular dynamics, etc.)
  - internet, biological/social networks
  - data where we just know relations (similarity,…)
  - data visualization
  - modern Computer Science likes this type of data
- and there are more coming …

A Dictionary between Machine Learning vs. Statistics

Machine Learning	Statistics
Supervised Learning	Regression Classification Ranking …
Unsupervised Learning	Dimensionality Reduction Clustering Density estimation …
Semi-supervised Learning	✗

Contributions

- Machine learning and computer science society had proposed many of scalable algorithms to deal with massive data sets
- Those algorithms are often found following consistency theory of statistics
- But sometimes traditional statistics does not work …

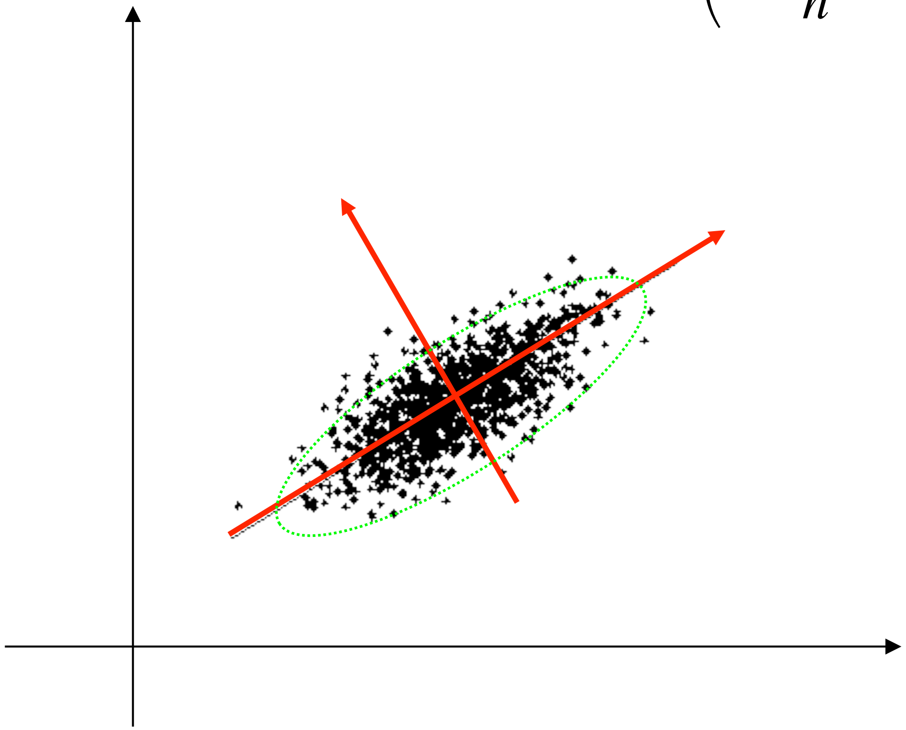
Principal Component Analysis (PCA)

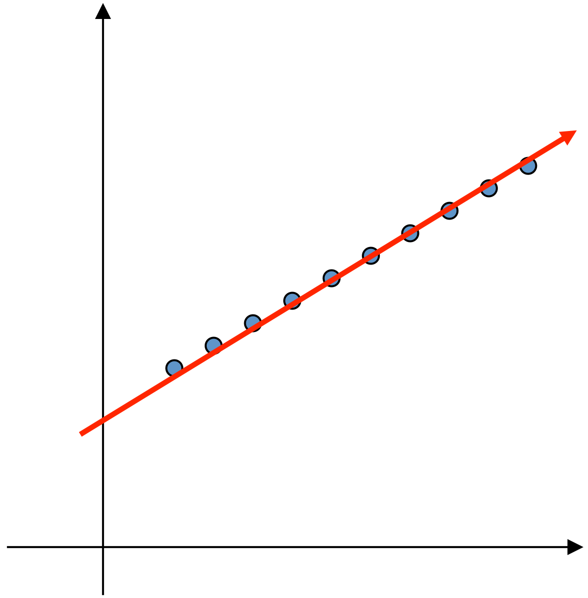
- Principal Component Analysis (PCA)

$$X_{n\times p} = [X_1 \quad X_2 \quad \ldots \quad X_p]$$

$$\Sigma_{ij} = [\mathrm{cov}(X_i, X_j)] = E[(X_i - \mu_i)(X_j - \mu_j)]$$

Eigen-decomposition of $\hat{\Sigma} = X^T\left(I - \frac{1}{n}ee^T\right)^2 X \rightarrow \Sigma$ for fixed p and n->∞





One Dimensional Manifold

PCA may go wrong if p ≥ n

- For *p/n* = $\gamma > 0$, assume *p*-dimensional X_i

$$X_i = \sum_{\nu=1}^{M} \sqrt{\lambda_\nu} u_{\nu i} \theta_\nu + \sigma Z_i, \qquad u_{\nu i} \approx N(0,1), \qquad Z_i \approx N_p(0, I_p)$$

where θ_ν are orthonormal, then PCA is inconsistent by Random Matrix Theory

$$\langle \hat{\theta}_\nu, \theta_\nu \rangle \to \begin{cases} 0 & \lambda_\nu \in [0, \sqrt{\gamma}] \\ \dfrac{1 - \gamma / \lambda_\nu^2}{1 + \gamma / \lambda_\nu} & \lambda_\nu > \sqrt{\gamma} \end{cases}$$

- Phase transition:
  - Below the threshold, estimation is orthogonal to the truth
  - Above the threshold, the angle decreases as eigenvalue grows, but always biased

Johnstone (2006) High Dimensional Statistical Inference and Random Matrices, arxiv.org/abs/math/0611589v1

A Dawn of the Science for High Dimensional Massive Data Sets

- When p>>n, PCA may work with additional requirements
  - Σ is sparse or fast decay
  - Σ^{-1} is sparse
  - … (e.g. see Tony Cai tutorials)
- Geometry and topology begin to enter this Odyssey
  - data concentrate on low-dimensional manifolds …

Geometric and Topological Methods

- Algebraic geometry for graphical models
  - Bernd Sturmfels (UC Berkeley)
  - Mathias Drton (U Chicago)
- Differential geometry for graphical models
  - Shun-ichi Amari (RIKEN, Japan)
  - John Lafferty (CMU)
- Integral geometry for statistical signal analysis
  - Jonathan Taylor (Stanford)
  - Rob Ghrist (UIUC - U Penn)
- Spectral kernel embedding:
  - LLE (Roweis, Saul etal), ISOMAP (Tennenbaum etal), Laplacian eigenmap (Niyogi, Belkin etal), Hessian LLE (Donoho etal), Diffusion map (Coifman, Singer etal)
- Computational topology for data analysis
  - Herbert Edelsbrunner (Duke - Institute of Sci. Tech. Austria), Gunnar Carlsson (Stanford), et al.

Two aspects in those works

Geometry and topology may play a role as

- characterizing local or global constraints (symmetry) in model spaces or data
    - algebraic and differential geometry for graphical models
    - spectral analysis for discrete groups (Risi Kondor)

- characterizing nonlinear distribution (sparsity) of data
    - spectral kernel embedding (nonlinear dimensionality reduction)
    - integral geometry for signal analysis
    - computational topology for data analysis

Let's focus on the second aspect.

Geometric and Topological Data Analysis

- General area of geometric data analysis attempts to give insight into data by imposing a geometry (metric) on it
  - manifold learning: global coordinate preserving local structure
  - metric learning: find a metric accounting for similarity

- Topological method is to study invariants under metric distortion
  - clustering as connected components
  - loops, holes

- Between them, lies in Hodge Theory, a bridge over geometry and topology
  - 0-dimensional Hodge Theory: Laplacian eigenmaps, Diffusion Maps
  - 1-dimensional Hodge Theory: Preference Aggregation, Game Theory

What we'll cover in this course?

I. Geometric Data Analysis: from PCA/MDS to LLE/ISOMAP

II. Geometric Data Analysis: diffusion geometry

III. Topological Data Analysis

IV. Hodge Theory in Data Analysis

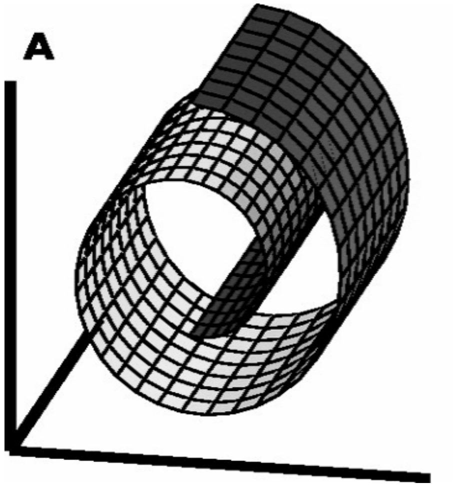
V. Seminar

Manifold learning: Spectral Kernel Embeddings

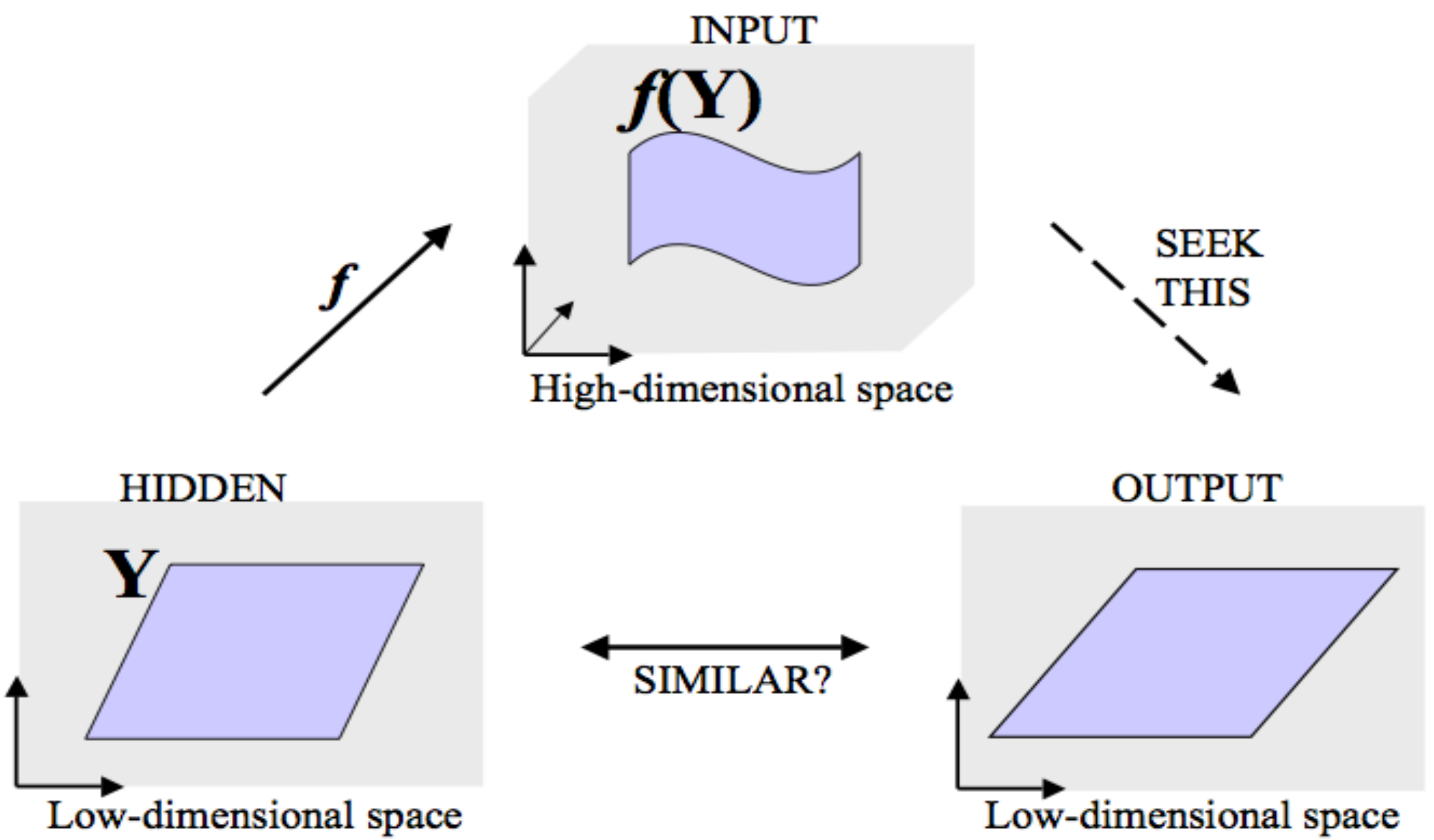
- Principle Component Analysis (PCA)
- Multi-Dimensional Scaling (MDS)
- Locally Linear Embedding (LLE)
- Isometric map (ISOMAP)
- Laplacian Eigenmaps
- Diffusion map
- Local Tangent Space Alignment (LTSA)
- …

