ISOMAP and LLE

姚 遠

2019

Fisher 1922

... the objective of statistical methods is the reduction of data. A quantity of data... is to be replaced by relatively few quantities which shall adequately represent ... the relevant information contained in the original data.
Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by an actual sample is irrelevant. It is the object of the statistical process employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data. –R.A.Fisher

Python scikit-learn Manifold learning Toolbox

http://scikit-learn.org/stable/modules/manifold.html

- PCA/MDS(SMACOF algorithm, not spectral method)
- ISOMAP/LLE (+MLLE)
- Hessian Eigenmap
- Laplacian Eigenmap
- LTSA
- tSNE

Matlab Dimensionality Reduction Toolbox

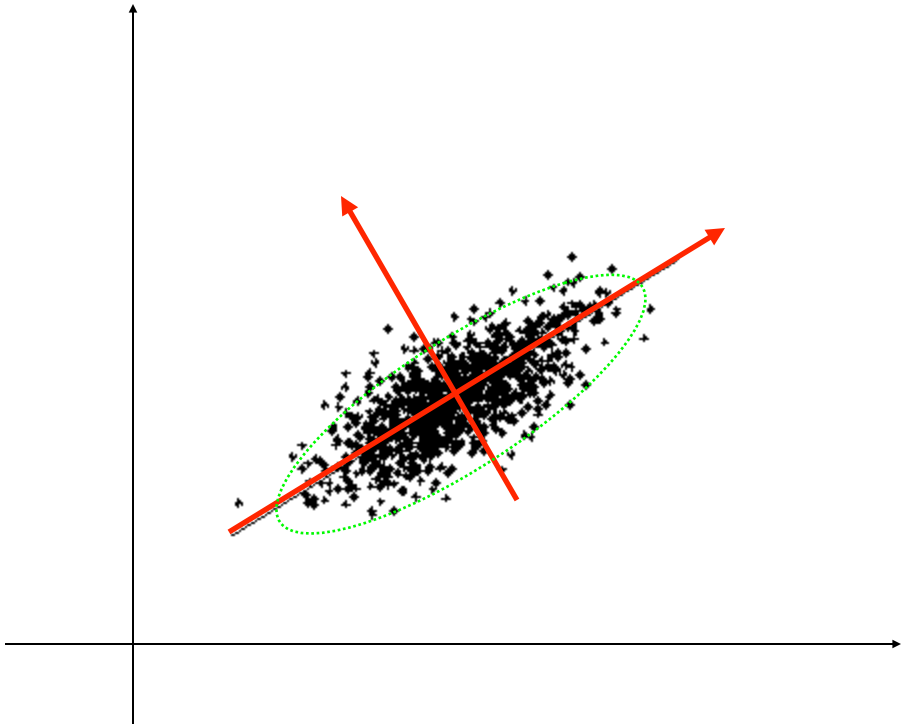
- http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
- Math.pku.edu.cn/teachers/yaoy/Spring2011/matlab/drtoolbox
  - Principal Component Analysis (PCA), Probabilistic PC
  - Factor Analysis (FA), Sammon mapping, Linear Discriminant Analysis (LDA)
  - Multidimensional scaling (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
  - …

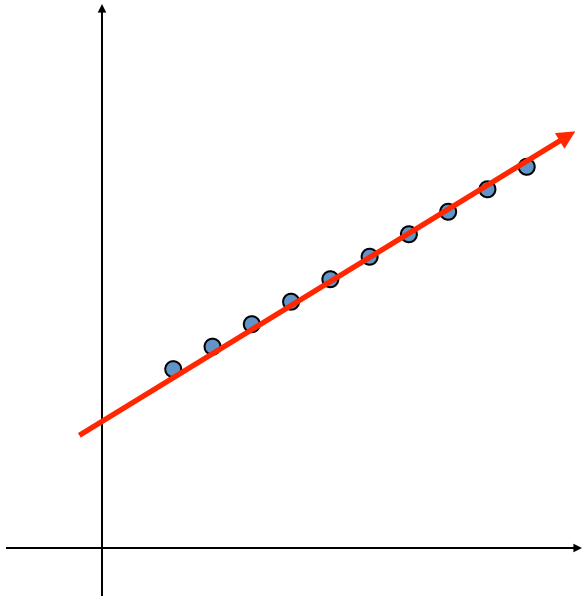
Recall: PCA

- Principal Component Analysis (PCA)

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

EigenValue Decomposition of XX^T

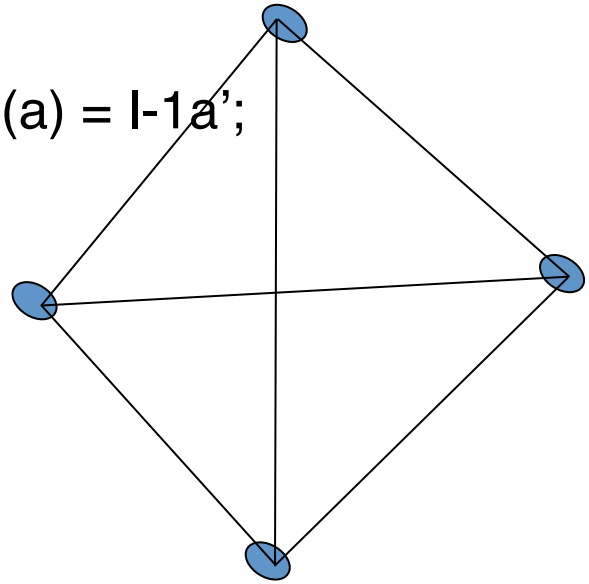




One Dimensional Manifold

Recall: MDS

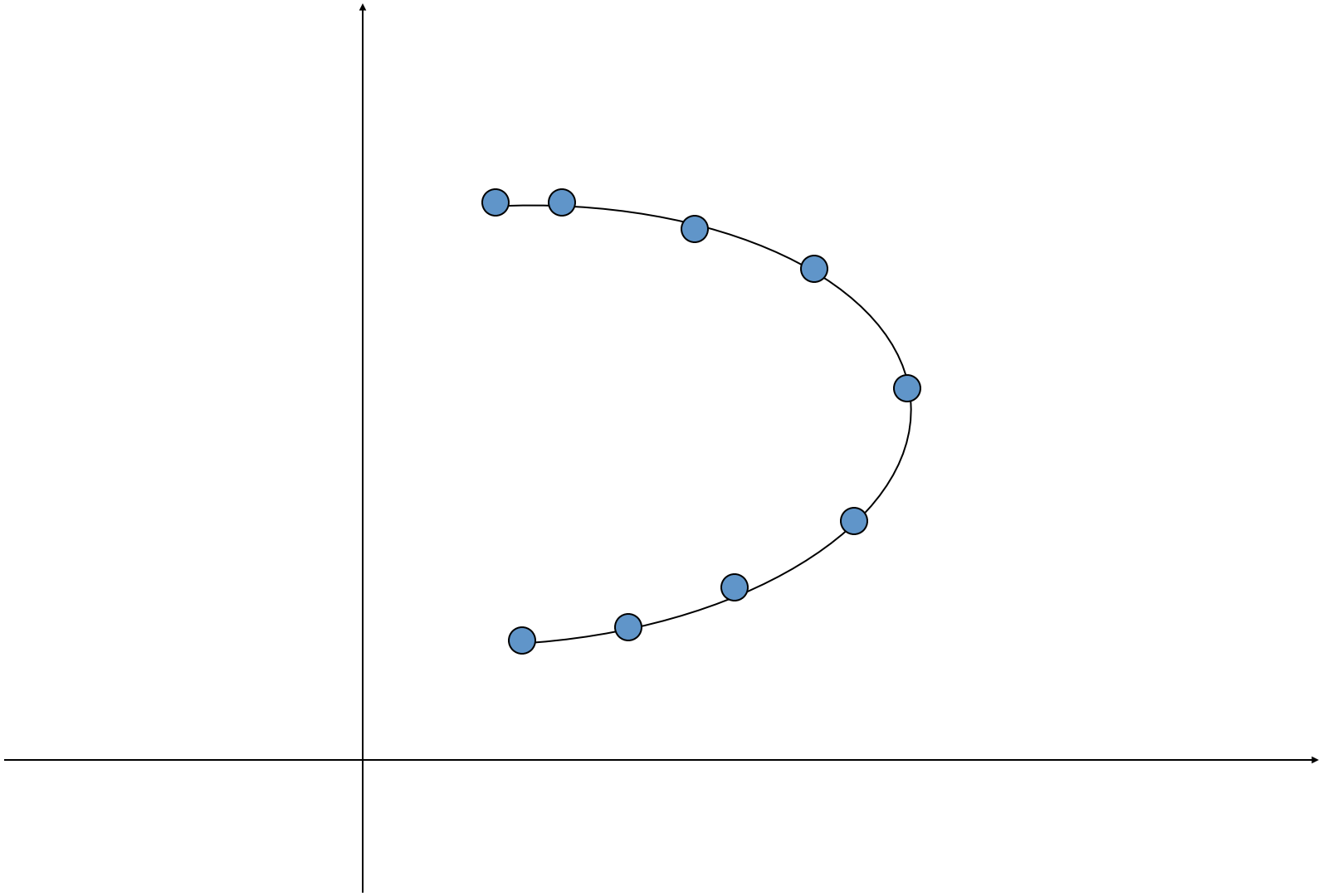
- Given pairwise distances *D*, where $D_{ij} = d_{ij}^2$, the squared distance between point i and j
  - Convert the pairwise distance matrix D (c.n.d.) into the dot product matrix B (p.s.d.)
    - B_{ij} (a) = -.5 H(a) D H'(a), Hölder matrix H(a) = I-1*a*';
    - a = 1_k: $B_{ij} = -.5 (D_{ij} - D_{ik} - D_{jk})$
    - a = 1/n: $B_{ij} = -\frac{1}{2}\left(D_{ij} - \frac{1}{N}\sum_{s=1}^{N} D_{sj} - \frac{1}{N}\sum_{t=1}^{N} D_{it} + \frac{1}{N^2}\sum_{s,t=1}^{N} D_{st} \right)$
  - Eigendecomposition of $B = YY^T$