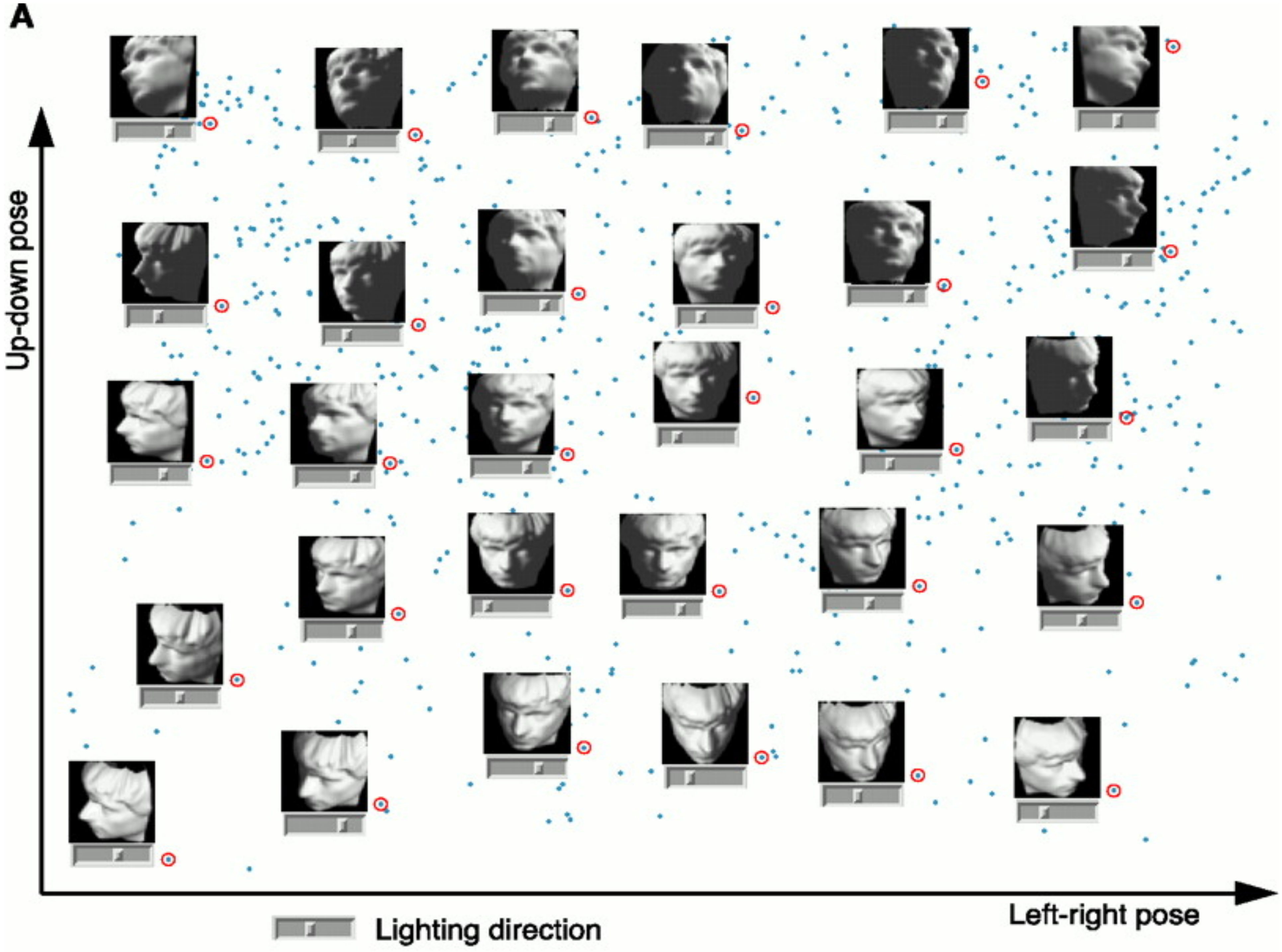
Dimensionality Reduction

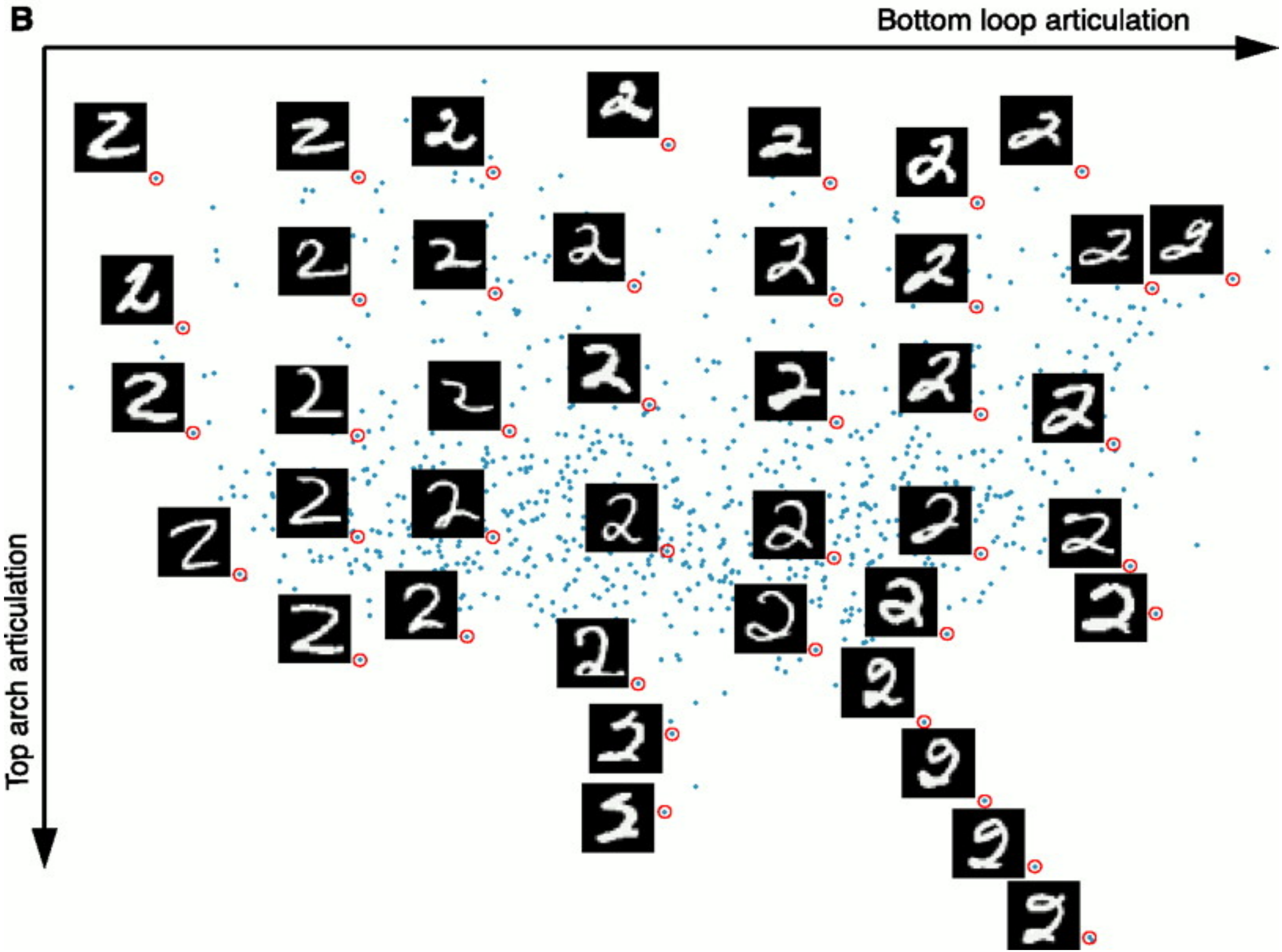
- Data are concentrated around low dimensional Manifolds
  - Linear： MDS, PCA
  - NonLinear： ISOMAP, LLE, …



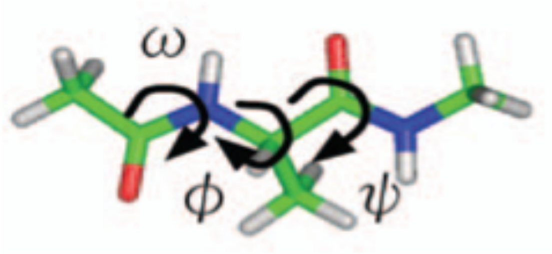
Generative Models in Manifold Learning

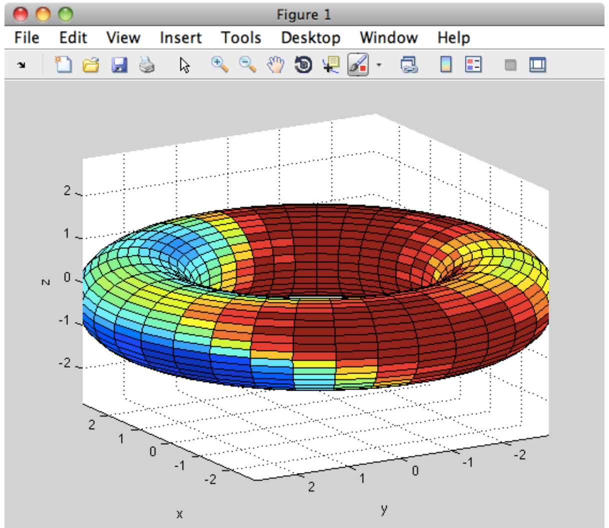


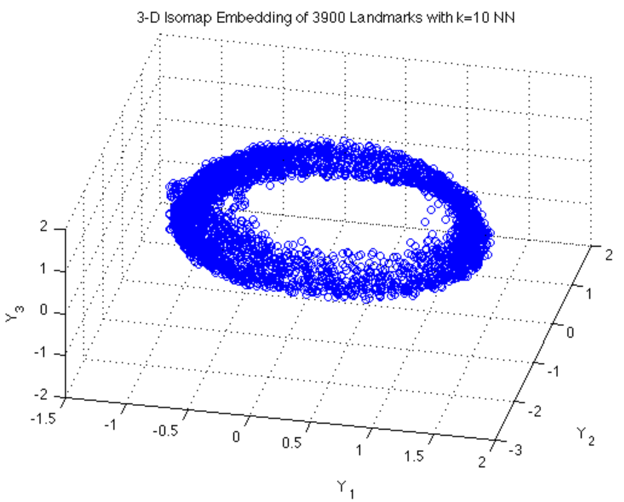

A
Up-down pose
Left-right pose
Lighting direction

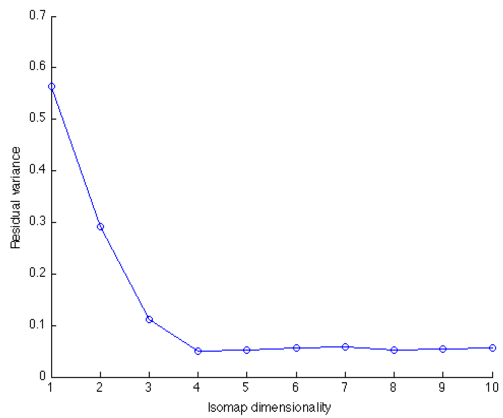

B
Bottom loop articulation
Top arch articulation

Biomolecular: Alanine-dipeptide









ISOMAP 3D embedding with RMSD metric on 3900 Kcenters

Distances & Mappings

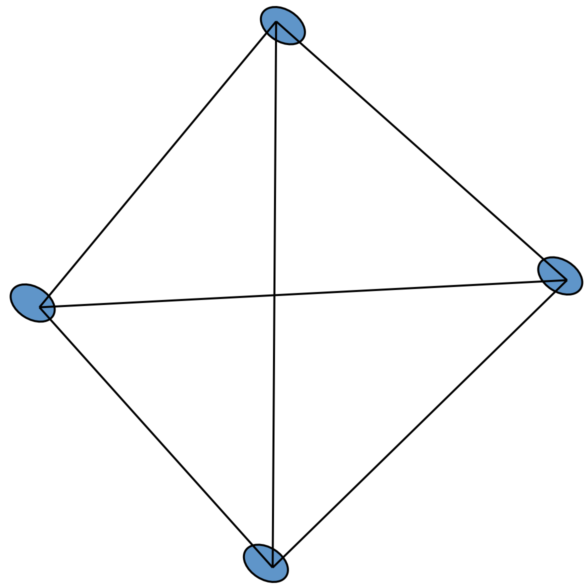
- Given an Euclidean embedding, it's easy to calculate the distances between the points :

$$d_{j,k} = \sqrt{\sum_a (x_{ja} - x_{ka})^2}$$

- Multi-Dimensional Scaling (MDS) operates the other way round:
  - Given the "distances" [data] find the embedding map [configuration] which generated them
  - MDS can do so when all but ordinal information has been jettisoned (fruit of the "non-metric revolution")
  - even when there are missing data and in the presence of considerable "noise"/error (MDS is robust).

PCA => MDS

- Here we are given pairwise distances instead of the actual data points.
  - First convert the pairwise distance matrix into the dot product matrix XX^T
  - After that same as PCA.

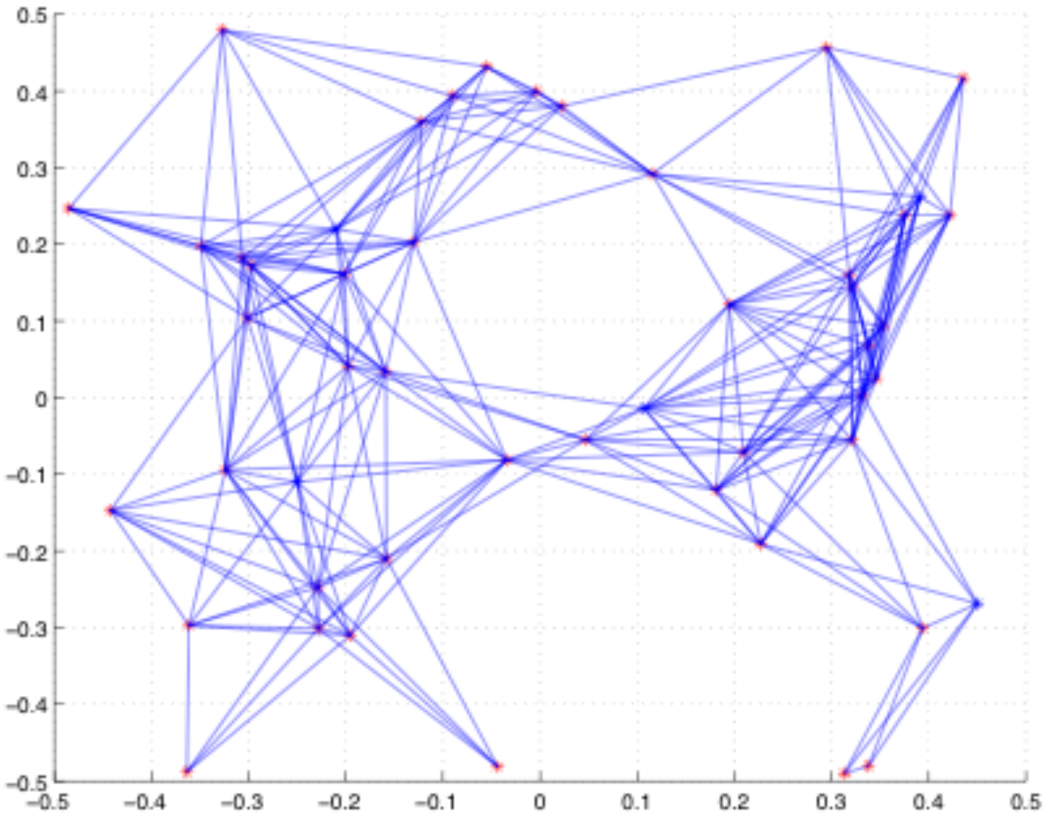


If we preserve the pairwise distances do we preserve the structure??

Matlab Commands

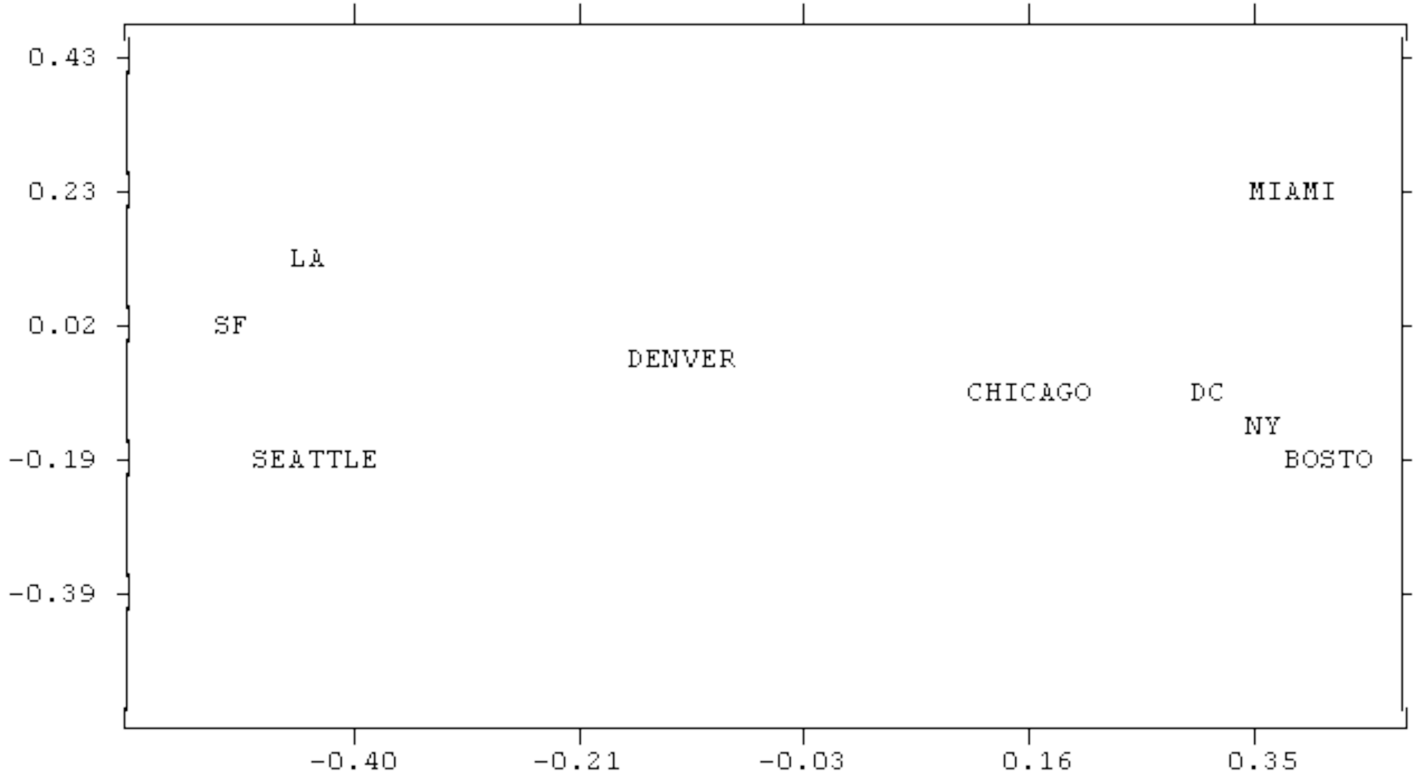
- STATS toolbox provides the following:
  - cmdscale - classical MDS
  - mdscale – nonmetric MDS

Example I: 50-node 2-D Sensor Network Localiztion



Example II City Map

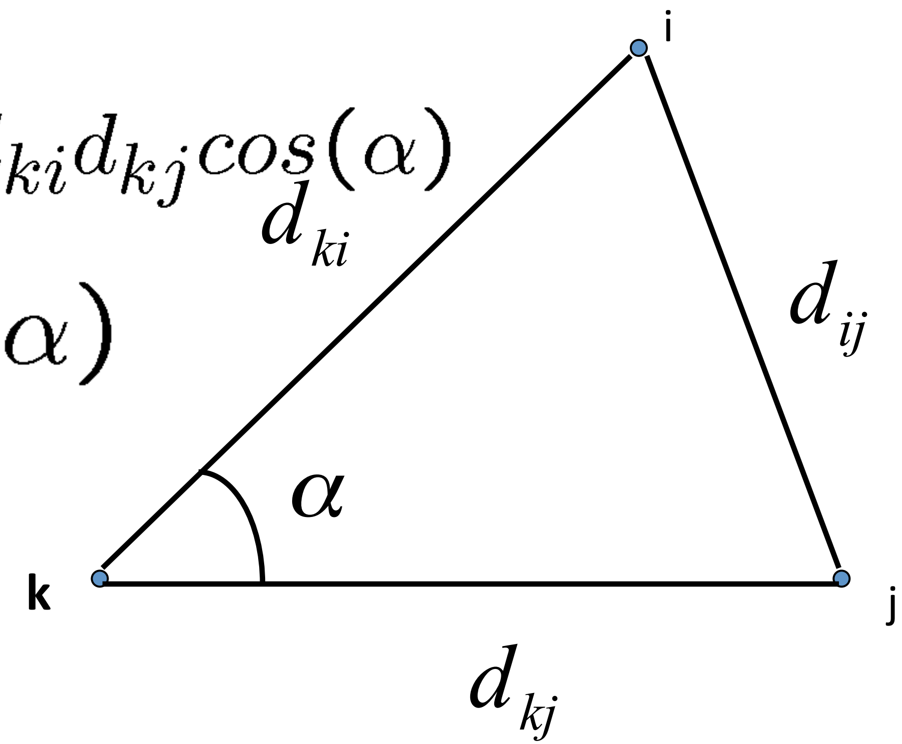
		1 BOST	2 NY	3 DC	4 MIAM	5 CHIC	6 SEAT	7 SF	8 LA	9 DENV
1	BOSTON	0	206	429	1504	963	2976	3095	2979	1949
2	NY	206	0	233	1308	802	2815	2934	2786	1771
3	DC	429	233	0	1075	671	2684	2799	2631	1616
4	MIAMI	1504	1308	1075	0	1329	3273	3053	2687	2037
5	CHICAGO	963	802	671	1329	0	2013	2142	2054	996
6	SEATTLE	2976	2815	2684	3273	2013	0	808	1131	1307
7	SF	3095	2934	2799	3053	2142	808	0	379	1235
8	LA	2979	2786	2631	2687	2054	1131	379	0	1059
9	DENVER	1949	1771	1616	2037	996	1307	1235	1059	0



How to get dot product matrix from pairwise distance matrix?

$$d_{ij}^2 = d_{ki}^2 + d_{kj}^2 - 2d_{ki}d_{kj}cos(\alpha)$$

$$b_{ij} = d_{ki}d_{kj}cos(\alpha)$$



$$b_{ij} = \frac{1}{2}(d_{ki}^2 + d_{kj}^2 - d_{ij}^2)$$

Origin centered MDS

- MDS—origin as one of the points and orientation arbitrary.

Centroid as origin

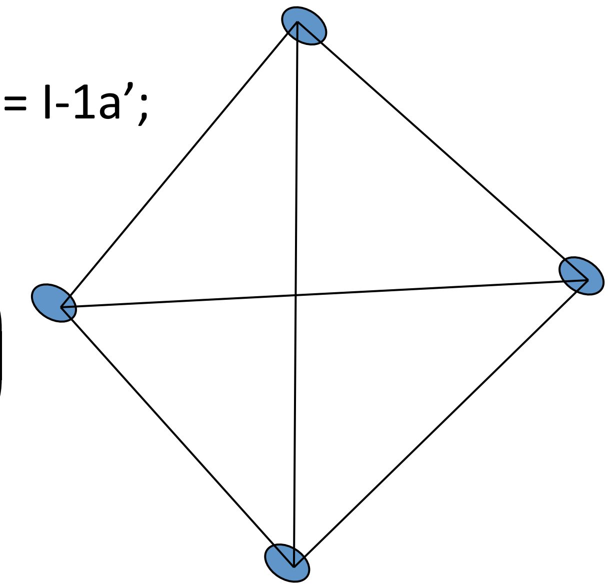
$$b^*_{ij} = -\frac{1}{2}\left[d^2_{ij} - \frac{1}{N}\sum_{l=1}^{N} d^2_{il} - \frac{1}{N}\sum_{m=1}^{N} d^2_{mj} + \frac{1}{N^2}\sum_{o=1}^{N}\sum_{p=1}^{N} d^2_{op}\right]$$

MDS Summary

- Given pairwise distances *D,* where $D_{ij} = d_{ij}^2$, the squared distance between point i and j
  - Convert the pairwise distance matrix D (c.n.d.) into the dot product matrix B (p.s.d.)
    - B_{ij} (a) = -.5 H(a) D H'(a), Hölder matrix H(a) = I-1a';
    - a = 1_k: $B_{ij} = -.5\ (D_{ij} - D_{ik} - D_{jk})$
    - a = 1/n:

$$B_{ij} = -\tfrac{1}{2}\left(D_{ij} - \tfrac{1}{N}\sum_{s=1}^{N} D_{sj} - \tfrac{1}{N}\sum_{t=1}^{N} D_{it} + \tfrac{1}{N^2}\sum_{s,t=1}^{N} D_{st} \right)$$

  - Eigendecomposition of $B = YY^T$



If we preserve the pairwise Euclidean distances do we preserve the structure??

Theory of Classical MDS: a few concepts

- An n-by-n matrix C is positive semi-definite (psd) if for all $v \in R^n$, $v'Cv \geq 0$.

- An n-by-n matrix C is conditionally negative definite (c.n.d) if for all $v \in R^n$ such that $\text{sum}_i\, v_i = 0$, $v'Cv \leq 0$.

Young-Householder-Shoenberg Lemma

- Let x be a signed distribution, i.e. x obeying $\mathrm{sum}_i\, x_i=1$ while x_i can be negative
- Householder centering matrix: $H(x) = I - 1x'$;
- Define $B(x) = -1/2\, H(x)\, C\, H'(x)$, for any C
- Theorem [Young/Householder, Schoenberg 1938b] For any signed distribution x,

B(x) p.s.d. iff C c.n.d.

Proof

- "$\Leftarrow$" first observe that if B(x) is p.s.d., then B(y) is also p.s.d. for any other signed distribution y, in view of the identity B(y) = H(y)B(x)H'(y), itself a consequence of H(y) = H(y)H(x). Also, for any z, z'B(x)z = −y'Cy/2, where the vector y = H'(x)z obeys $\mathrm{sum}_i\, y_i = 0$ for any z, showing necessity.
- "$\Rightarrow$" Also, y = H'(x)y whenever $\mathrm{sum}_i\, y_i = 0$, and hence y'B(x)y = −y'Cy/2, thus demonstrating sufficiency.

Classical MDS Theorem

- Let D be n-by-n symmetric matrix. Define a zero diagonal matrix C to be $C_{ij} = D_{ij} - 0.5\ D_{ii} - 0.5\ D_{jj}$. Then we have
  - $B(x) := -0.5\ H(x)\ D\ H'(x) = -0.5\ H(x)\ C\ H'(x)$
  - $C_{ij} = B_{ii}(x) + B_{jj}(x) - 2B_{ij}(x)$
  - D is c.n.d. iff C is c.n.d.
  - If C is c.n.d., then C is isometrically embeddable, i.e. $C_{ij} = \mathrm{sum}_k(Y_{ik} - Y_{jk})^2$ where

$Y = U\,\Lambda^{1/2}$, with eigendecomp $B(x) = U\,\Lambda\,U'$

Proof

- the first identity in follows from $H(x)1 = 0$;
- the second one from $B_{ii}(x) + B_{jj}(x) - 2B_{ij}(x) = C_{ij} - 0.5C_{ii} - 0.5C_{jj}$, itself a consequence of the definition $B_{ij}(x) = -0.5C_{ij} + \gamma_i + \gamma_j$ for some vector γ;
- the next assertion follows from $u'Du = u'Cu$ whenever $\text{sum}_i\, u_i = 0$;
- the last one can be shown to amount to the second identity by direct substitution.

Remarks

- P.S.D. B(x) (or c.n.d. C) defines a unique squared Euclidean distance D, which satisfies:
- $B_{ij}(x) = -0.5\ (D_{ij} - D_{ik} - D_{jk})$, with freedom x to choose center
- B = Y Y', the scalar product matrix of n-by-d Euclidean coordinate matrix Y

Gaussian Kernels

- Theorem. Let D_{ij} be a squared Euclidean distance. Then for any $\lambda \geq 0$, $B_{ij}(\lambda)=\exp(-\lambda D_{ij})$ is p.s.d., and $C_{ij}(\lambda) = 1 - \exp(-\lambda D_{ij})$ is a squared Euclidean distance (c.n.d. with zero diagonal).

- So Gaussian kernels are p.s.d. and 1 – gaussian kernel is a squared Euclidean distance.

Schoenberg Transform

- A Schoenberg transformation is a function ϕ (D) from R+ to R+ of the form (Schoenberg 1938a)

$$\phi(d) = \int_0^\infty \frac{1-\exp(-\lambda d)}{\lambda} g(\lambda) d\lambda$$

- where g(λ)dλ is a non-negative measure on [0, ∞) such that

$$\int_1^\infty \frac{g(\lambda)}{\lambda} d\lambda < \infty$$

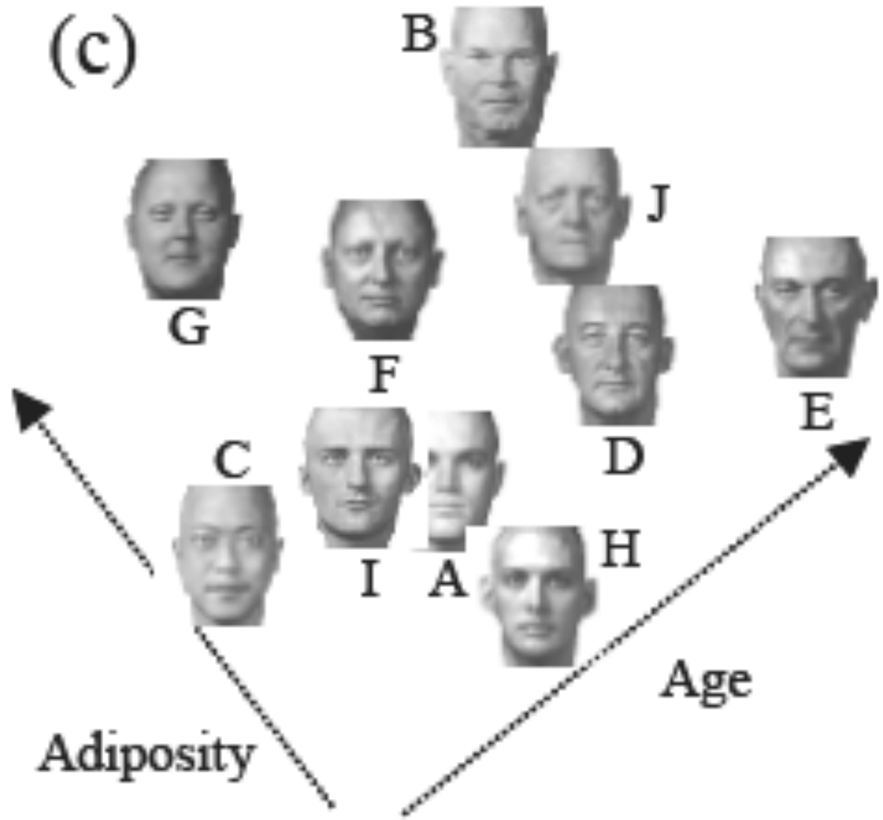
Schoenberg Theorem [1938a]

- Let D be a n x n matrix of squared Euclidean distances. Define the components of the n x n matrix C_{ij} as $C_{ij} = \phi(Dij)$. Then C is a squared Euclidean distance iff ϕ is the Schoenberg Transformation,

$$\phi(d) = \int_{0}^{\infty} \frac{1-\exp(-\lambda d)}{\lambda} g(\lambda) d\lambda$$

Extension I: Nonmetric MDS

- Instead of pairwise distances we can use paiwise “dissimilarities”.
- When the distances are Euclidean MDS is equivalent to PCA.
- Eg. Face recognition, wine tasting
- Can get the significant cognitive dimensions.



Extension II: MDS with missing values (Graph Realization Problem)

Given a graph $G = (V, E)$ and sets of non–negative weights, say $\{d_{ij} : (i,j) \in E\}$ on edges, the goal is to compute a realization of G in the Euclidean space $\mathbf{R}^d$ for a given low dimension d. That is,

- to place the vertexes of G in $\mathbf{R}^d$ such that
- the Euclidean distance between a pair of adjacent vertexes (i,j) equals to (or bounded by) the prescribed weight $d_{ij} \in E$.

Classical MDS (complete graph with squared distance d_{ij}):
Young/Householder 1938, Shoenberg 1938

MDS with Uncertainty
Quadratic equality and inequality system

Given graph $G = (V, E)$ and $d_{ij} \in E$, find $\mathbf{x}_j \in \mathbf{R}^d$ such that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad d_{ij}^2, \; \forall \, (i, j) \in E, \; i < j.$$

MDS in Sensor Network Localization: anchor points

Or given $\mathbf{a}_k \in \mathbf{R}^d$, $d_{ij} \in N_x$, and $\hat{d}_{kj} \in N_a$, find $\mathbf{x}_i \in \mathbf{R}^d$ such that

$$
\begin{aligned}
\|\mathbf{x}_i - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad & d_{ij}^2, \ \forall\, (i,j) \in N_x, \ i < j, \\
\|\mathbf{a}_k - \mathbf{x}_j\|^2 \quad (\leq) = (\geq) \quad & \hat{d}_{kj}^2, \ \forall\, (k,j) \in N_a;
\end{aligned}
$$

that is, edge (ij) (or (kj)) connects sensors i and j (or anchor k and sensor j) with the Euclidean length equal to d_{ij} (or $\hat{d}_{kj}$).