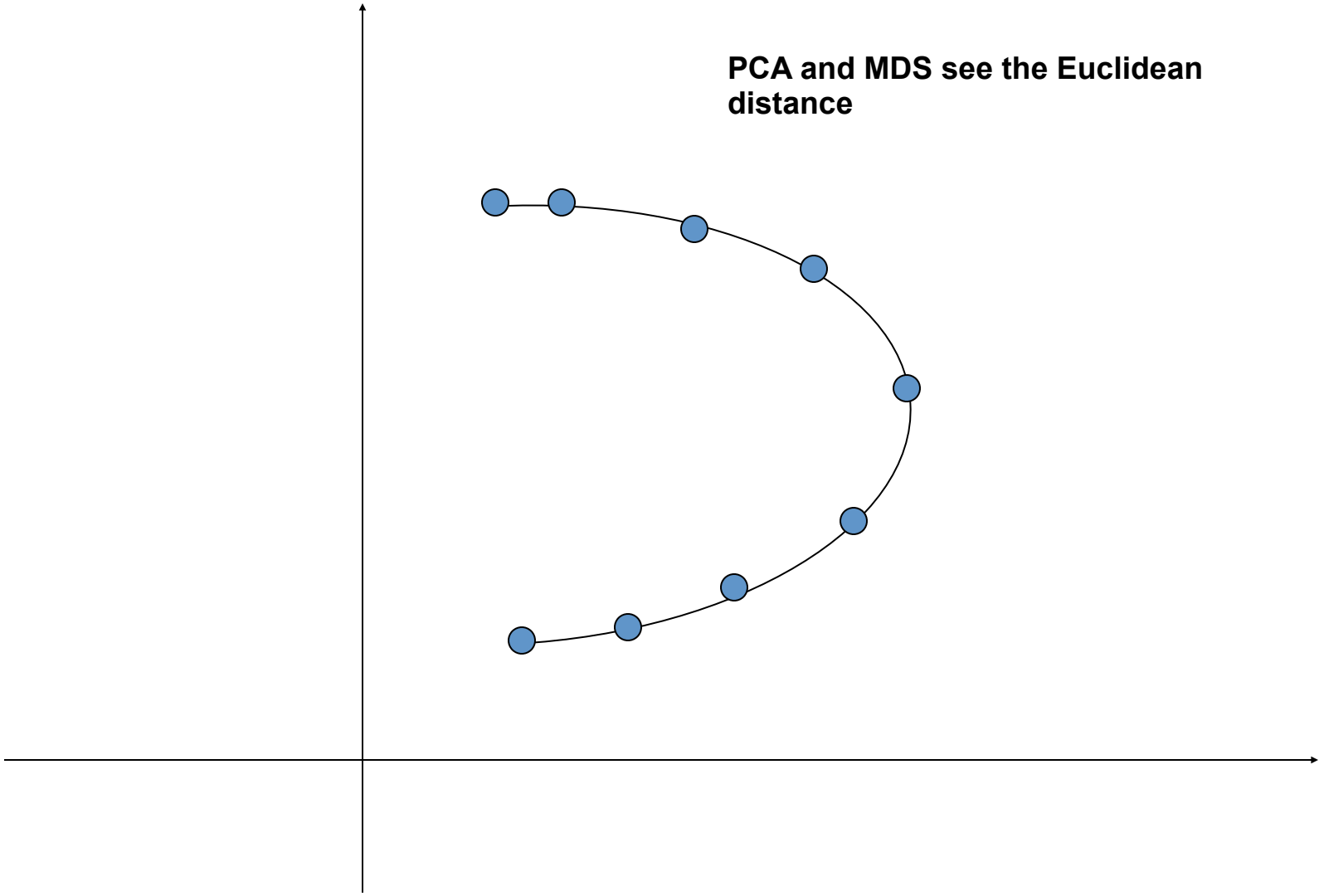
If we preserve the pairwise Euclidean distances do we preserve the structure??

Nonlinear Manifolds..

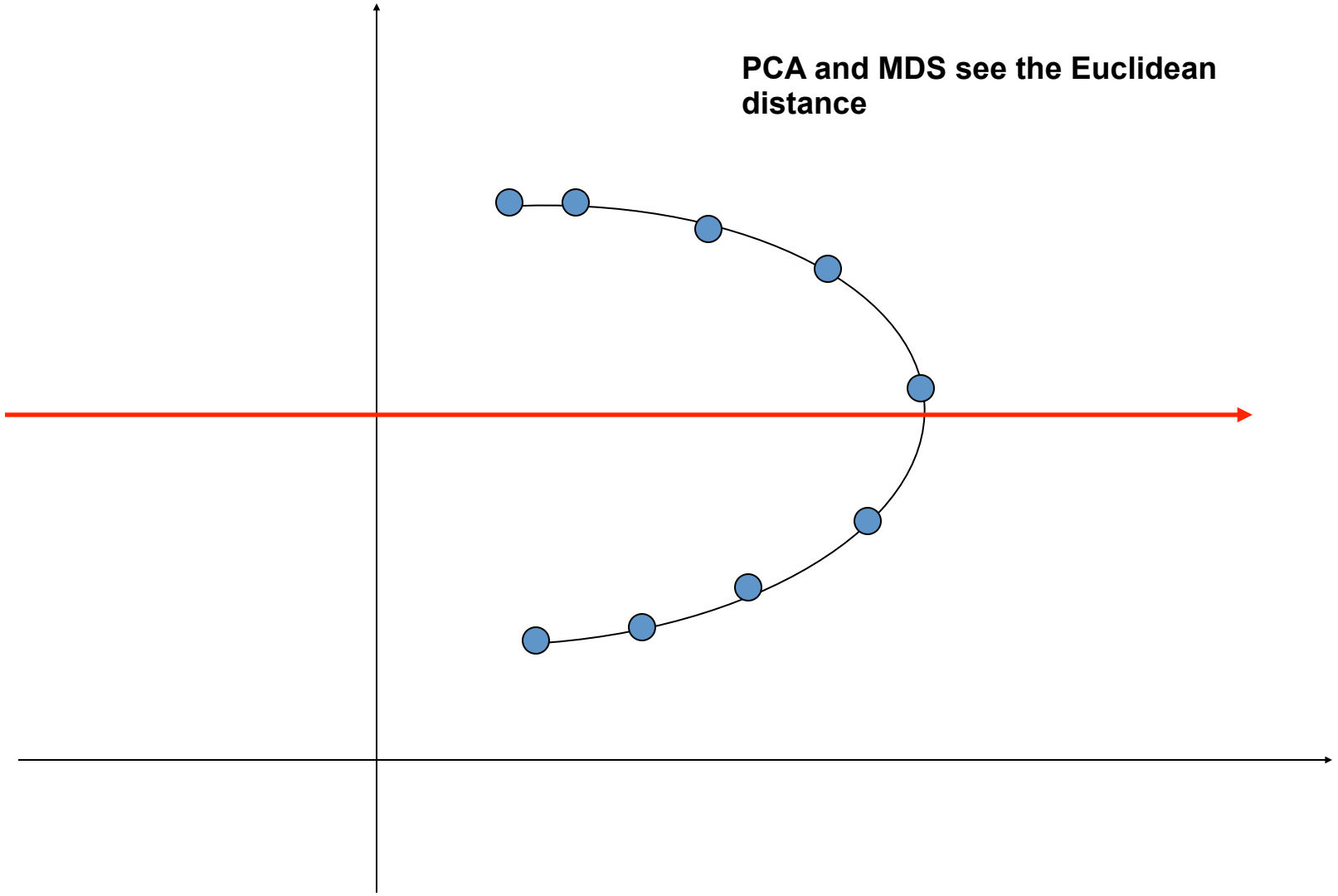
Nonlinear Manifolds..



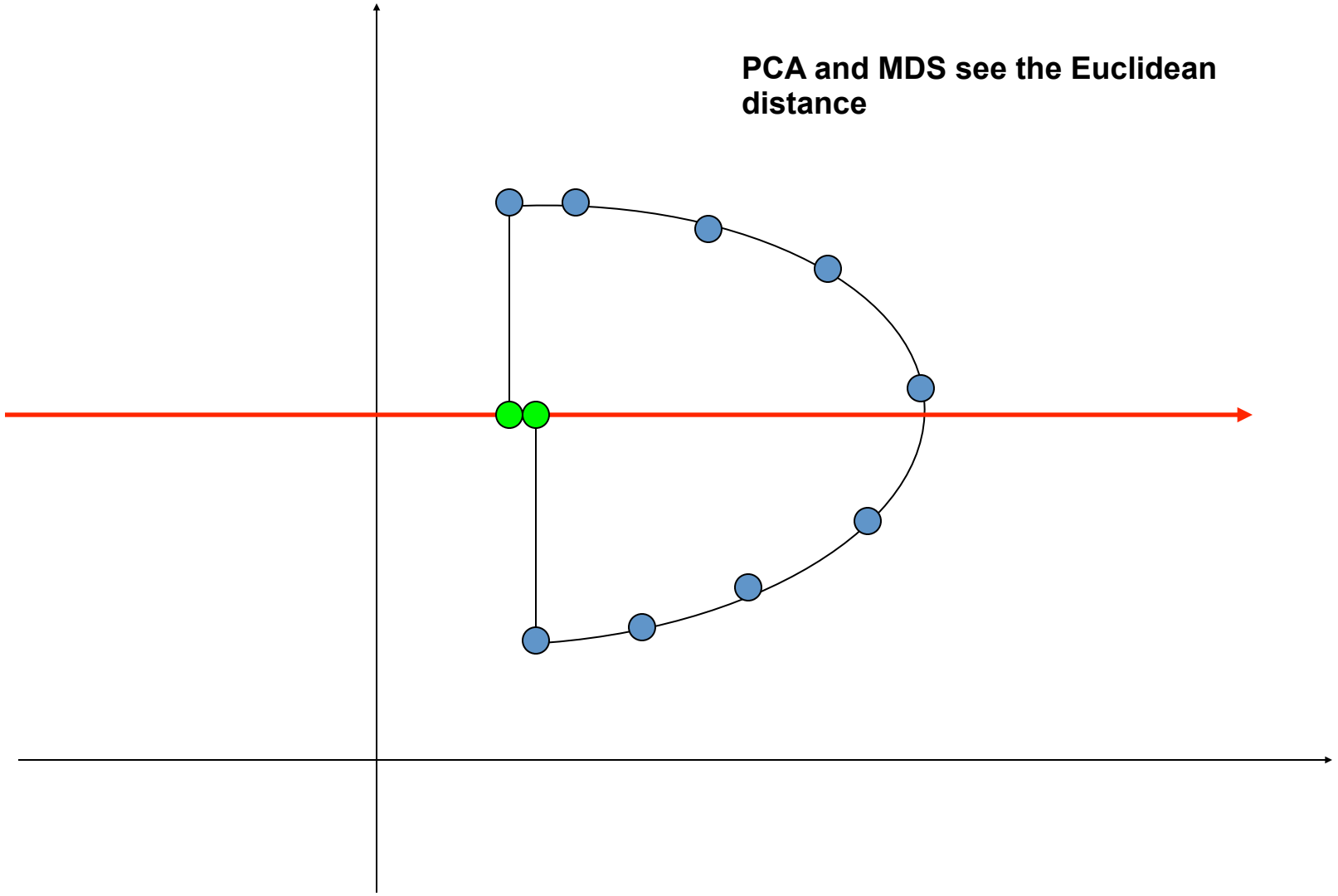
Nonlinear Manifolds..