Key Problems

Recall the system:

$$\begin{aligned} \|\mathbf{x}_i - \mathbf{x}_j\|^2 &= d_{ij}^2, \ \forall\, (i,j) \in N_x, \ i < j, \\ \|\mathbf{a}_k - \mathbf{x}_j\|^2 &= \hat{d}_{kj}^2, \ \forall\, (k,j) \in N_a, \end{aligned}$$

- Does the system have a localization or realization of all $\mathbf{x}_j$'s?
- Is the localization unique or the framework $(G, D, \mathbf{x})$ is rigid, and it can be certified?
- Is the system partially localizable or rigid with a certification?

Nonlinear Least Squares

$$\min_{\mathbf{x}} \quad \sum_{(i,j)\in N_x} (\|\mathbf{x}_i - \mathbf{x}_j\|^2 - d_{ij}^2)^2 + \sum_{(k,j)\in N_a} (\|\mathbf{a}_k - \mathbf{x}_j\|^2 - \hat{d}_{kj}^2)^2$$

Or

$$\min_{\mathbf{x}} \quad \sum_{(i,j)\in N_x} (\|\mathbf{x}_i - \mathbf{x}_j\| - d_{ij})^2 + \sum_{(k,j)\in N_a} (\|\mathbf{a}_k - \mathbf{x}_j\| - \hat{d}_{kj})^2.$$

A difficult global optimization problem.

Matrix Represenation I

For simplicity, let $d = 2$ and $X = [\mathbf{x}_1\ \mathbf{x}_2\ ...\ \mathbf{x}_n]$ be the $2 \times n$ matrix that needs to be determined and $\mathbf{e}_j$ be the vector of all zero except 1 at the jth position. Then

$$\mathbf{x}_i - \mathbf{x}_j = X(\mathbf{e}_i - \mathbf{e}_j) \quad \text{and} \quad \mathbf{a}_k - \mathbf{x}_j = [I \quad X](\mathbf{a}_k; -\mathbf{e}_j)$$

so that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = (\mathbf{e}_i - \mathbf{e}_j)^T X^T X(\mathbf{e}_i - \mathbf{e}_j)$$

$$\|\mathbf{a}_k - \mathbf{x}_j\|^2 = (\mathbf{a}_k; -\mathbf{e}_j)^T [I \quad X]^T [I \quad X](\mathbf{a}_k; -\mathbf{e}_j) =$$

$$(\mathbf{a}_k; -\mathbf{e}_j)^T \begin{pmatrix} I & X \\ X^T & X^T X \end{pmatrix} (\mathbf{a}_k; -\mathbf{e}_j).$$

Matrix Represenation II

Or, equivalently,

$$
\begin{aligned}
(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)^T \bullet Y &= d_{ij}^2, \ \forall\ i,j \in N_x,\ i < j, \\
(\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T \bullet \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix} &= \hat{d}_{kj}^2, \ \forall\ k,j \in N_a, \\
Y &= X^T X.
\end{aligned}
$$

SDP Relaxation [Biswas-Ye 2004]

Relax

$$Y = X^T X$$

to

$$Y \succeq X^T X;$$

or equivalently to

$$Z := \begin{pmatrix} I & X \\ X^T & Y \end{pmatrix} \succeq \mathbf{0}.$$

SDP Standard Form [Biswas-Ye 2004]

Find a symmetric matrix $Z \in \mathcal{S}^{(n+2)}$ such that

$$
\begin{aligned}
Z_{1:2,1:2} &= I \\
(\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)(\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)^T \bullet Z &= d_{ij}^2, \ \forall\ i, j \in N_x, \ \ i < j, \\
(\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T \bullet Z &= \hat{d}_{kj}^2, \ \forall\ k, j \in N_a, \\
Z &\succeq \mathbf{0}.
\end{aligned}
$$

- This is an SDP problem,
- if every sensor point is connected, directly or indirectly, to an anchor point, then the solution set must be bounded,
- a solution matrix Z has rank at least 2,
- it's 2 if and only if $Y = X^T X$ and it solves the original problem.

SDP Dual Form [Biswas-Ye 2004]

$$\begin{array}{ll} \text{minimize} & I \bullet V + \sum_{i<j \in N_x} w_{ij} d_{ij}^2 + \sum_{k,j \in N_a} \hat{w}_{kj} \hat{d}_{kj}^2 \\ \text{subject to} & \begin{pmatrix} V & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} + \sum_{i<j \in N_x} w_{ij} (\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)(\mathbf{0}; \mathbf{e}_i - \mathbf{e}_j)^T \\ & + \sum_{k,j \in N_a} w_{kj} (\mathbf{a}_k; -\mathbf{e}_j)(\mathbf{a}_k; -\mathbf{e}_j)^T = S \succeq 0, \end{array}$$

where variable matrix $V \in \mathcal{S}^2$, variable w_{ij} is the (stress) weight on edge between $\mathbf{x}_i$ and $\mathbf{x}_j$, and $\hat{w}_{kj}$ is the (stress) weight on edge between $\mathbf{a}_k$ and $\mathbf{x}_j$.

- The dual is always feasible since $V = \mathbf{0}$ and all $w_{\cdot}$s equal 0 is a feasible solution.
- The rank of any optimal dual stress matrix S is less or equal to n.

Unique Localizability

A sensor network is uniquely-localizable (UL) if there exists a unique localization in $\mathbf{R}^2$ and there is no nontrivial localizations, $x_j \in \mathbf{R}^h, \; j = 1, ..., n$, where $h > 2$, such that

$$\|\mathbf{x}_i - \mathbf{x}_j\|^2 = d_{ij}^2, \; \forall \; i, j \in N_x, \; i < j,$$
$$\|(\mathbf{a}_k; \mathbf{0}) - \mathbf{x}_j\|^2 = \hat{d}_{kj}^2, \; \forall \; k, j \in N_a.$$

It basically says that the problem cannot be localized in a higher dimension space where anchor points are simply augmented to $(\mathbf{a}_k; \mathbf{0}) \in \mathbf{R}^h, \; k = 1, ..., m.$

UL is called universal rigidity.

ULP is localizable in polynomial time [So-Ye 2005]

Theorem

The following statements are equivalent:

1. *The sensor network is uniquely-localizable;*
2. *The max-rank solution of the SDP relaxation has rank* 2;
3. *The solution matrix has* $Y = X^T X$ *or* $Trace(Y - X^T X) = 0$.

Moreover, the localization of a uniquely localizable instance can be computed approximately in a time polynomial in n, d, *and the accuracy* $\log(1/\epsilon)$.

UL Graphs

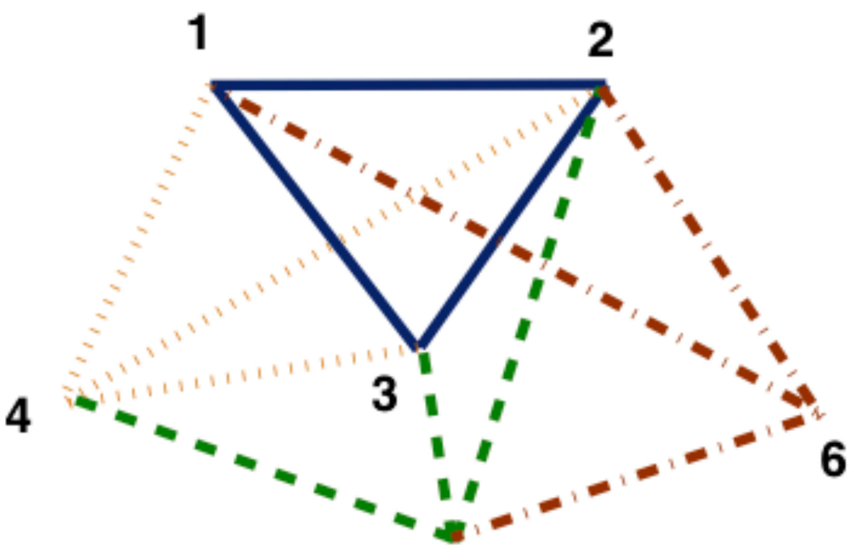
Theorem

Let the SNL problem have at least 3 anchors and they, together with all sensors, are in general positions.

- *If G is complete or every edge length is specified, then the sensor network is uniquely-localizable (Schoenberg 1942).*
- *If one sensor with its edge lengths to at least 3 anchors specified, then it is uniquely-localizable (So and Y 2005).*
- *The trilateration graph is uniquely-localizable (So 2007 and Zhu, So and Y 2009). Moreover, it is the sparsest graph (with only $3n$ edges) that is uniquely-localizable.*

d-lateration Graphs

A $d+1$-Lateration ordering in dimension d for a graph G is an ordering of the vertexes $1, \cdots, d+1, d+2, \cdots, n$ such that K_{d+1}, the complete graph of the first $d+1$ vertexes, is in G, and every vertex $j > d+1$ has $d+1$ edges connected to its preceding vertexes on the sequence.



Open Problems in Sensor Network Localization

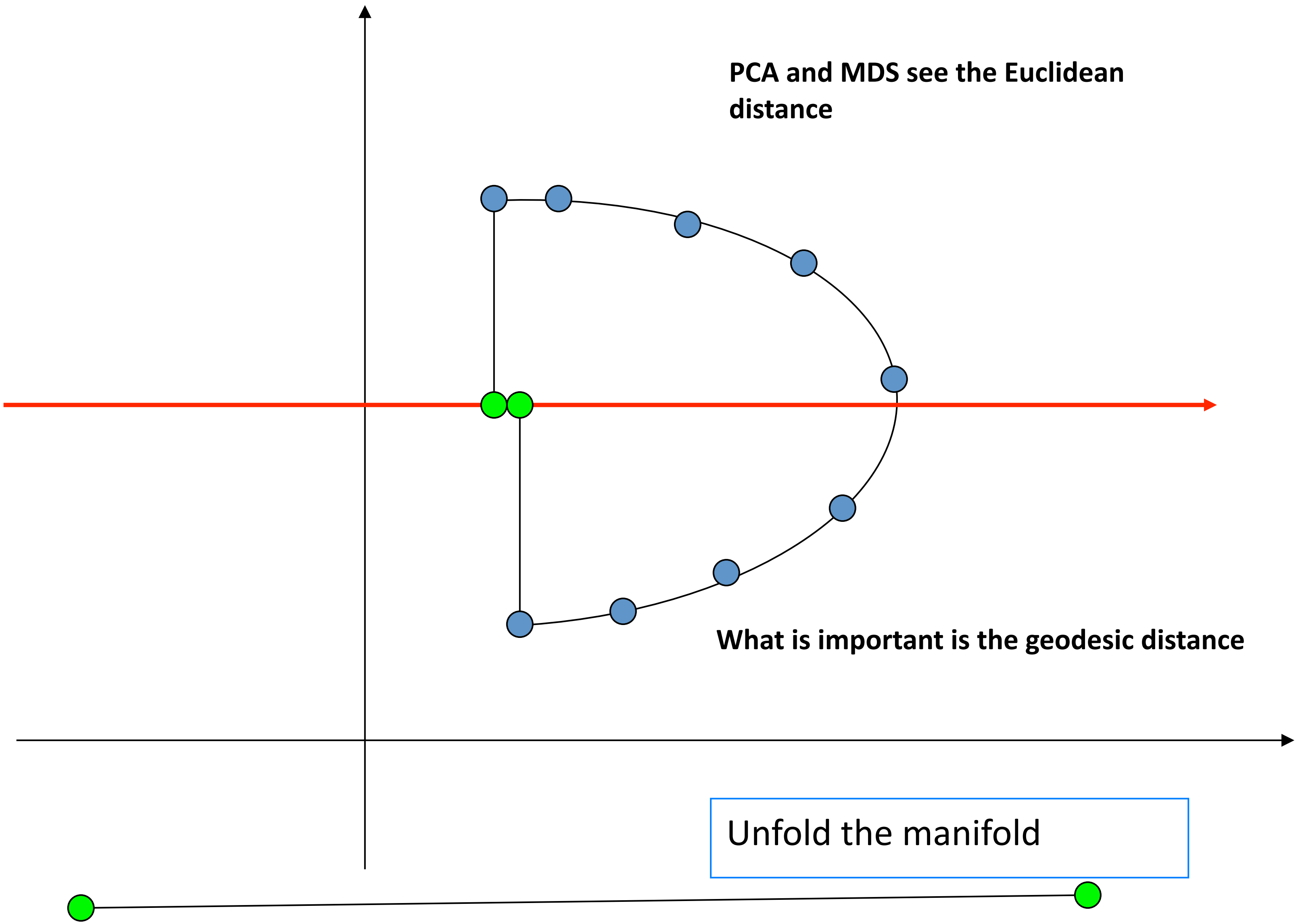
Recall that an SNL problem is strongly localizable if there exists a rank n (dual) stress matrix.

- Let all points be in generic position. Then
  - The existence of a rank n stress matrix implies that the SNL problem is uniquely localizable or universally rigid (Connelly 1999, Alfakih 2010).
  - An SNL problem is universally rigid if and only if there exists a rank n stress matrix (Gortler and Thurston, 2009).
- Let all points be in general position. Then
  - The existence of a rank n stress matrix implies that the SNL problem is universally rigid (Alfakih and Y 2010).
  - An SNL problem that contains a $(d+1)$-lateration graph is universally rigid if and only if there exists a rank n stress matrix (Alfakih, Taheri and Y 2010).

Reference

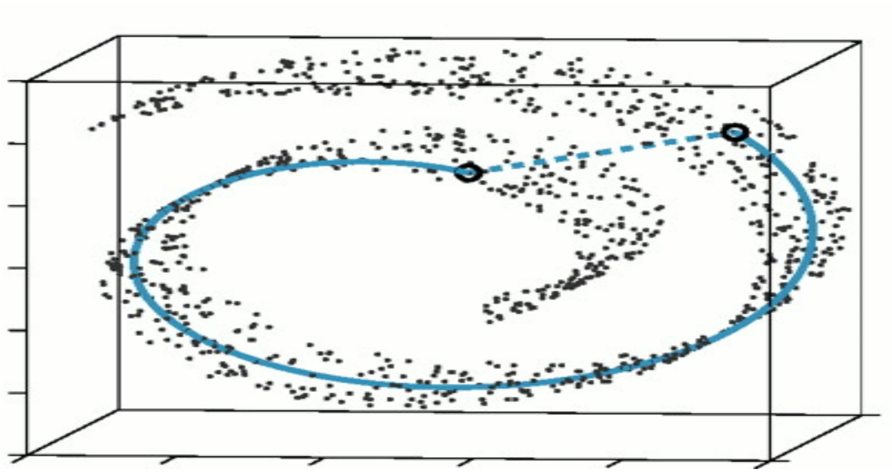
- Schoenberg, I. J. (1938a) Metric Spaces and Completely Monotone Functions. The Annals of Mathematics 39, 811–841
- Schoenberg, I. J. (1938b) Metric Spaces and Positive Definite Functions. Transactions of the American Mathematical Society 44, 522–536
- Francois Bavaud (2010) On the Schoenberg Transformations in Data Analysis: Theory and Illustrations, arXiv:1004.0089v2
- T. F. Cox and M. A. A. Cox, Multidimensional Scaling. New York: Chapman & Hall, 1994.
- Yinyu Ye, Semidefinite Programming and its Low-Rank Theorems, ICCM 2010

Nonlinear Manifolds..



Intrinsic Description..

- To preserve *structure*, preserve the *geodesic* distance and not the *Euclidean* distance.



Two Basic Geometric Embedding Methods

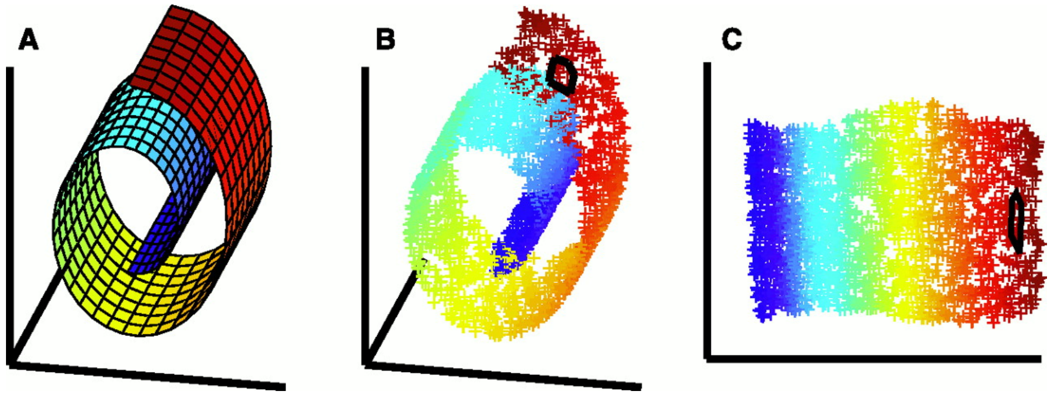
- Tenenbaum-de Silva-Langford Isomap Algorithm
  - Global approach.
  - On a low dimensional embedding
    - Nearby points should be nearby.
    - Faraway points should be faraway.

- Roweis-Saul Locally Linear Embedding Algorithm
  - Local approach
    - Nearby points nearby

Isomap

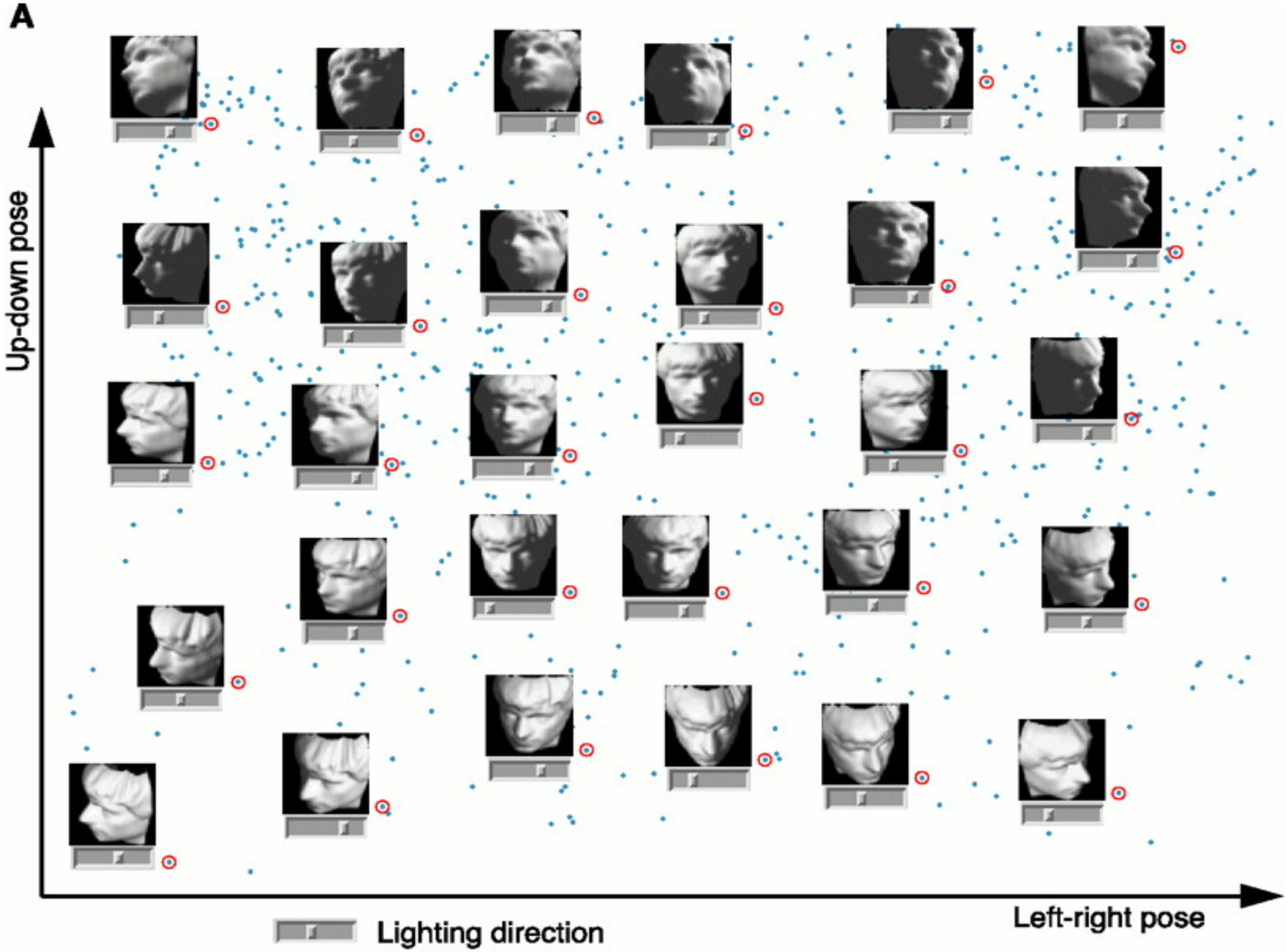
- **Estimate the geodesic distance between faraway points.**
- For neighboring points Euclidean distance is a good approximation to the geodesic distance.
- For faraway points estimate the distance by a series of short hops between neighboring points.
  - Find shortest paths in a graph with edges connecting neighboring data points

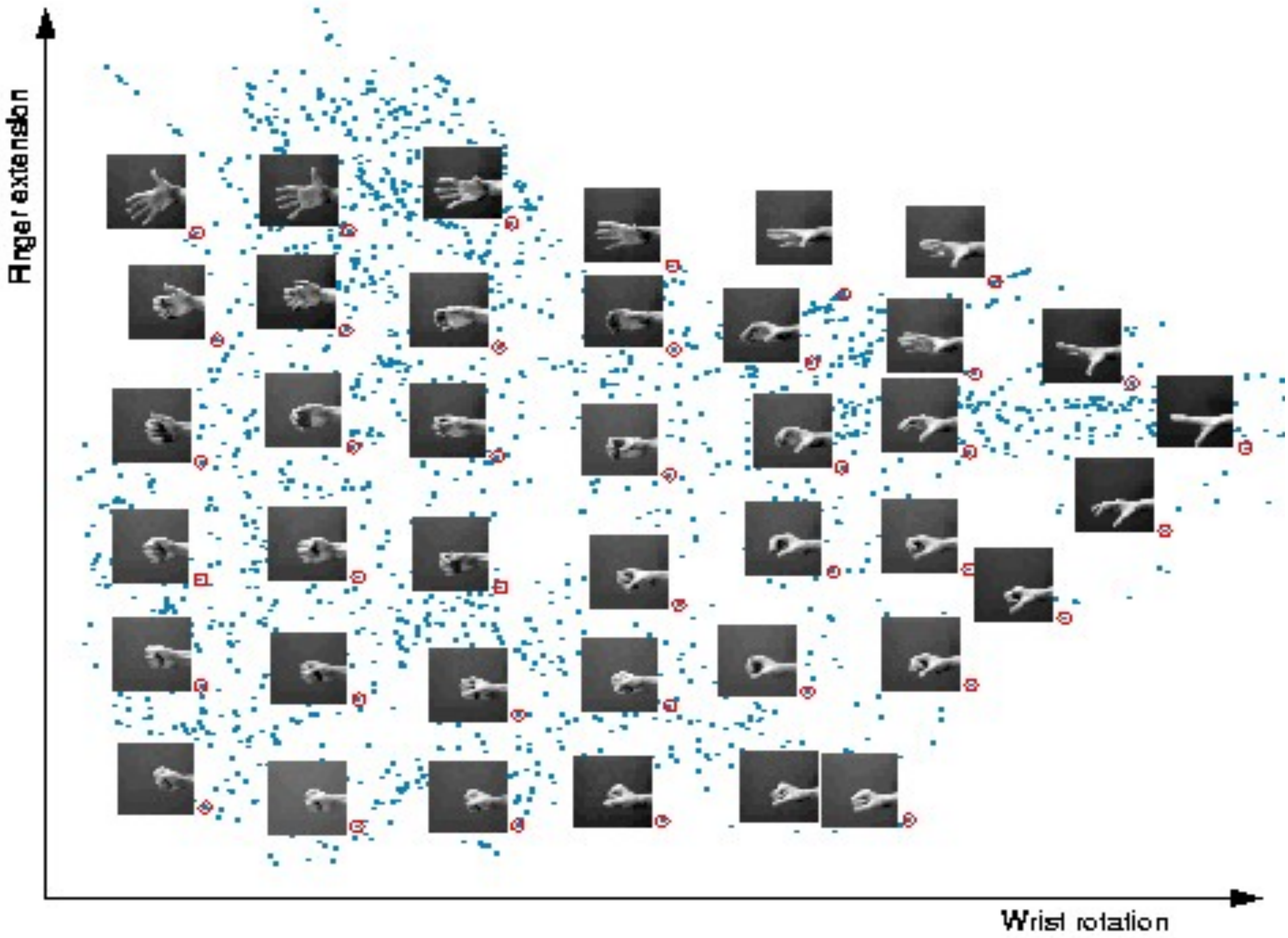
Once we have all pairwise geodesic distances use classical metric MDS

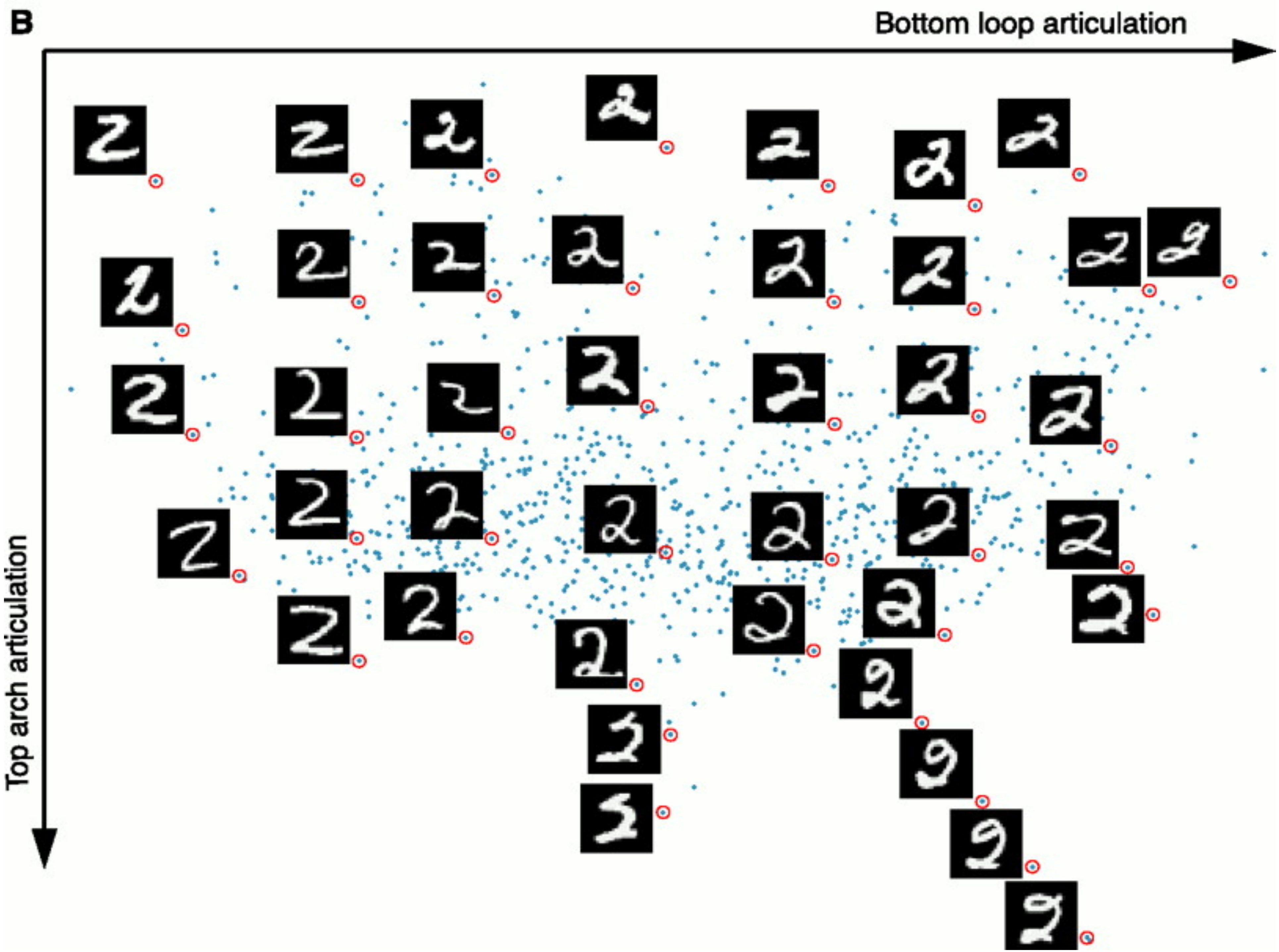


Isomap - Algorithm

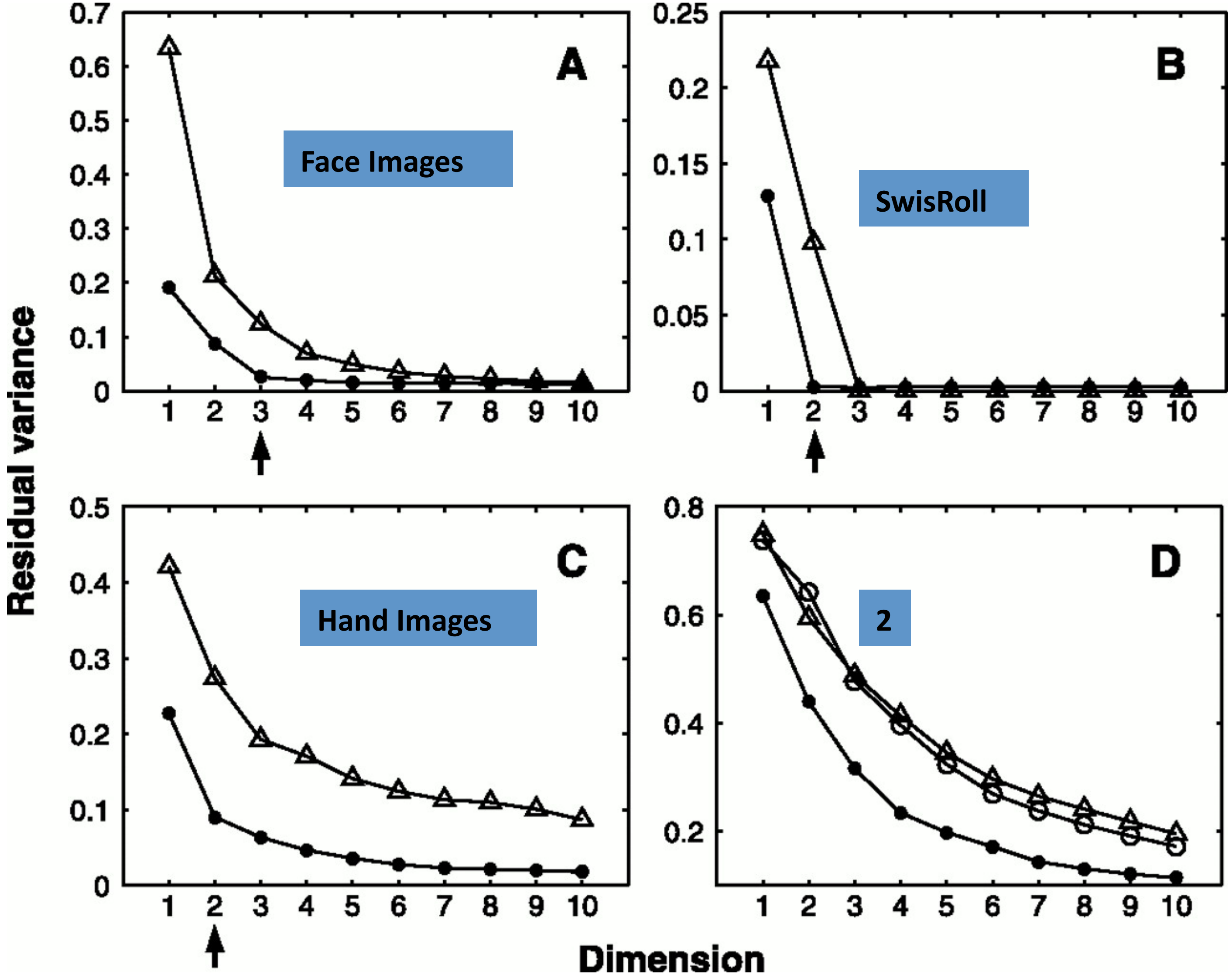
- Determine the neighbors.
  - All points in a fixed radius.
  - K nearest neighbors
- Construct a neighborhood graph.
  - Each point is connected to the other if it is a K nearest neighbor.
  - Edge Length equals the Euclidean distance
- Compute the shortest paths between two nodes
  - Floyd's Algorithm ($O(N^3)$)
  - Dijkstra's Algorithm ($O(kN^2 \log N)$)
- Construct a lower dimensional embedding.
  - Classical MDS