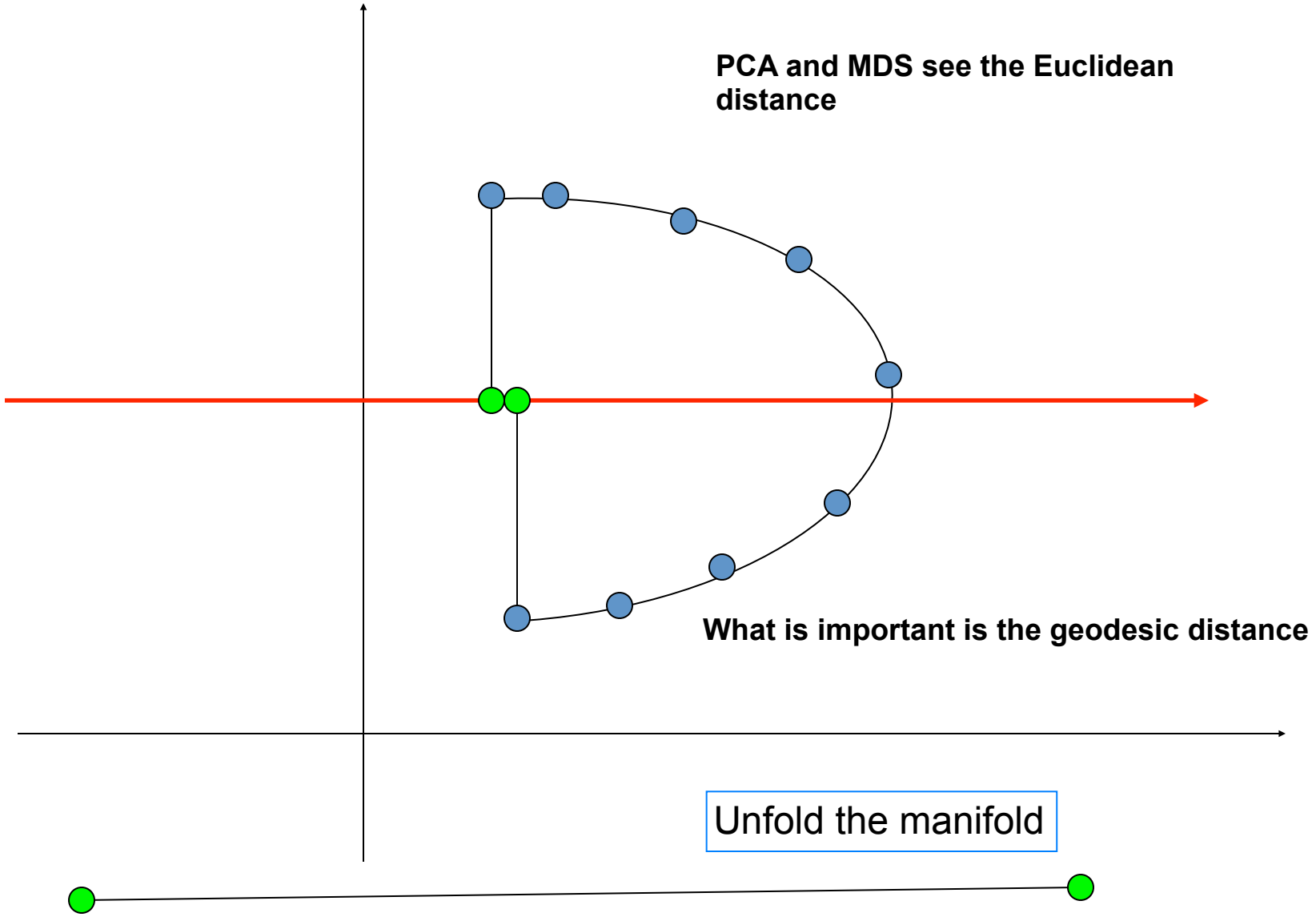
Nonlinear Manifolds..



Nonlinear Manifolds..

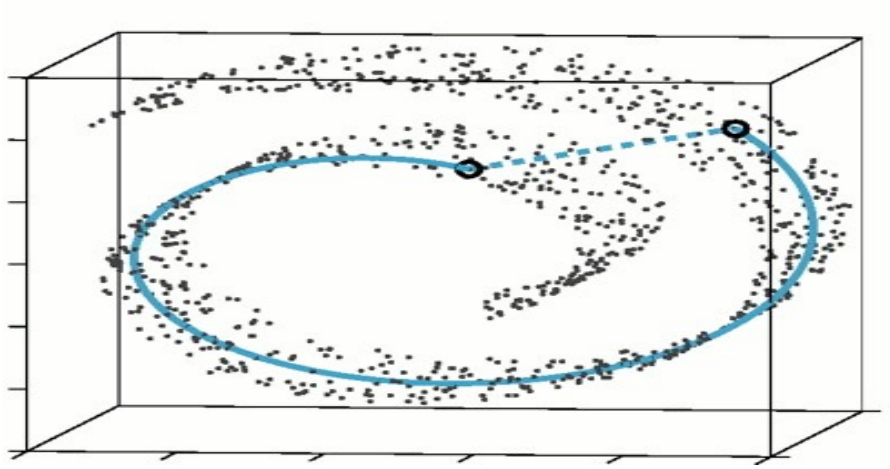


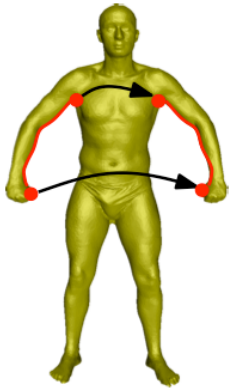
Nonlinear Manifolds..



Intrinsic Description..

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





Manifold Learning

Learning when data $\sim \mathcal{M} \subset \mathbb{R}^N$

- Clustering: $\mathcal{M} \to \{1, \ldots, k\}$

  connected components, min cut

- Classification/Regression: $\mathcal{M} \to \{-1, +1\}$ or $\mathcal{M} \to \mathbb{R}$

  P on $\mathcal{M} \times \{-1, +1\}$ or P on $\mathcal{M} \times \mathbb{R}$

- Dimensionality Reduction: $f : \mathcal{M} \to \mathbb{R}^n \quad n << N$

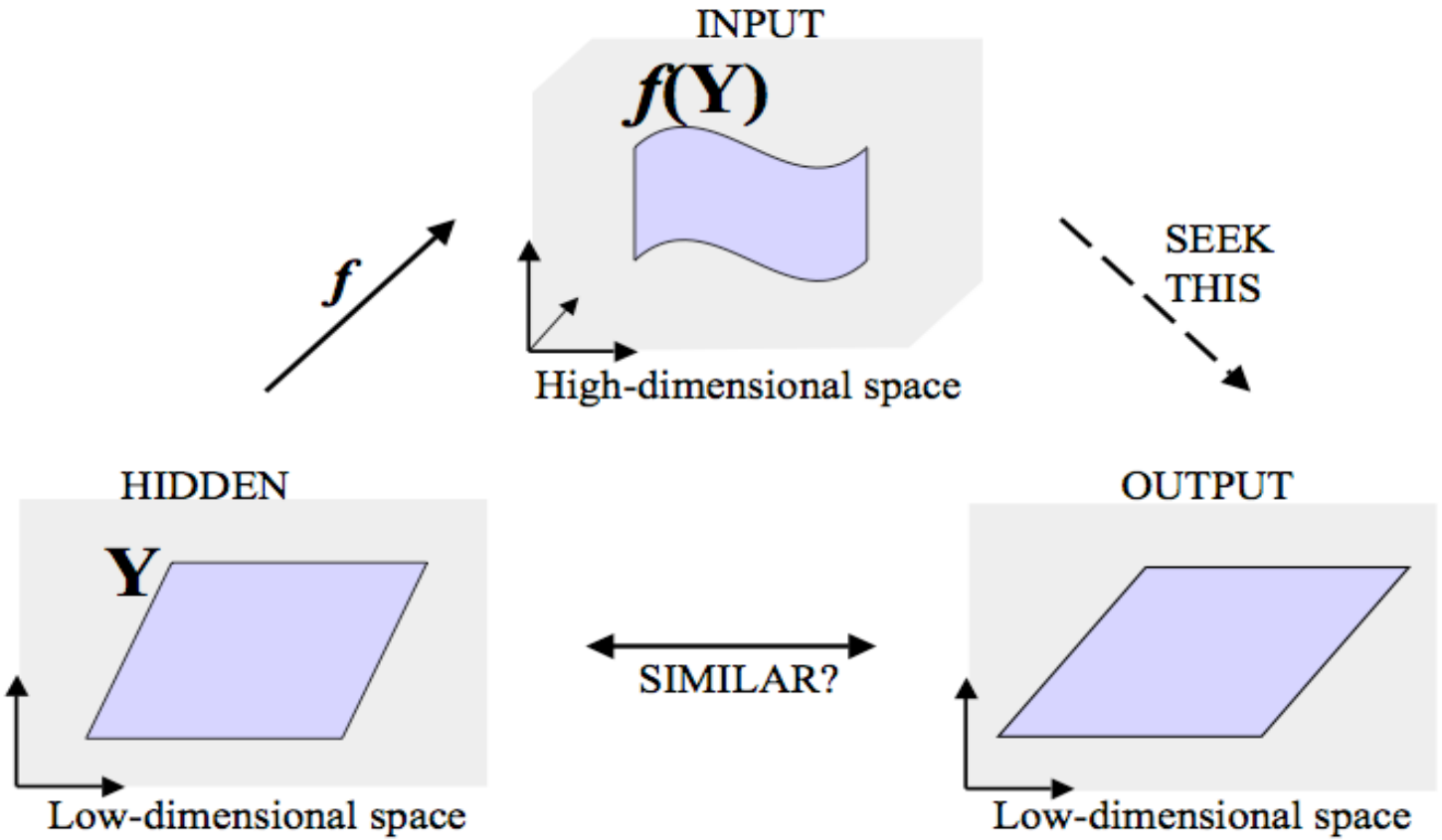
- $\mathcal{M}$ unknown: what can you learn about $\mathcal{M}$ from data?

  e.g. dimensionality, connected components

  holes, handles, homology

  curvature, geodesics

Generative Models in Manifold Learning



Spectral Geometric Embedding

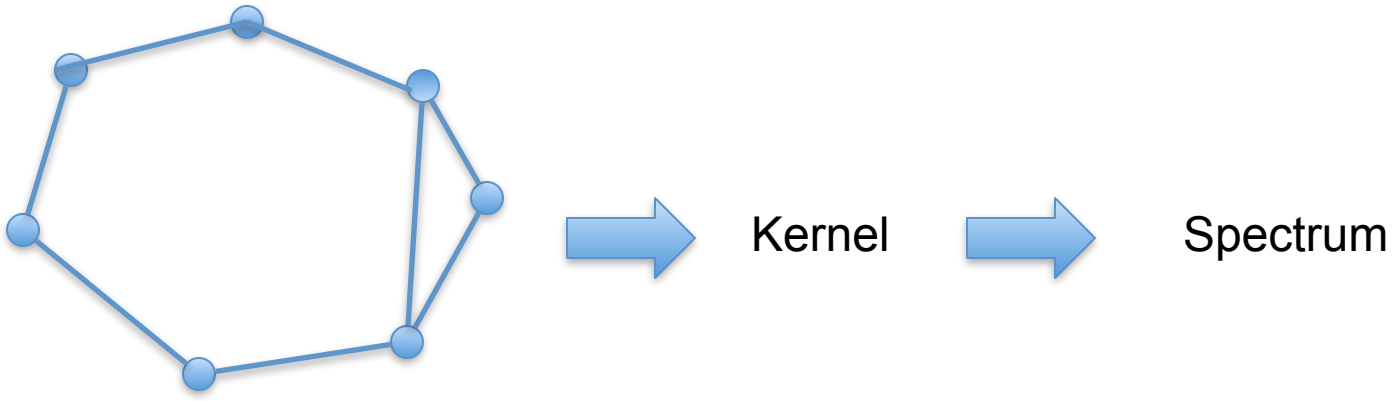
Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$,
Find $y_1, \ldots, y_n \in \mathbb{R}^d$ where $d << N$