A
Up-down pose
Left-right pose
Lighting direction

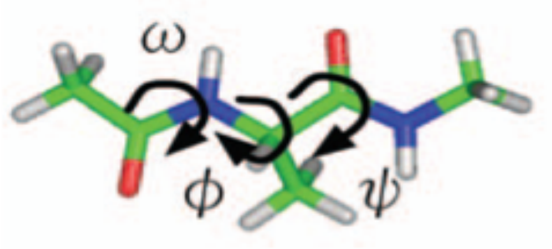

Finger extension
Wrist rotation

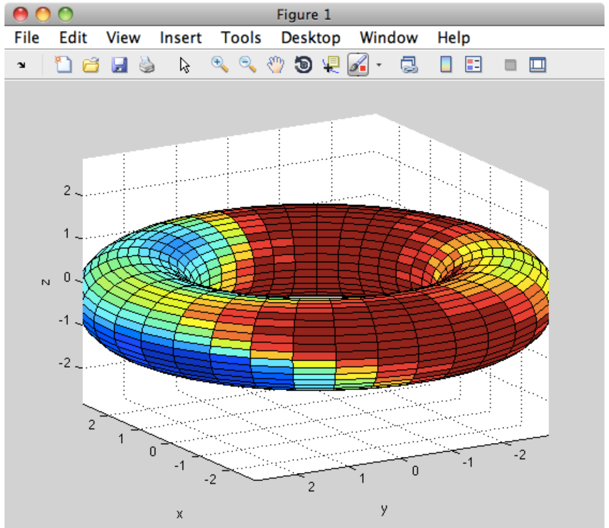

B
Bottom loop articulation
Top arch articulation

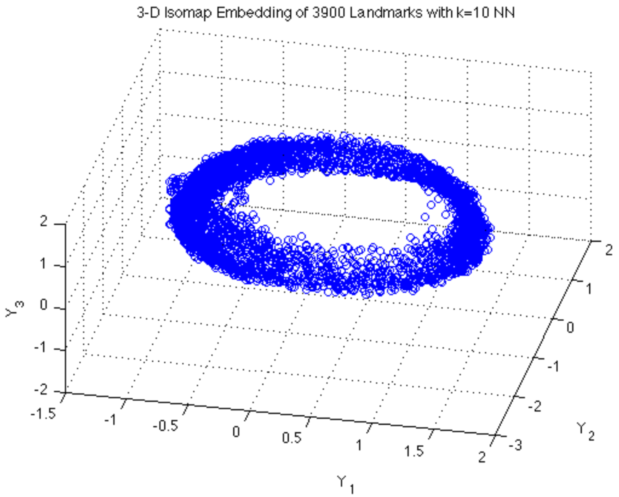
Residual Variance

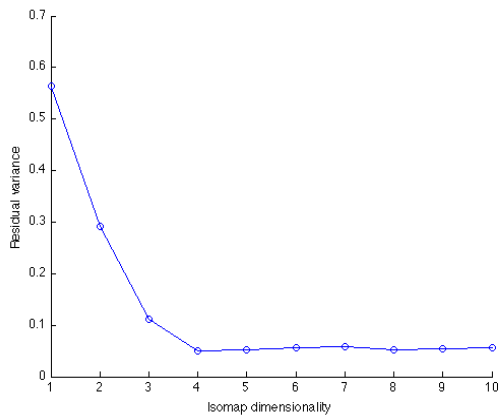


Biomolecular: Alanine-dipeptide









ISOMAP 3D embedding with RMSD metric on 3900 Kcenters

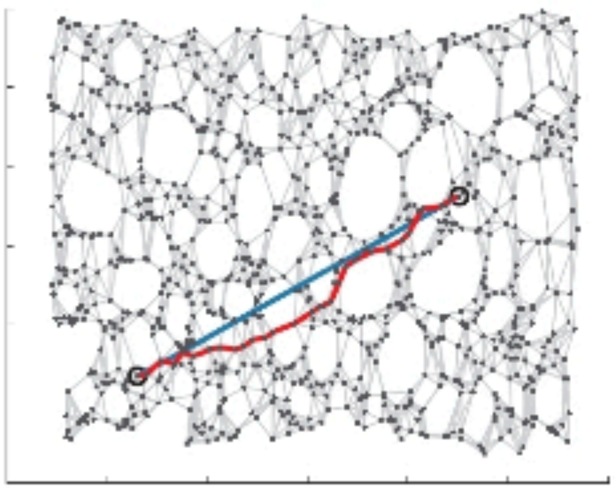
Convergence of ISOMAP

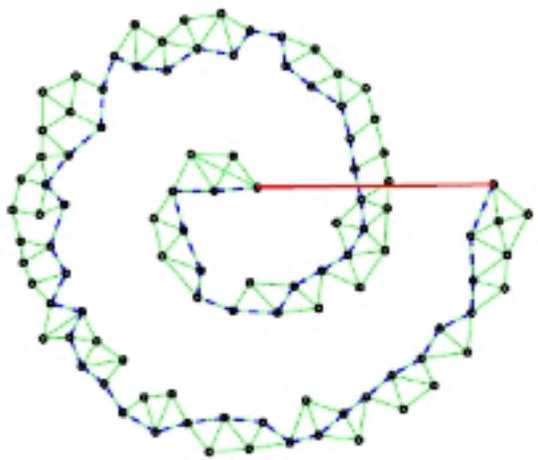
- ISOMAP has provable convergence guarantees;
- Given that $\{x_i\}$ is sampled sufficiently dense, ISOMAP will approximate closely the original distance as measured in manifold M;
- In other words, actual geodesic distance approximations using graph G can be arbitrarily good;
- Let's examine these theoretical guarantees in more detail …

Possible Issues

- It is not immediately obvious that G should give a good approximation to geodesic distances.
- Degenerate cases could lead to zig-zagging behavior that could add a significant amount of overhead.

c





Two step approximations

- Convergence proof hinges on the idea that we can approximate geodesic distance in M by short Euclidean distance hops.

Let's define the following for two points $x, y \in M$:

$$d_M(x, y) = \inf_{\gamma}\{length(\gamma)\}$$

$$d_G(x, y) = \min_{P}(\|x_0 - x_1\| + \ldots + \|x_{p-1} - x_p\|)$$

$$d_S(x, y) = \min_{P}(d_M(x_0, x_1) + \ldots + d_M(x_{p-1}, x_p))$$

where γ varies over the set of smooth arcs connecting x to y in M and P varies over all paths along the edges of G starting at data point $x = x_0$ and ending at $y = x_p$.

- We will show $d_M \approx d_S$ and $d_S \approx d_G$, which will imply the desired result that $d_G \approx d_M$.

Proposition 1. We have the inequalities:

$$\begin{aligned} d_M(x,y) &\leq d_S(x,y) \\ d_G(x,y) &\leq d_S(x,y) \end{aligned}$$

Proof. The first expression is just the triangle inequality for the metric d_M. The second inequality holds because the Euclidean distances $\|x_i - x_{i+1}\|$ are smaller than the arc-length distances $d_M(x_i, x_{i+1})$. $\square$

Dense-sampling Theorem [Bernstein, de Silva, Langford, and Tenenbaum 2000]

Theorem 1: Let $\epsilon, \delta > 0$ with $4\delta < \epsilon$. Suppose G contains all edges $e = (x, y)$ for which $d_M(x, y) < \epsilon$. Furthermore, assume for every point $m \in M$ there is a data point x_i such that $d_M(m, x_i) < \delta$ (δ-sampling condition).

Then for all pairs of data points x,y we have:

$$d_M(x, y) \leq d_S(x, y) \leq (1 + 4\delta/\epsilon) d_M(x, y)$$

Proof

Proof of Theorem 1

$$d_M(x, y) \leq d_S(x, y) \leq (1 + 4\delta/\epsilon) d_M(x, y)$$

Proof:

- The left hand side of the inequality follows directly from the triangle inequality.
- Let γ be any piecewise-smooth arc connecting x to y with $\ell = \mathit{length}(\gamma)$.
- If $\ell \leq \epsilon - 2\delta$ then x and y are connected by an edge in G which we can use as our path.

Proof

Proof (cont'd)

- If $\ell > \epsilon - 2\delta$ then we can write $\ell = \ell_0 + (\ell_1 + \ldots + \ell_1) + \ell_0$ where $\ell_1 = \epsilon - 2\delta$ and $\epsilon - 2\delta \geq \ell_0 \geq (\epsilon - 2\delta)/2$.
- This splits up arc γ into a sequence of points $\gamma_0 = x, \gamma_1, \ldots, \gamma_p = y$. Each point γ_i lies within a distance δ of a sample data point x_i. *Claim:* The path $xx_1x_2 \ldots x_{p-1}y$ satisfies our requirements.

$$\begin{aligned}
d_M(x_i, x_{i+1}) &\leq d_M(x_i, \gamma_i) + d_M(\gamma_i, \gamma_{i+1}) + d_M(\gamma_{i+1}, x_{i+1}) \\
&\leq \delta + \ell_1 + \delta \\
&= \epsilon \\
&= \ell_1 \epsilon/(\epsilon - 2\delta)
\end{aligned}$$

Proof

Proof (cont'd)

- Similarly $d_M(x, x_1) \leq \ell_0 \epsilon/(\epsilon - 2\delta) \leq \epsilon$ and the same holds for $d_M(x_{p-1}, y)$.

$$\begin{aligned} d_M(x_0, x_1) + \ldots + d_M(x_{p-1}, x_p) &\leq \ell\epsilon/(\epsilon - 2\delta) \\ &< \ell(1 + 4\delta/\epsilon) \end{aligned}$$

- The last inequality utilizes the fact that $1/(1-t) < 1 + 2t$ for $0 < t < 1/2$.
- Finally, we take the inf over all γ giving $\ell = d_M(x, y)$.
- Thus, we see that $d_S \approx d_M$ arbitrarily well given both the graph construction and δ-sampling conditions.

The Second Approximation

$d_S \approx d_G$

- We would like to now show the other approximate equality: $d_S \approx d_G$. First let's make some definitions:
    1. The *minimum radius of curvature* $r_0 = r_0(M)$ is defined by $\frac{1}{r_0} = \max_{\gamma,t} \|\gamma''(t)\|$ where γ varies over all unit-speed geodesics in M and t is in the domain D of γ.
        - Intuitively, geodesics in M curl around 'less tightly' than circles of radius less than $r_0(M)$.
    2. The *minimum branch separation* $s_0 = s_0(M)$ is the largest positive number for which $\|x - y\| < s_0$ implies $d_M(x, y) \leq \pi r_0$ for any $x, y \in M$.

Lemma: If γ is a geodesic in M connecting points x and y, and if $\ell = length(\gamma) \leq \pi r_0$, then:

$$2r_0 sin(\ell/2r_0) \leq \|x - y\| \leq \ell$$

Remarks

- We will take this Lemma without proof as it is somewhat technical and long.
- Using the fact that $sin(t) \geq t - t^3/6$ for $t \geq 0$ we can write down a weakened form of the Lemma:

$$(1 - \ell^2/24r_0^2)\ell \leq \|x - y\| \leq \ell$$

- We can also write down an even more weakened version valid for $\ell \leq \pi r_0$:

$$(2/\pi)\ell \leq \|x - y\| \leq \ell$$

- We can now show $d_G \approx d_S$.

Theorem 2 [Bernstein, de Silva, Langford, and Tenenbaum 2000]

Theorem 2: Let $\lambda > 0$ be given. Suppose data points $x_i, x_{i+1} \in M$ satisfy:

$$\|x_i - x_{i+1}\| < s_0$$
$$\|x_i - x_{i+1}\| \leq (2/\pi) r_0 \sqrt{24\lambda}$$

Suppose also there is a geodesic arc of length $\ell = d_M(x_i, x_{i+1})$ connecting x_i to x_{i+1}. Then:

$$(1 - \lambda)\ell \leq \|x_i - x_{i+1}\| \leq \ell$$

Proof

Proof of Theorem 2

- By the first assumption we can directly conclude $\ell \leq \pi r_0$.
- This fact allows us to apply the Lemma using the weakest form combined with the second assumption gives us:

$$\ell \leq (\pi/2)\,\|x_i - x_{i+1}\| \leq r_0\sqrt{24\lambda}$$

- Solving for λ in the above gives: $1 - \lambda \leq (1 - \ell^2/24r_0^2)$. Applying the weakened statement of the Lemma then gives us the desired result.
- Combining Theorem 1 and 2 shows $d_M \approx d_G$. This leads us then to our main theorem...

Main Theorem [Bernstein, de Silva, Langford, and Tenenbaum 2000]

Theorem 1: Let M be a compact submanifold of $\mathbf{R}^n$ and let $\{x_i\}$ be a finite set of data points in M. We are given a graph G on $\{x_i\}$ and positive real numbers $\lambda_1, \lambda_2 < 1$ and $\delta, \epsilon > 0$. Suppose:

1. G contains all edges (x_i, x_j) of length $\|x_i - x_j\| \leq \epsilon$.
2. The data set $\{x_i\}$ statisfies a δ*-sampling condition* – for every point $m \in M$ there exists an x_i such that $d_M(m, x_i) < \delta$.
3. M is *geodesically convex* – the shortest curve joining any two points on the surface is a geodesic curve.
4. $\epsilon < (2/\pi) r_0 \sqrt{24\lambda_1}$, where r_0 is the *minimum radius of curvature of M* – $\frac{1}{r_0} = \max_{\gamma,t} \|\gamma''(t)\|$ where γ varies over all unit-speed geodesics in M.
5. $\epsilon < s_0$, where s_0 is the *minimum branch separation* of M – the largest positive number for which $\|x - y\| < s_0$ implies $d_M(x, y) \leq \pi r_0$.
6. $\delta < \lambda_2 \epsilon / 4$.