- ISOMAP (Tenenbaum, et al, 00)
- LLE (Roweis, Saul, 00)
- Laplacian Eigenmaps (Belkin, Niyogi, 01)
- Local Tangent Space Alignment (Zhang, Zha, 02)
- Hessian Eigenmaps (Donoho, Grimes, 02)
- Diffusion Maps (Coifman, Lafon, et al, 04)

Related: Kernel PCA (Schoelkopf, et al, 98)

Meta-Algorithm

- Construct a neighborhood graph
- Construct a positive semi-definite kernel
- Find the spectrum decomposition



Two Basic Geometric Embedding Methods: Science 2000

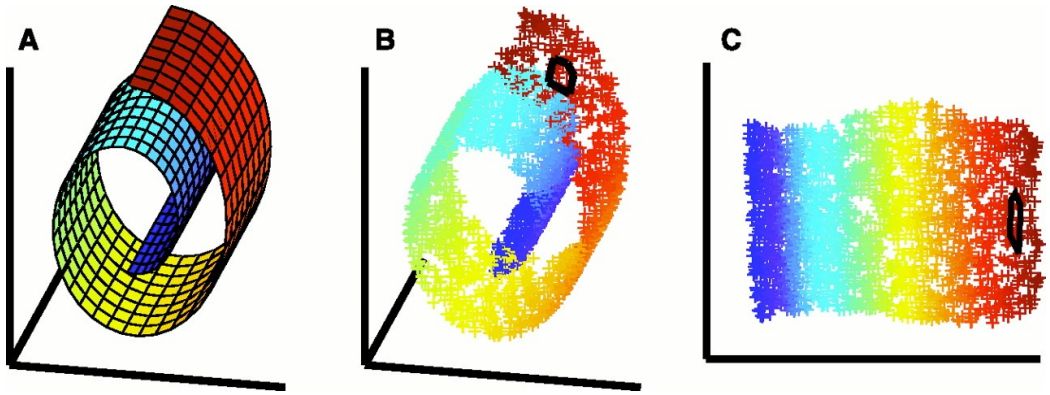
- Tenenbaum-de Silva-Langford Isomap Algorithm
  - Global approach.
  - On a low dimensional embedding
    - Nearby points should be nearby.
    - Faraway points should be faraway.

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

Isomap

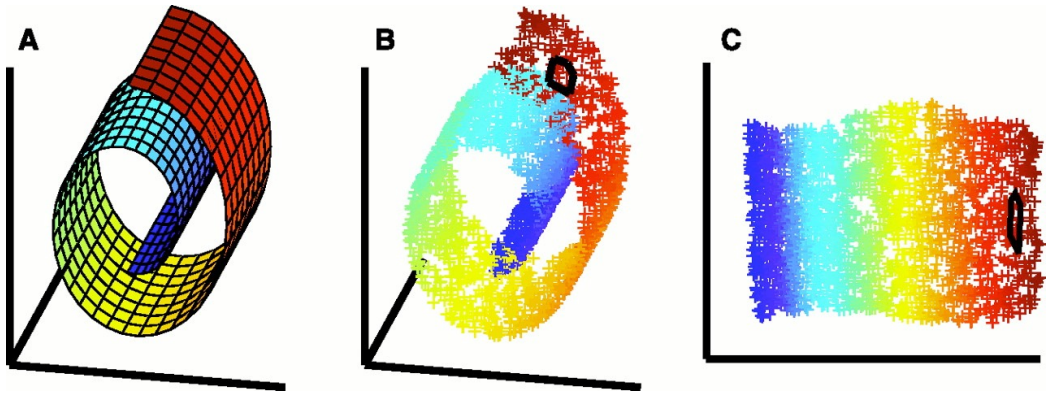
Isomap

- **Estimate the geodesic distance between faraway points.**



Isomap

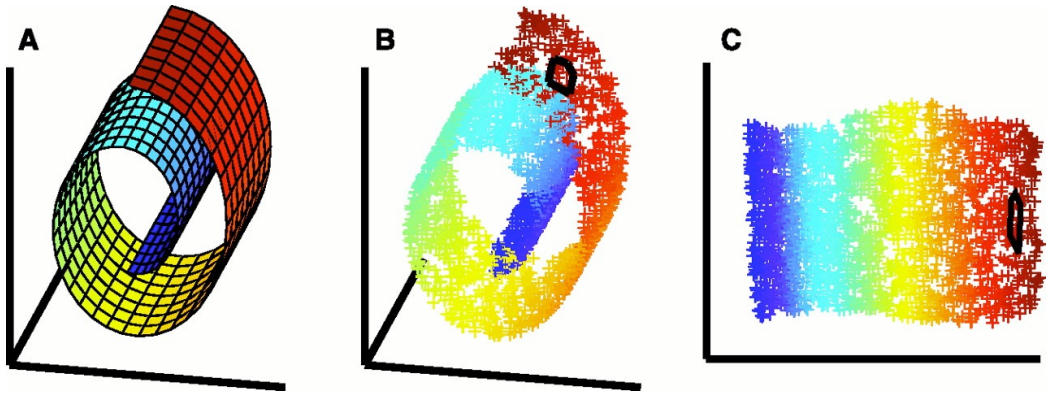
- **Estimate the geodesic distance between faraway points.**
- For neighboring points Euclidean distance is a good approximation to the geodesic distance.
- For faraway points estimate the distance by a series of short hops between neighboring points.
  - Find shortest paths in a graph with edges connecting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



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

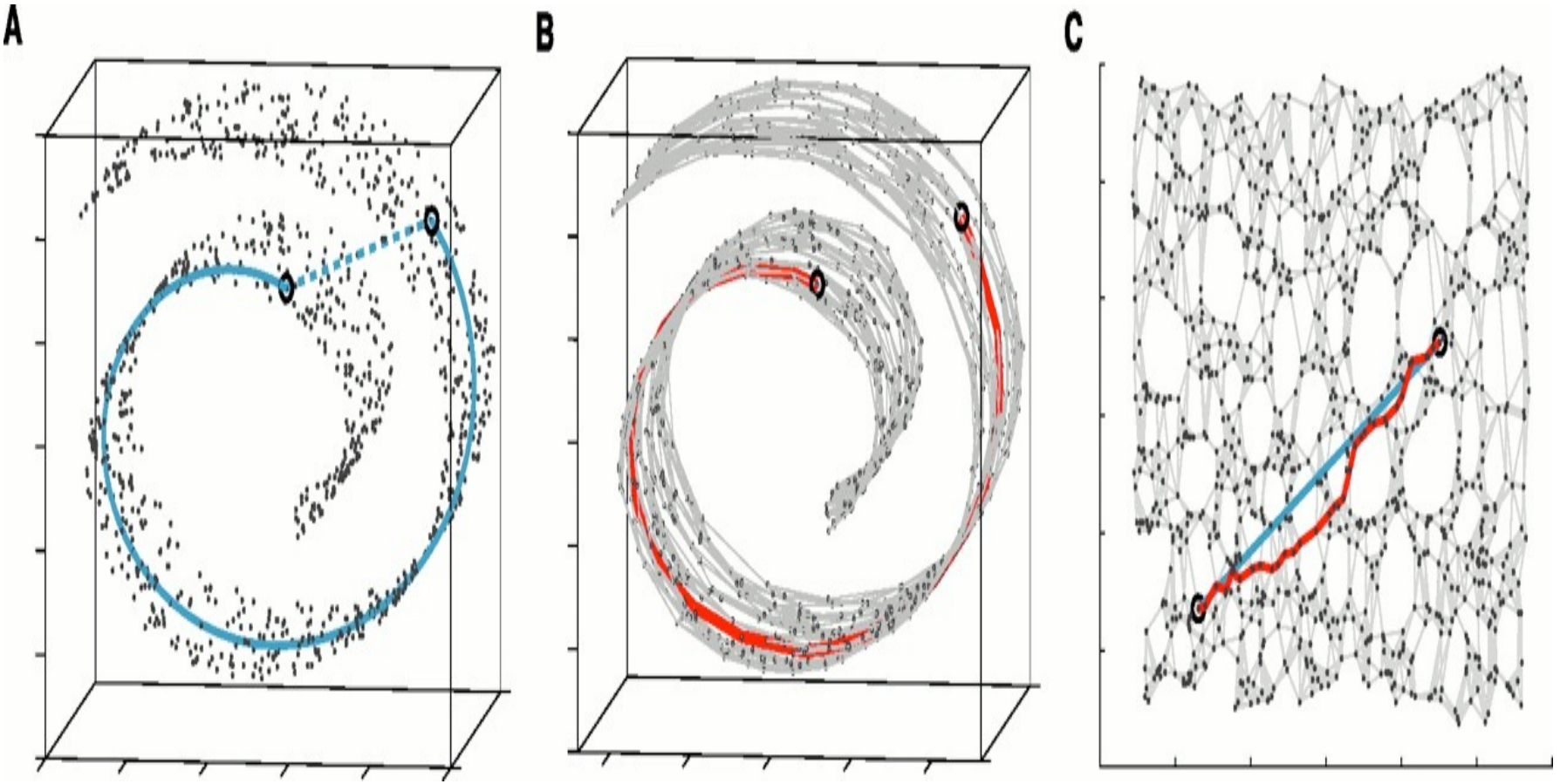
Once we have all pairwise geodesic distances use classical metric MDS