Then the following is valid for all $x, y \in M$,

$$(1 - \lambda_1) d_M(x, y) \leq d_G(x, y) \leq (1 + \lambda_2) d_M(x, y)$$

Probabilistic Result

- So, short Euclidean distance hops along G approximate well actual geodesic distance as measured in M.
- What were the main assumptions we made? The biggest one was the δ-sampling density condition.
- A probabilistic version of the Main Theorem can be shown where each point x_i is drawn from a density function. Then the approximation bounds will hold with high probability. Here's a truncated version of what the theorem looks like now:

Asymptotic Convergence Theorem: Given $\lambda_1, \lambda_2, \mu > 0$ then for density function α sufficiently large:

$$1 - \lambda_1 \leq \frac{d_G(x, y)}{d_M(x, y)} \leq 1 + \lambda_2$$

will hold with probability at least $1 - \mu$ for any two data points x, y.

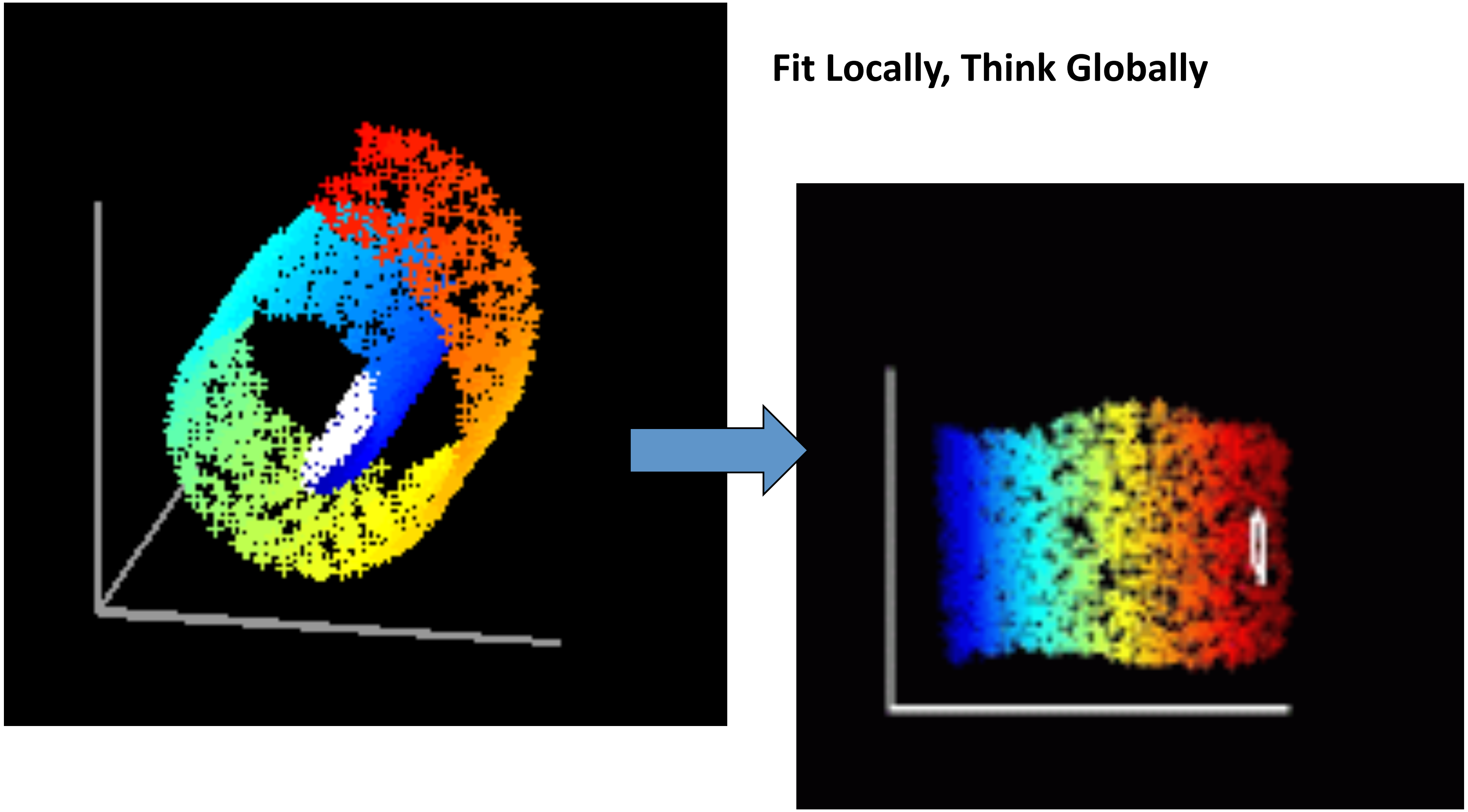
A Shortcoming of ISOMAP

- One need to compute pairwise shortest path between all sample pairs (i,j)
  - Global
  - Non-sparse
  - Cubic complexity $O(N^3)$

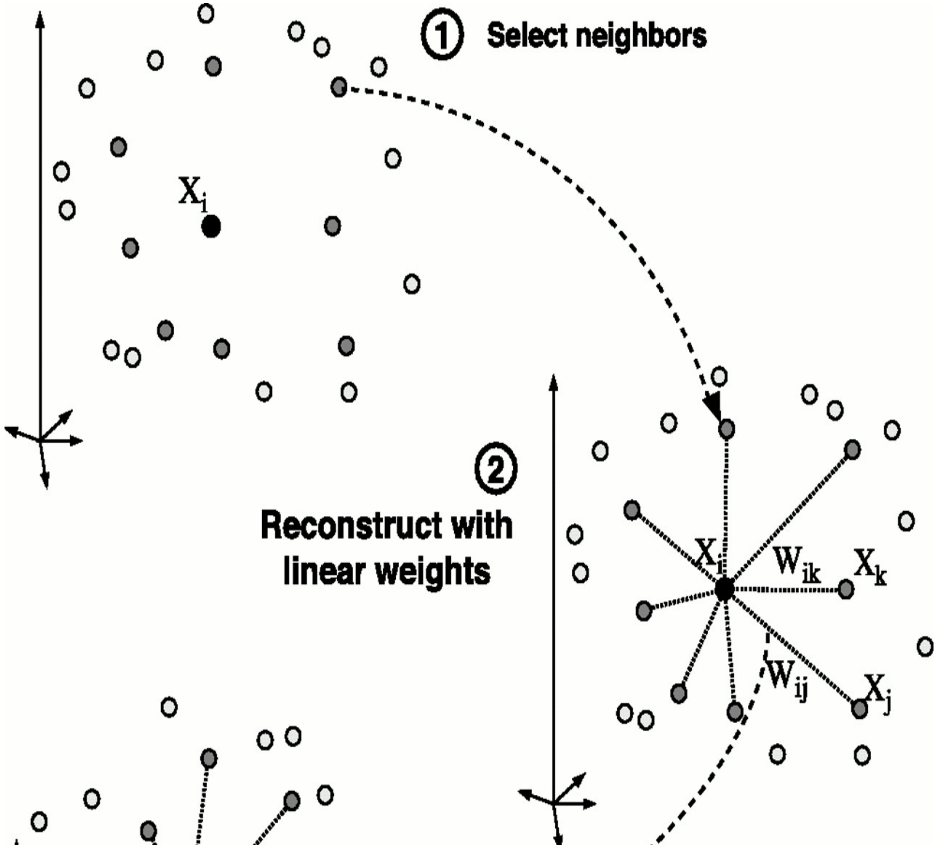
Locally Linear Embedding

manifold is a topological space which is locally Euclidean."

Fit Locally, Think Globally



Fit Locally…



We expect each data point and its neighbours to lie on or close to a locally linear patch of the manifold.

Each point can be written as a linear combination of its neighbors. The weights choosen to minimize the reconstruction Error.

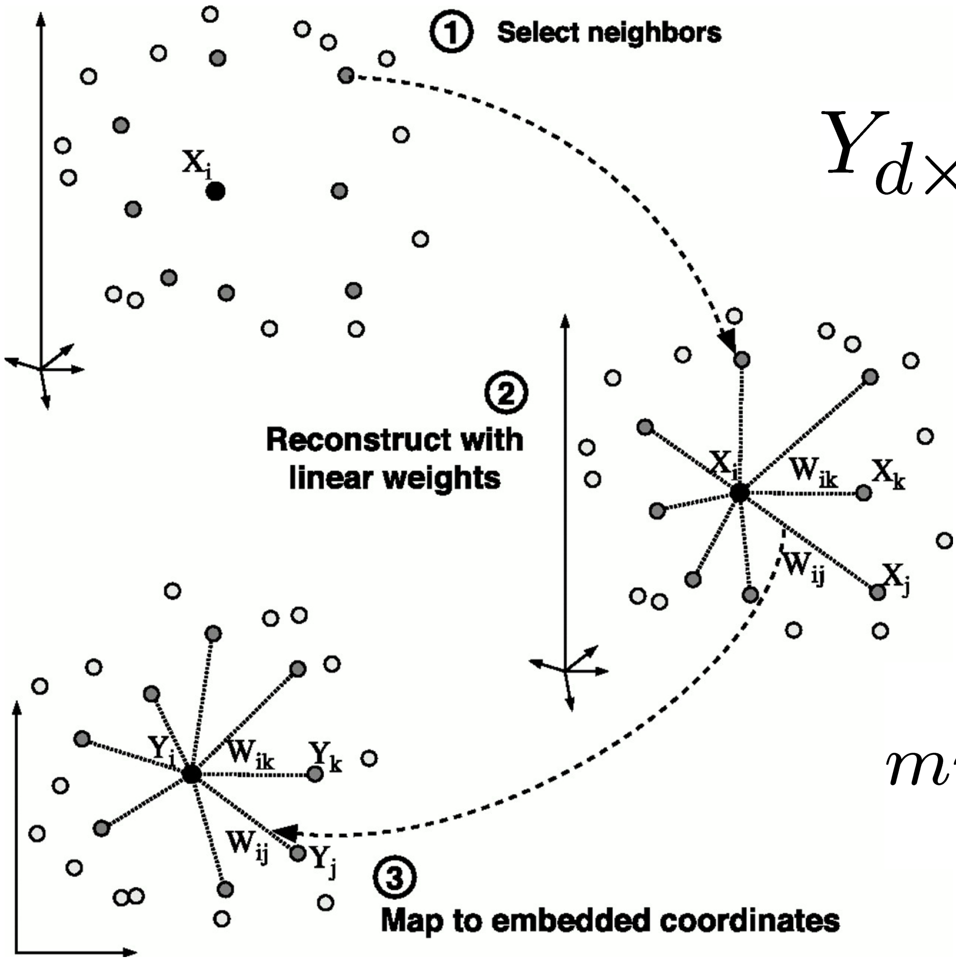
$$min_W \| X_i - \sum_{j=1}^{K} W_{ij} X_j \|^2 \qquad (1)$$

Derivation on board

Important property…

- The weights that minimize the reconstruction errors are invariant to rotation, rescaling and translation of the data points.
  - Invariance to translation is enforced by adding the constraint that the weights sum to one.
- **The same weights that reconstruct the datapoints in D dimensions should reconstruct it in the manifold in d dimensions.**
  - The weights characterize the intrinsic geometric properties of each neighborhood.

Think Globally...



$$Y_{d \times N} = [Y_1|Y_2|...|Y_N]$$

$$min_Y \sum_{i=1}^{N} \| Y_i - YW_i \|^2$$

Algorithm (K-NN)

- Local fitting step (with centering):
  - Consider a point x_i
  - Choose its K(i) neighbors η_j whose origin is at x_i
  - Compute the (sum-to-one) weights w_{ij} which minimizes

$$\Psi_i(w) = \left\| x_i - \sum_{j=1}^{K(i)} w_{ij}\eta_j \right\|^2, \quad \sum_j w_{ij} = 1, \quad x_i = 0$$

    - Contruct neighborhood inner product: $C_{jk} = \langle \eta_j, \eta_k \rangle$
    - Compute the weight vector $w_i=(w_{ij})$, where 1 is K-vector of all-one and λ is a regularization parameter

$$w_i = (C + \lambda I)^{-1} 1$$

    - Then normalize w_i to a *sum-to-one* vector.

Algorithm (K-NN)

- Local fitting step (without centering):
  - Consider a point x_i
  - Choose its K(i) neighbors x_j
  - Compute the (sum-to-one) weights w_{ij} which minimizes

$$\Psi_i(w) = \left\| x_i - \sum_{j=1}^{K(i)} w_{ij} x_j \right\|^2,$$

    - Contruct neighborhood inner product: $C_{jk} = \langle \eta_j, \eta_k \rangle$
    - Compute the weight vector w_i=(w_{ij}), where $v_{ik} = \langle \eta_k, \quad x_i \rangle$

$$w_i = C^+ v_i, \qquad v_i = (v_{ik}) \in R^{K(i)}$$

Algorithm continued

- Global embedding step:
  - Construct N-by-N weight matrix *W:* $W_{ij} = \begin{cases} w_{ij}, & j \in N(i) \\ 0, & otherwise \end{cases}$
  - Compute d-by-N matrix Y which minimizes

$$\phi(Y) = \sum_i \left\| Y_i - \sum_{j=1}^{N} W_{ij} Y_j \right\|^2 = Y(I-W)^T (I-W) Y^T$$

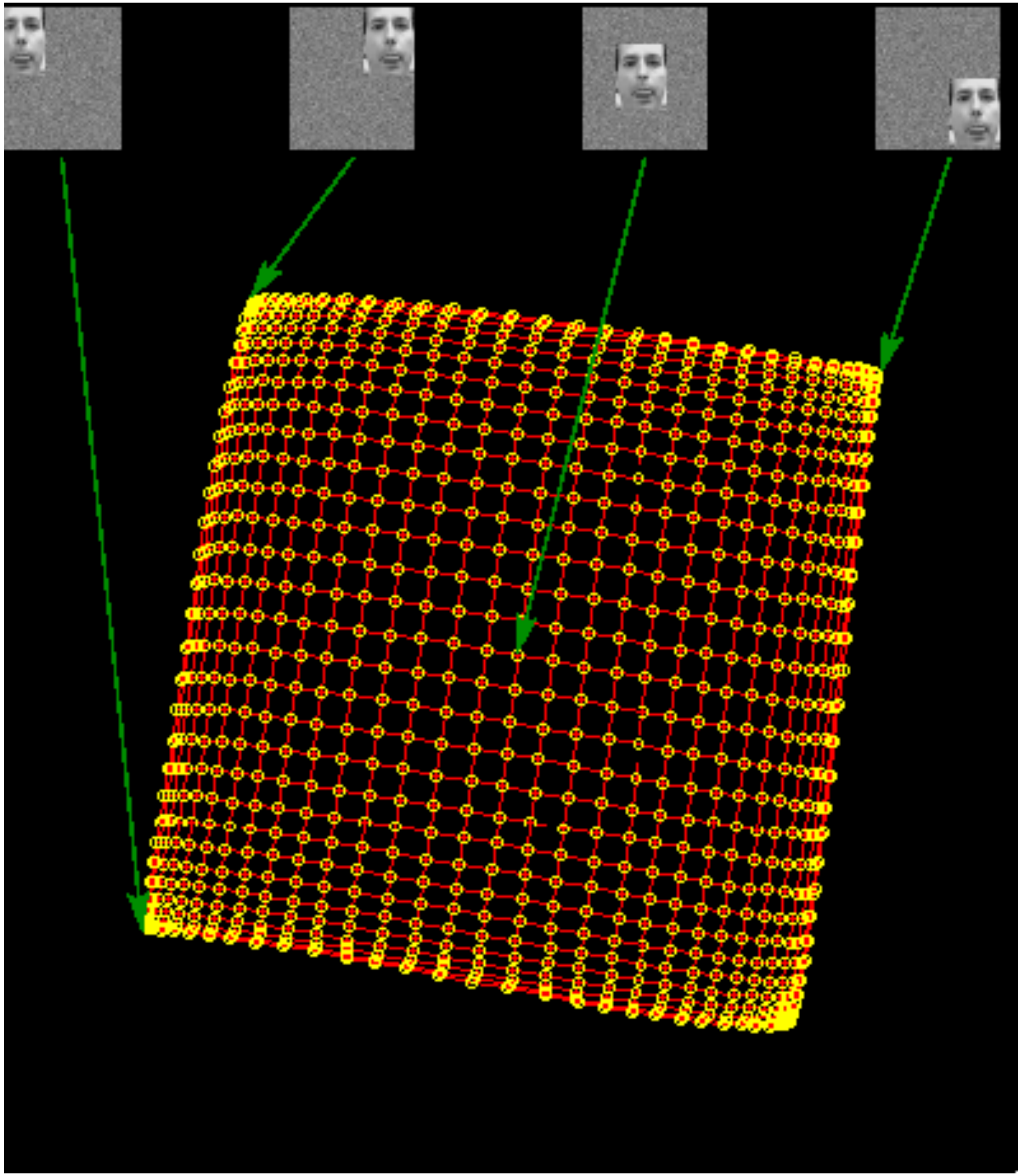
    - Compute: $B = (I-W)^T (I-W)$
    - Find d+1 bottom eigenvectors of *B,* $v_n, v_{n-1}, \ldots, v_{n-d}$
    - Let d-dimensional embedding Y =[$v_{n-1}, v_{n-2}, \ldots v_{n-1}$]