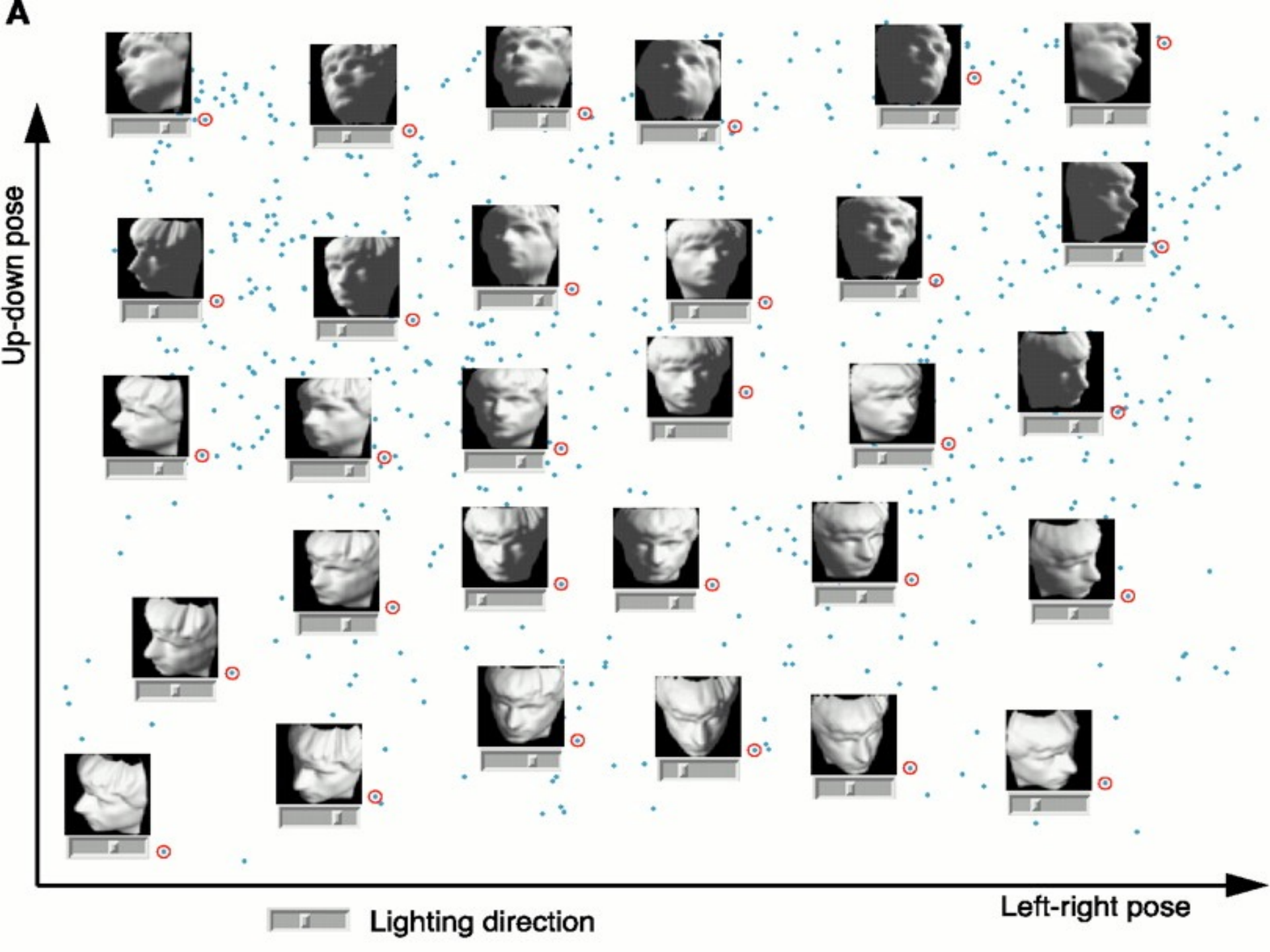


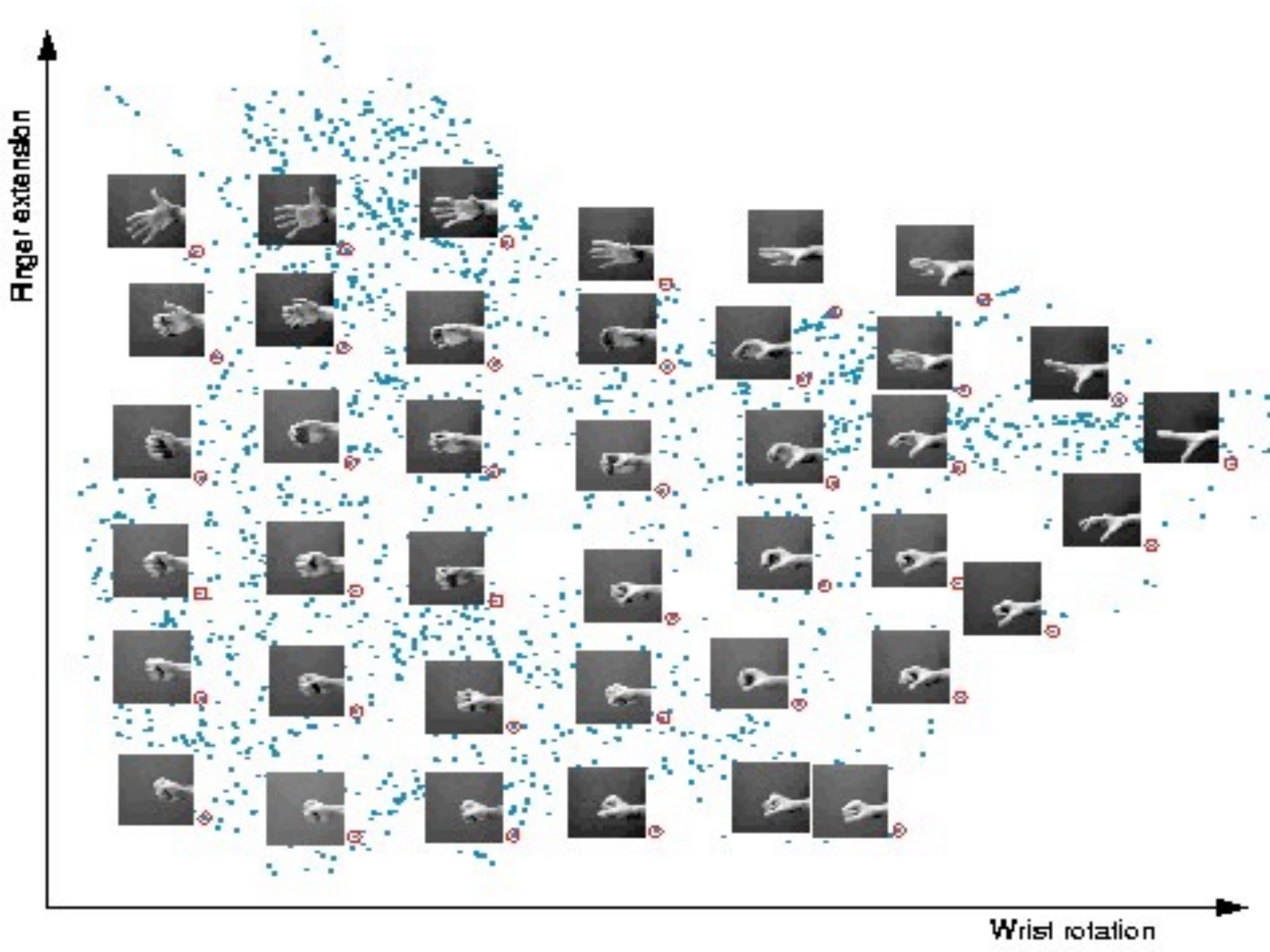
Isomap - Algorithm

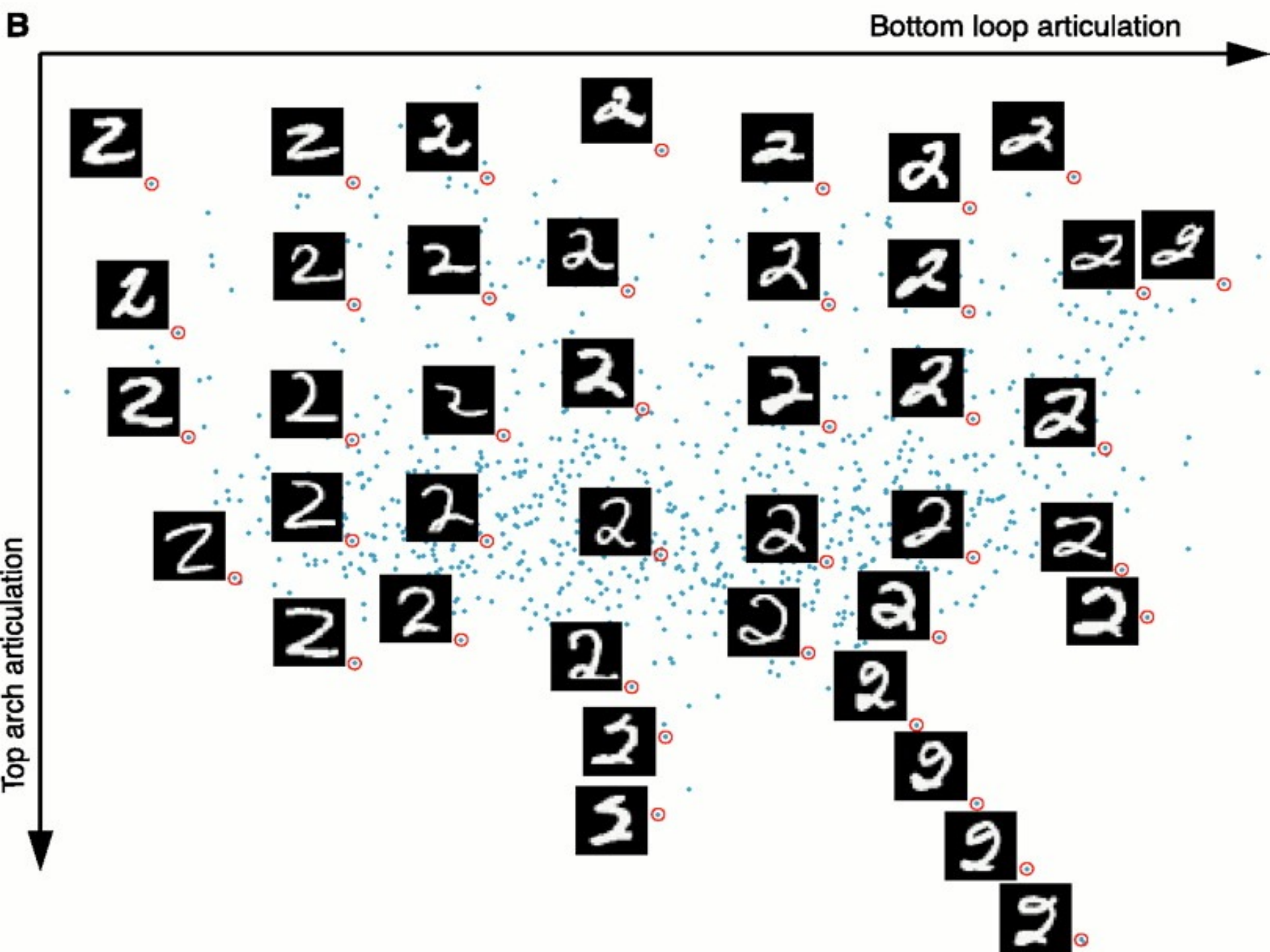
- Construct an n-by-n neighborhood graph
  - connecting points whose distances are within a fixed radius.
  - K nearest neighbor graph
- Compute the shortest path (geodesic) distances between nodes: D
  - Floyd's Algorithm (O(N^3))
  - Dijkstra's Algorithm (O($kN^2 logN$))
- Construct a lower dimensional embedding.
  - Classical MDS (K = -0.5 H D H' = U S U')

Isomap

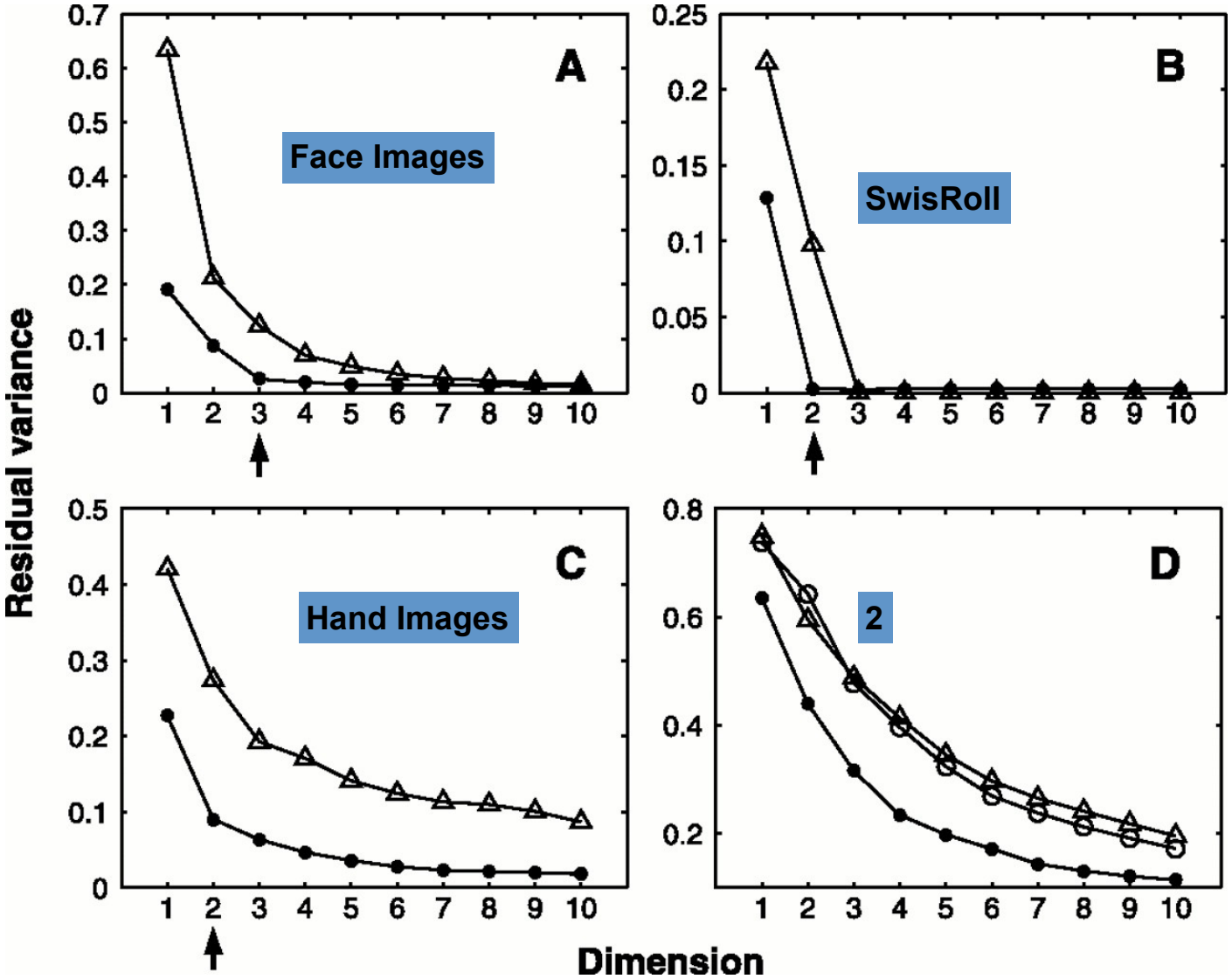



A
Up-down pose
Left-right pose
Lighting direction

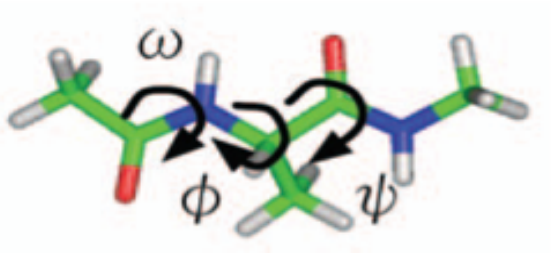

Finger extension
Wrist rotation

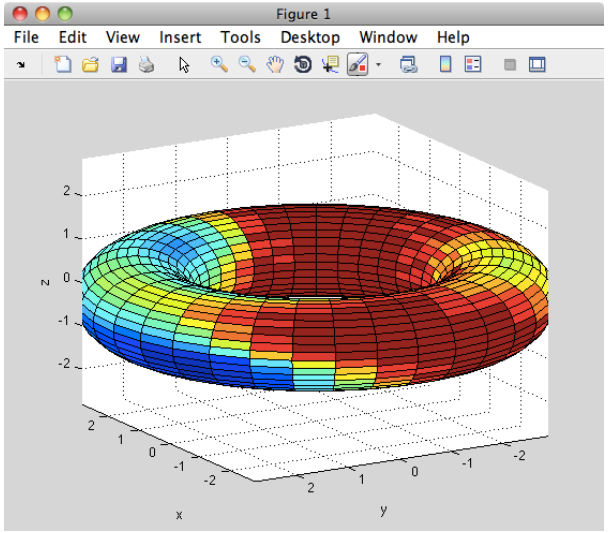

B
Bottom loop articulation
Top arch articulation

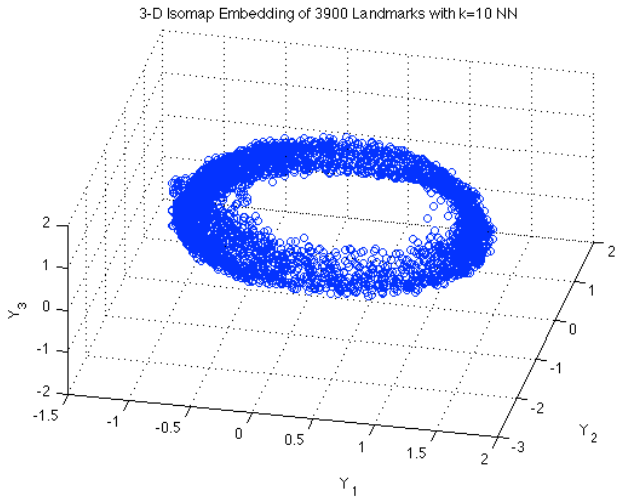
Residual Variance vs. Intrinsic Dimension

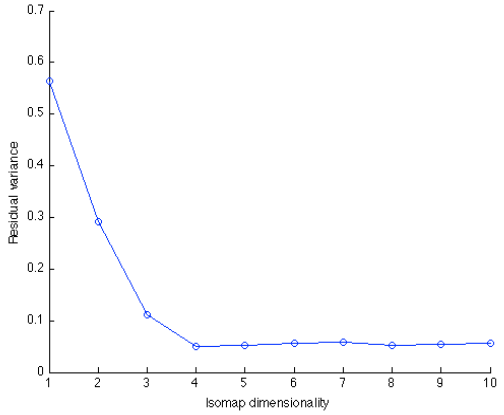


ISOMAP on Alanine-dipeptide









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, graph shortest path distance will approximate closely the original geodesic distance as measured in manifold M;
- But ISOMAP may suffer from nonconvexity such as holes on manifolds

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 $d_G \approx d_M$.

Convergence Theorem [Bernstein, de Silva, Langford,

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 $O(N^3)$

Landmark ISOMAP: Nystrom Extension Method

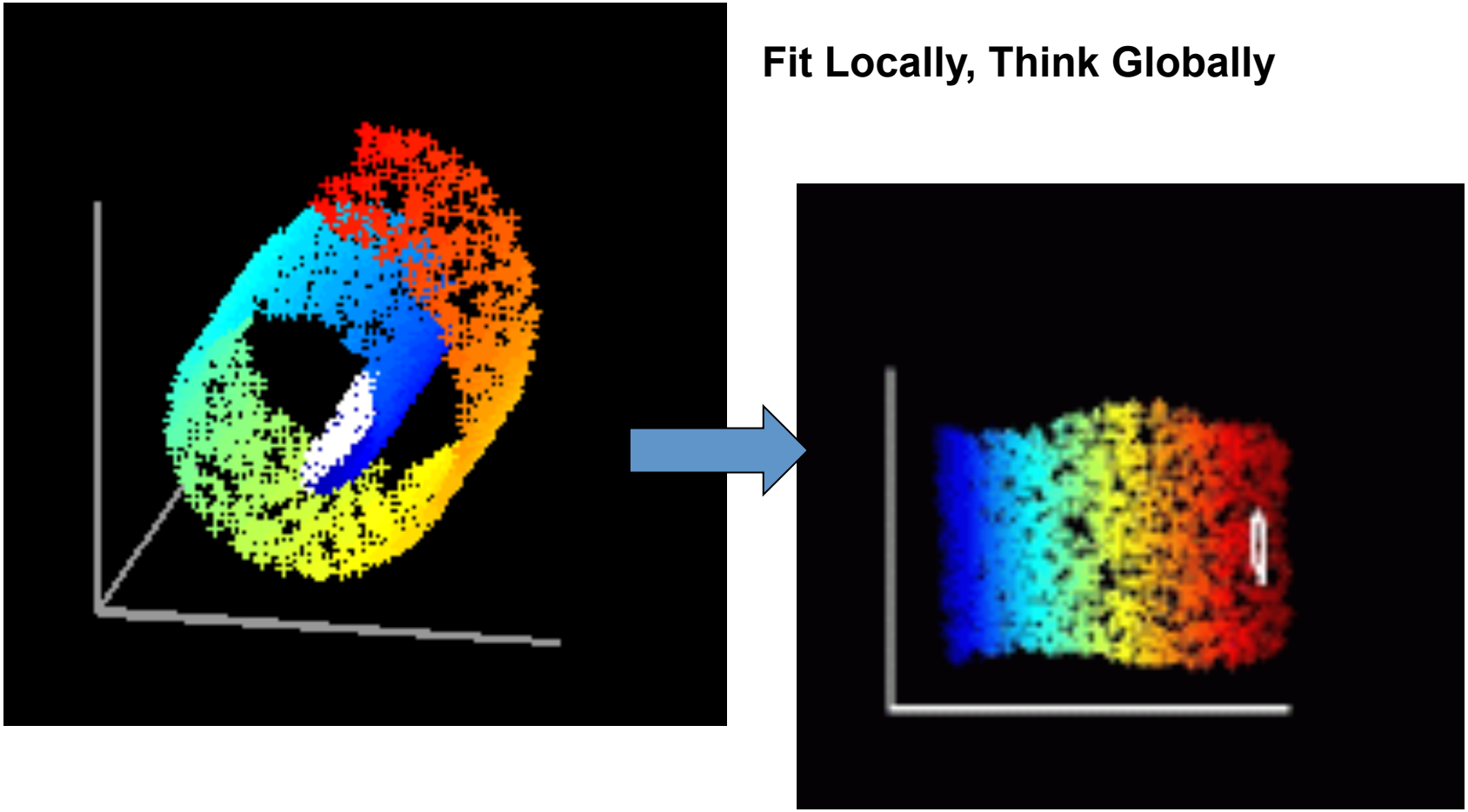
- ISOMAP out of the box is not scalable. Two bottlenecks:
  - All pairs shortest path - $O(kN^2 \log N)$.
  - MDS eigenvalue calculation on a full NxN matrix - $O(N^3)$.
  - For contrast, LLE is limited by a sparse eigenvalue computation - $O(dN^2)$.
- Landmark ISOMAP (L-ISOMAP) Idea:
  - Use $n << N$ *landmark* points from $\{x_i\}$ and compute a n *x* N matrix of geodesic distances, D_n, from each data point to the landmark points only.
  - Use new procedure Landmark-MDS (LMDS) to find a Euclidean embedding of all the data – utilizes idea of triangulation similar to GPS.
- Savings: L-ISOMAP will have shortest paths calculation of $O(knN \log N)$ and LMDS eigenvalue problem of $O(n^2 N)$.