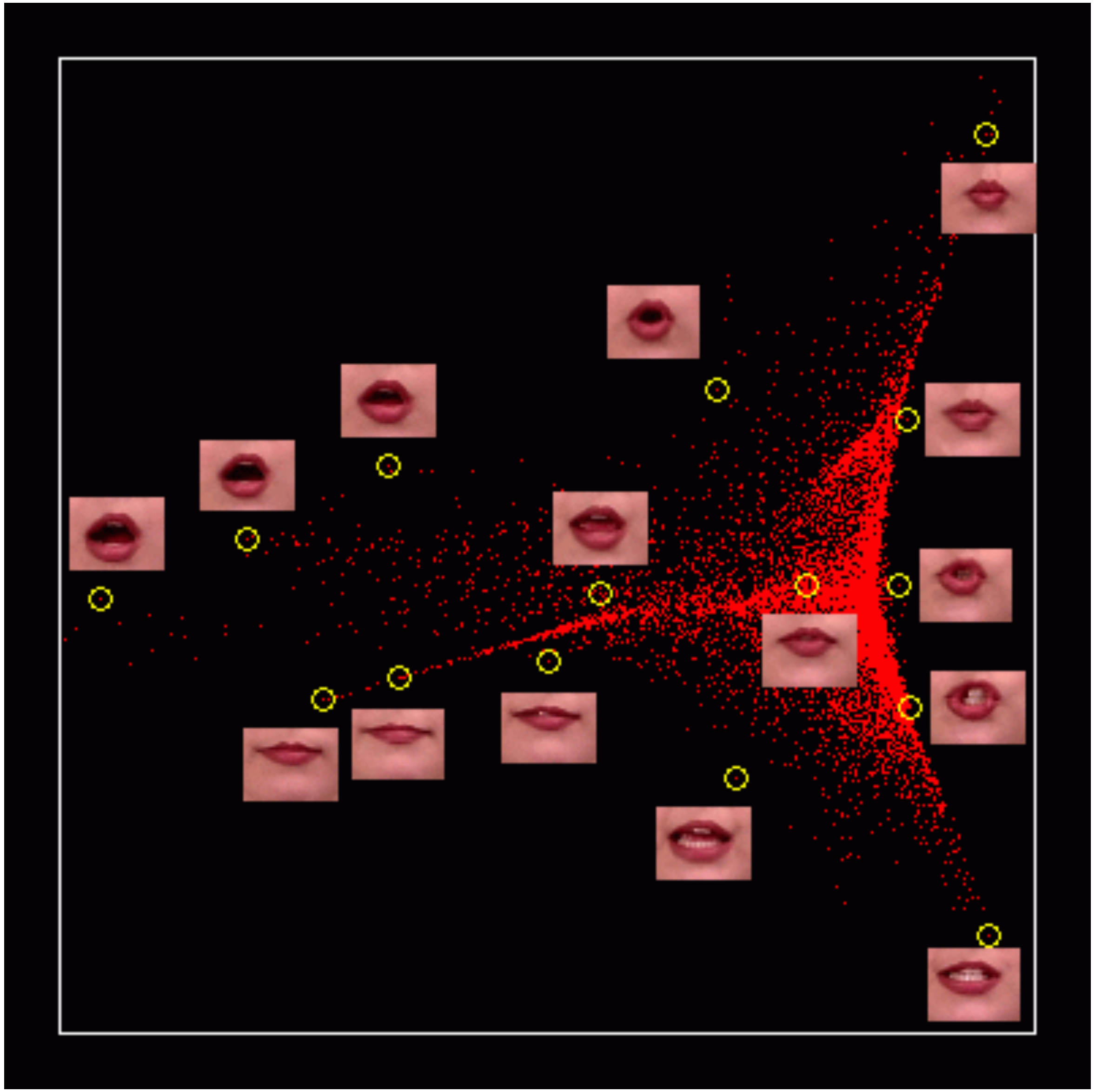
Remarks on LLE

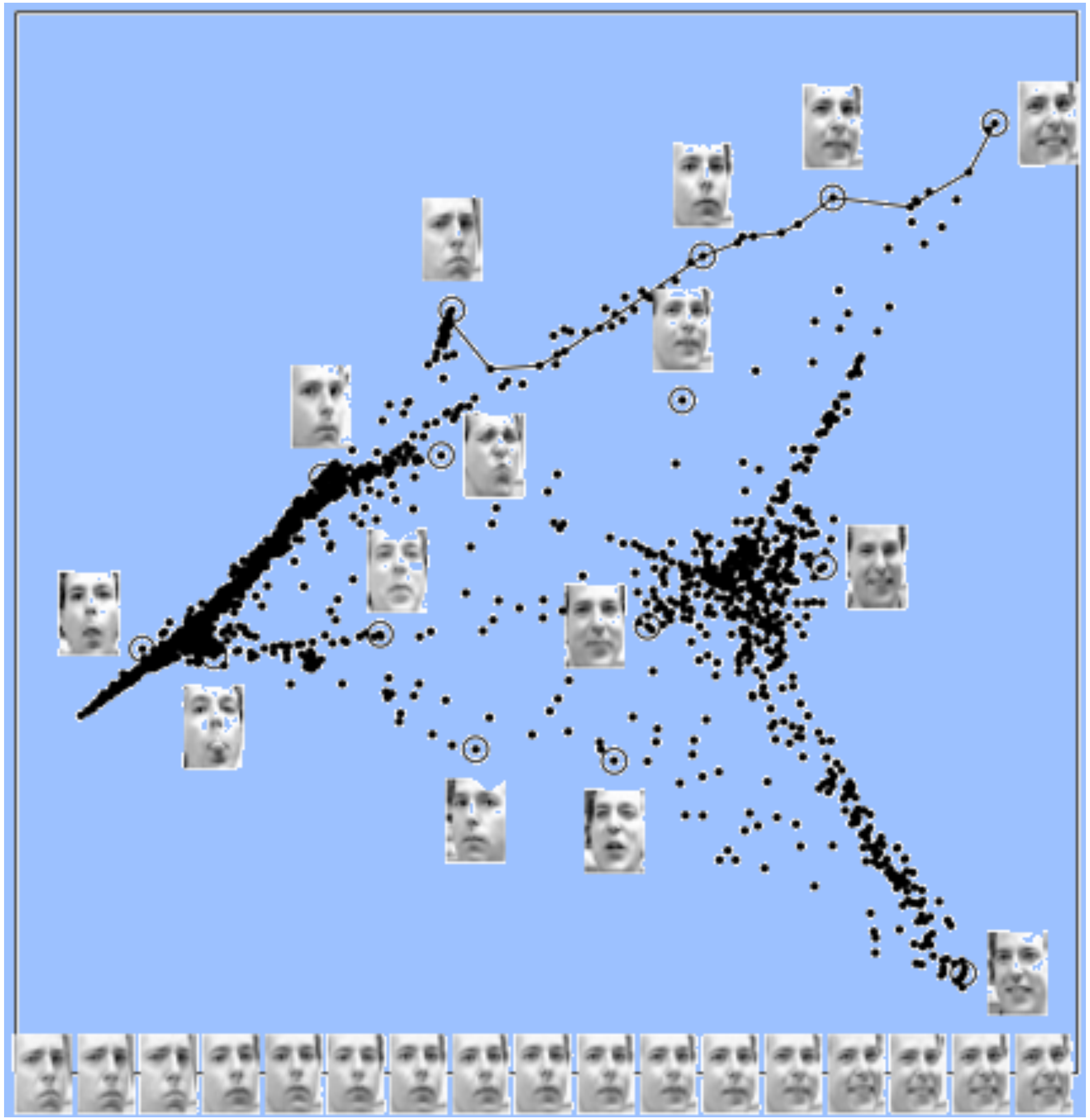
- Searching k-nearest neighbors is of O(kN)

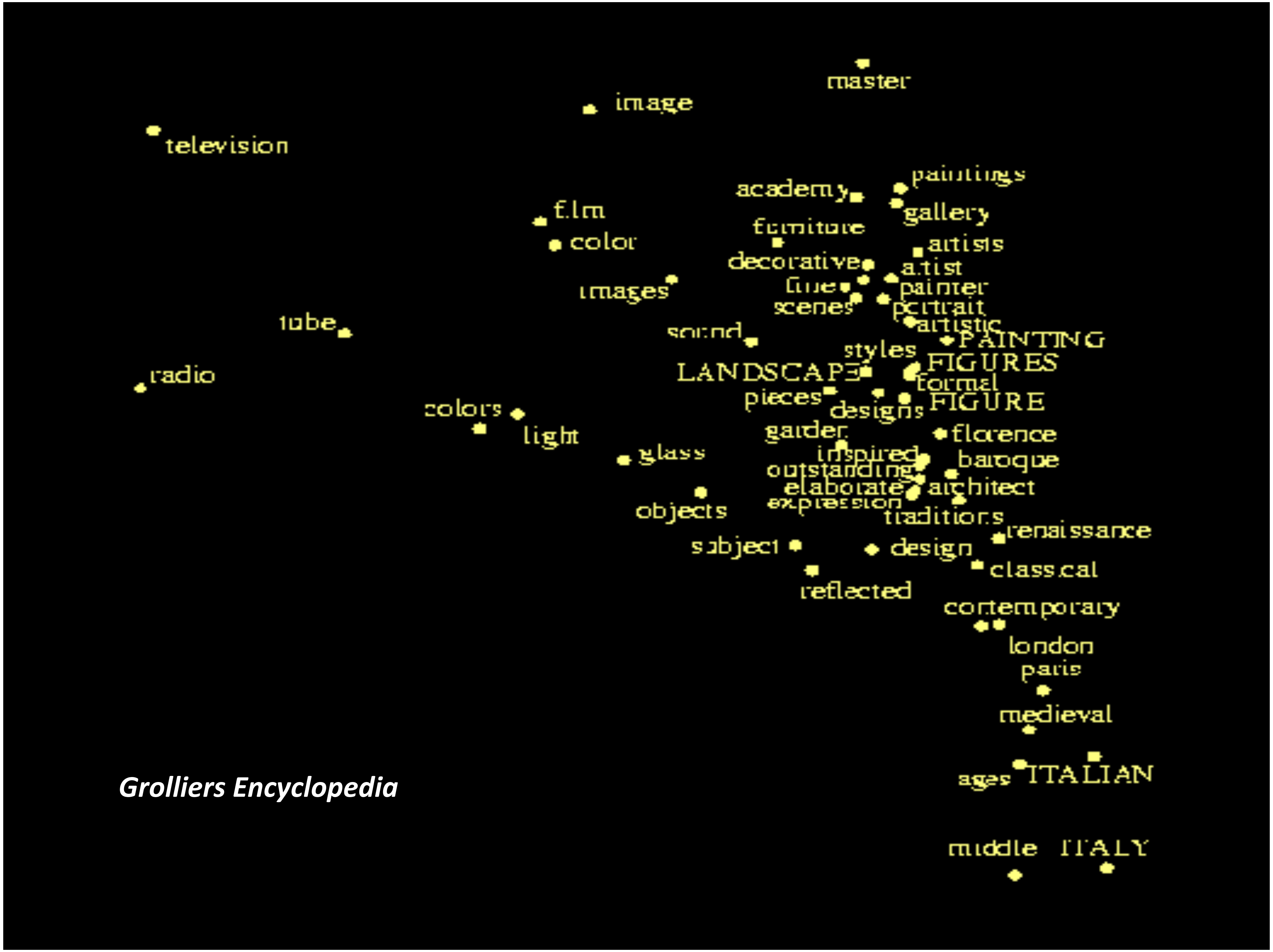
- W is sparse, kN/N^2=k/N nozeros
- W might be negative, additional nonnegative constraint can be imposed
- $B=(I-W)^T(I-W)$ is positive semi-definite (p.s.d.)

- Open Problem: exact reconstruction condition?









Grolliers Encyclopedia

Summary..

ISOMAP	LLE
Do MDS on the geodesic distance matrix.	Model local neighborhoods as linear a patches and then embed in a lower dimensional manifold.
Global approach	Local approach
Might not work for nonconvex manifolds with holes	Nonconvex manifolds with holes
Extensions: Landmark, Conformal & Isometric ISOMAP	Extensions: Hessian LLE, Laplacian Eigenmaps etc.

Both needs manifold finely sampled.

Reference

- Tenenbaum, de Silva, and Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290:2319-2323, 22 Dec. 2000.
- Roweis and Saul, Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290:2323-2326, 22 Dec. 2000.
- M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum. Graph Approximations to Geodesics on Embedded Manifolds. Technical Report, Department of Psychology, Stanford University, 2000.
- V. de Silva and J.B. Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. Neural Information Processing Systems 15 (NIPS'2002), pp. 705-712, 2003.
- V. de Silva and J.B. Tenenbaum. Unsupervised learning of curved manifolds. Nonlinear Estimation and Classification, 2002.
- V. de Silva and J.B. Tenenbaum. Sparse multidimensional scaling using landmark points. Available at: http://math.stanford.edu/~silva/public/publications.html

Matlab Dimensionality Reduction Toolbox

- http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
- Math.pku.edu.cn/teachers/yaoy/Spring2011/matlab/drtoolbox
  - Principal Component Analysis (PCA), Probabilistic PC
  - Factor Analysis (FA), Sammon mapping, Linear Discriminant Analysis (LDA)
  - Multidimensional scaling (MDS), Isomap, Landmark Isomap
  - Local Linear Embedding (LLE), Laplacian Eigenmaps, Hessian LLE, Conformal Eigenmaps
  - Local Tangent Space Alignment (LTSA), Maximum Variance Unfolding (extension of LLE)
  - Landmark MVU (LandmarkMVU), Fast Maximum Variance Unfolding (FastMVU)
  - Kernel PCA
  - Diffusion maps
  - …