Landmark Choice

- Random
- MiniMax: k-center
- Hierarchical landmarks: cover-tree
- Nyström extension method

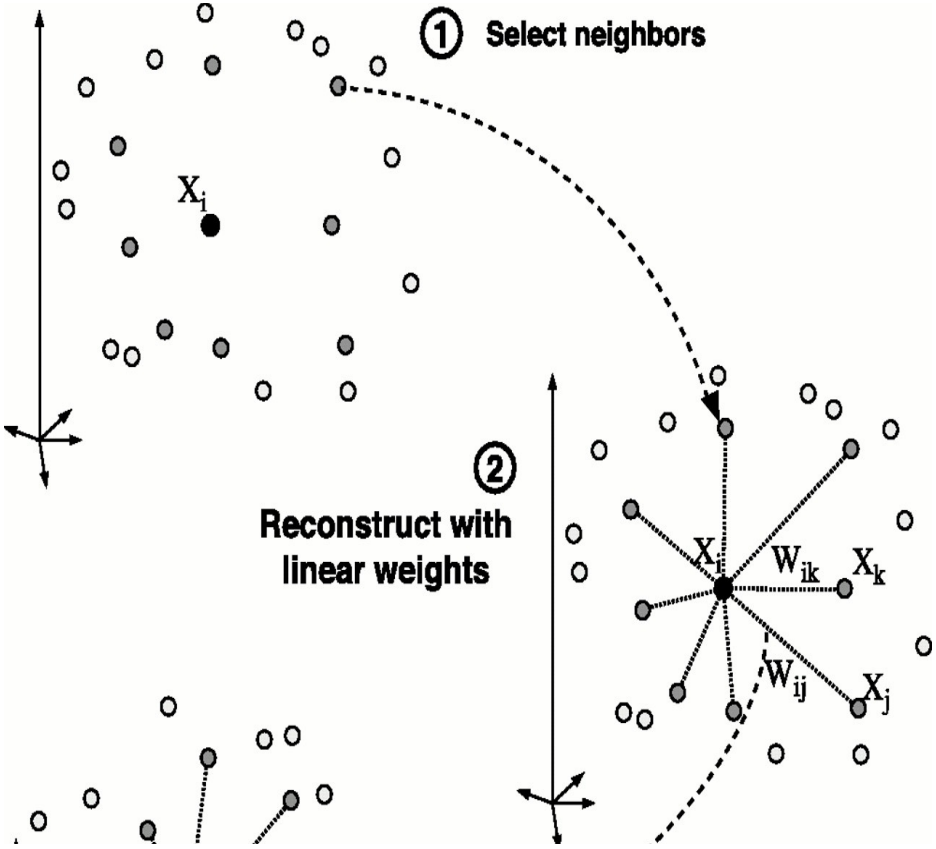
Locally Linear Embedding

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 are chosen to minimize the reconstruction Error.

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

Derivation on board

Important property…

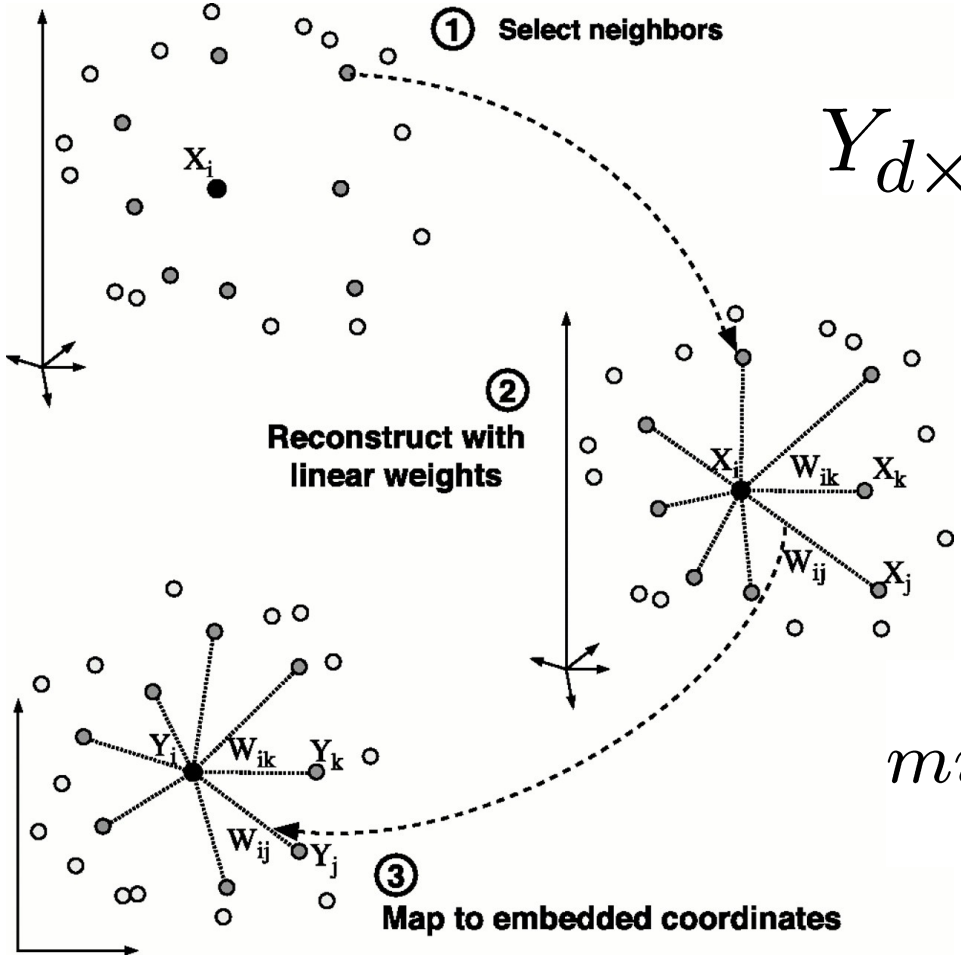
Important property…

- 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.

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…



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

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

LLE Algorithm (I)

(1) Construct a neighborhood graph $G = (V, E)$ such that

$V = \{x_i : i = 1, \ldots, n\}$

$E = \{(i, j) : \text{ if } j \text{ is a neighbor of } i, \text{ i.e. } j \in \mathcal{N}_i\}$, *e.g.* k-nearest neighbors, ϵ-neighbors

(2) Local fitting:

Pick up a point x_i and its neighbors $\mathcal{N}_i$

Compute the local fitting weights

$$\min_{\sum_{j \in \mathbb{N}_i} w_{ij} = 1} \|x_i - \sum_{j \in \mathcal{N}_i} w_{ij} x_j\|^2,$$

LLE Algorithm (II)

(2) Local fitting:

Pick up a point x_i and its neighbors $\mathcal{N}_i$

Compute the local fitting weights

$$\min_{\sum_{j\in\mathbb{N}_i} w_{ij}=1} \|x_i - \sum_{j\in\mathcal{N}_i} w_{ij}x_j\|^2,$$

which is equivalent to

$$\min_{\sum_{j\in\mathbb{N}_i} w_{ij}=1} \| \sum_{j\in\mathcal{N}_i} w_{ij}(x_j - x_i)\|^2,$$

that is, finding a linear combination (possibly *not unique*!) for the subspace spanned by $\{(x_j - x_i) : j \in \mathcal{N}_i\}$. This can be done by Lagrange multiplier method, *i.e.* solving

$$\min_{w_{ij}} \frac{1}{2}\| \sum_{j\in\mathcal{N}_i} w_{ij}(x_j - x_i)\|^2 + \lambda(1 - \sum_{j\in\mathcal{N}_i} w_{ij}).$$

Let $w_i = [w_{ij_1}, \ldots w_{ij_k}]^T \in \mathbb{R}^k$, $\bar{X}_i = [x_{j_1} - x_i, \ldots, x_{j_k} - x_i]$, and the local Gram (covariance) matrix $C_i(j,k) = \langle x_j - x_i, x_k - x_i\rangle$, whence the weights are

$$w_i = \lambda C_i^\dagger \mathbf{1}, \tag{80}$$

where the Lagrange multiplier equals to the following normalization parameter

$$\lambda = \frac{1}{\mathbf{1}^T C_i^\dagger \mathbf{1}}, \tag{81}$$

and $C_i^\dagger$ is a Moore-Penrose (pseudo) inverse of C_i. Note that C_i is often ill-conditioned and to find its Moore-Penrose inverse one can use regularization method $(C_i + \mu I)^{-1}$ for some $\mu > 0$.

LLE Algorithm (III)

(3) Global alignment

Define a n-by-n weight matrix W:

$$W_{ij} = \begin{cases} w_{ij}, & j \in \mathcal{N}_i \\ 0, & otherwise \end{cases}$$

Compute the global embedding d-by-n embedding matrix Y,

$$\min_Y \sum_i \|y_i - \sum_{j=1}^{n} W_{ij} y_j\|^2 = \text{trace}(Y(I-W)^T(I-W)Y^T)$$

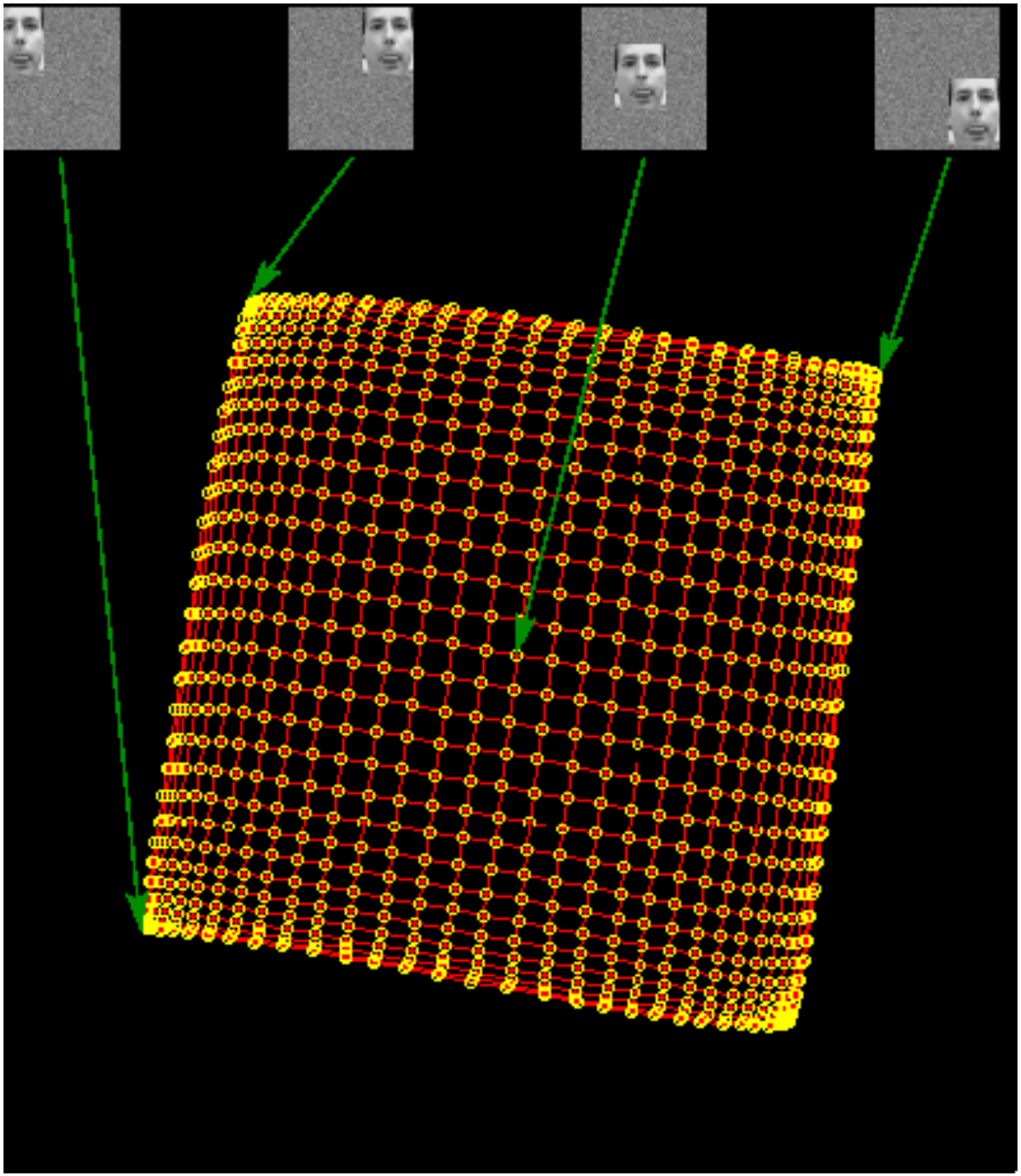
In other words, construct a positive semi-definite matrix $B = (I - W)^T(I-W)$ and find $d+1$ smallest eigenvectors of B, $v_0, v_1, \ldots, v_d$ associated smallest eigenvalues $\lambda_0, \ldots, \lambda_d$. Drop the smallest eigenvector which is the constant vector explaining the degree of freedom as translation and set $Y = [v_1/\sqrt{(}\lambda_1), \ldots, v_d/\sqrt{\lambda_d}]^T$.

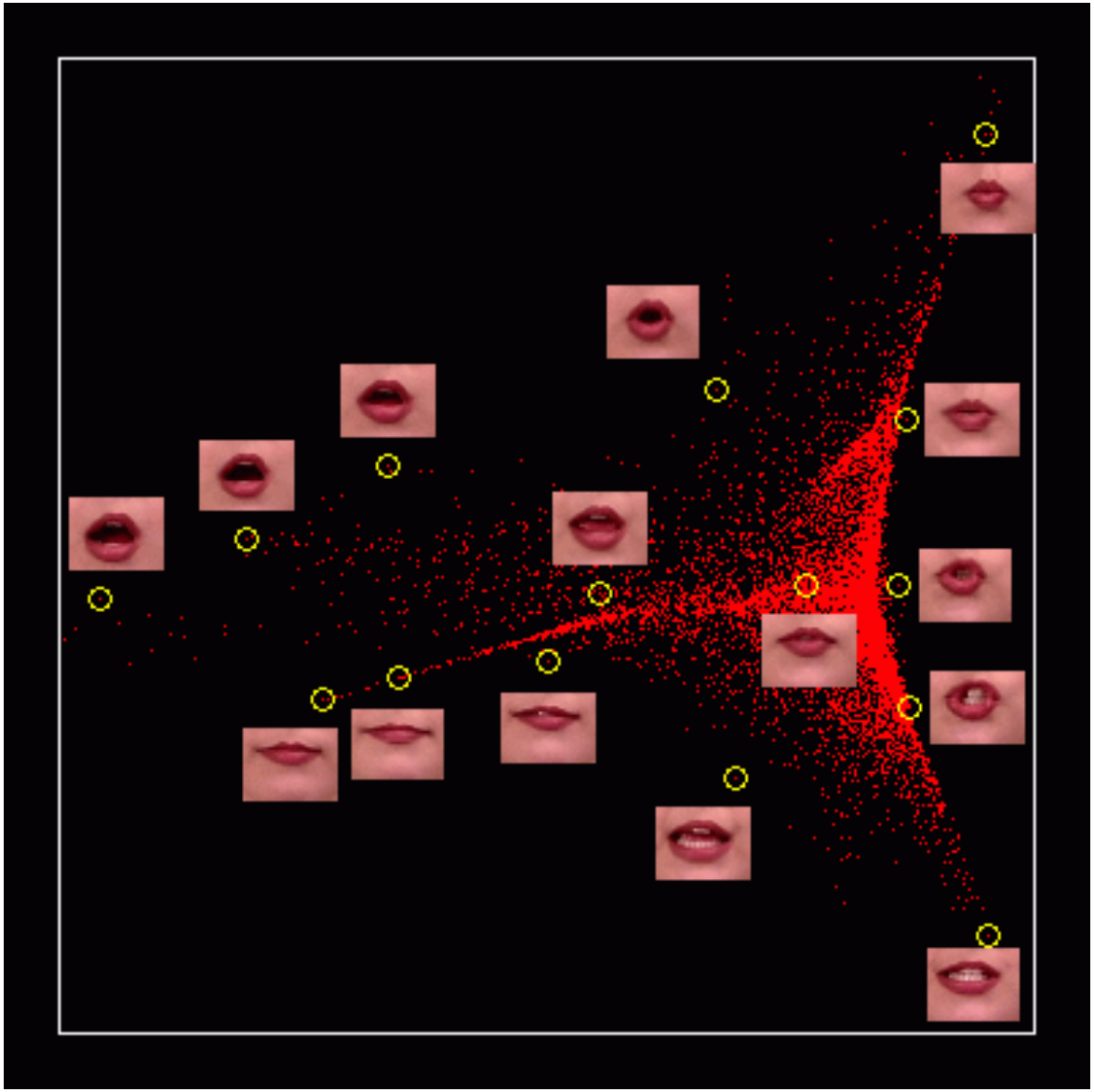
Remarks on LLE

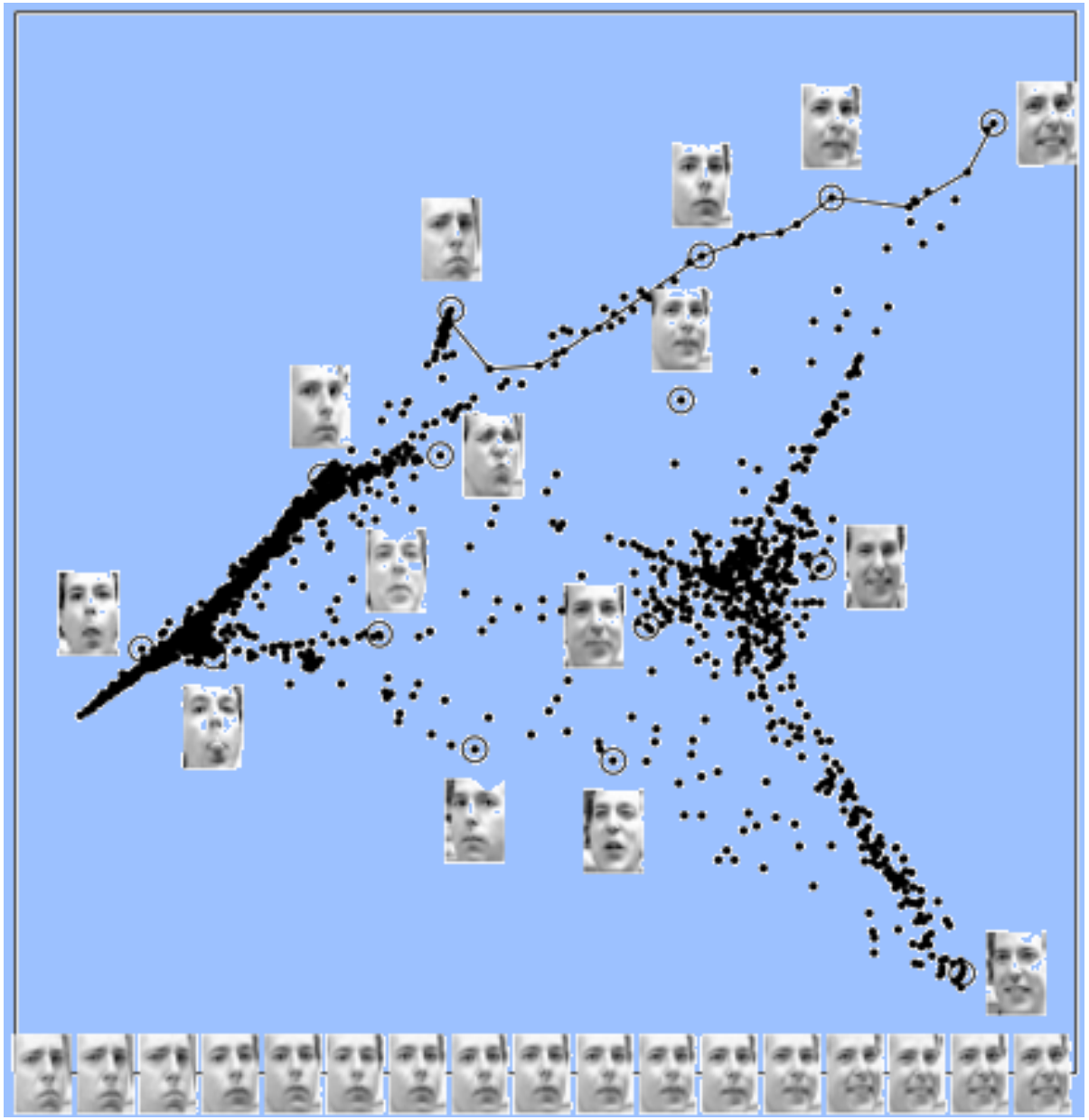
- Searching k-nearest neighbors is of O(kN)

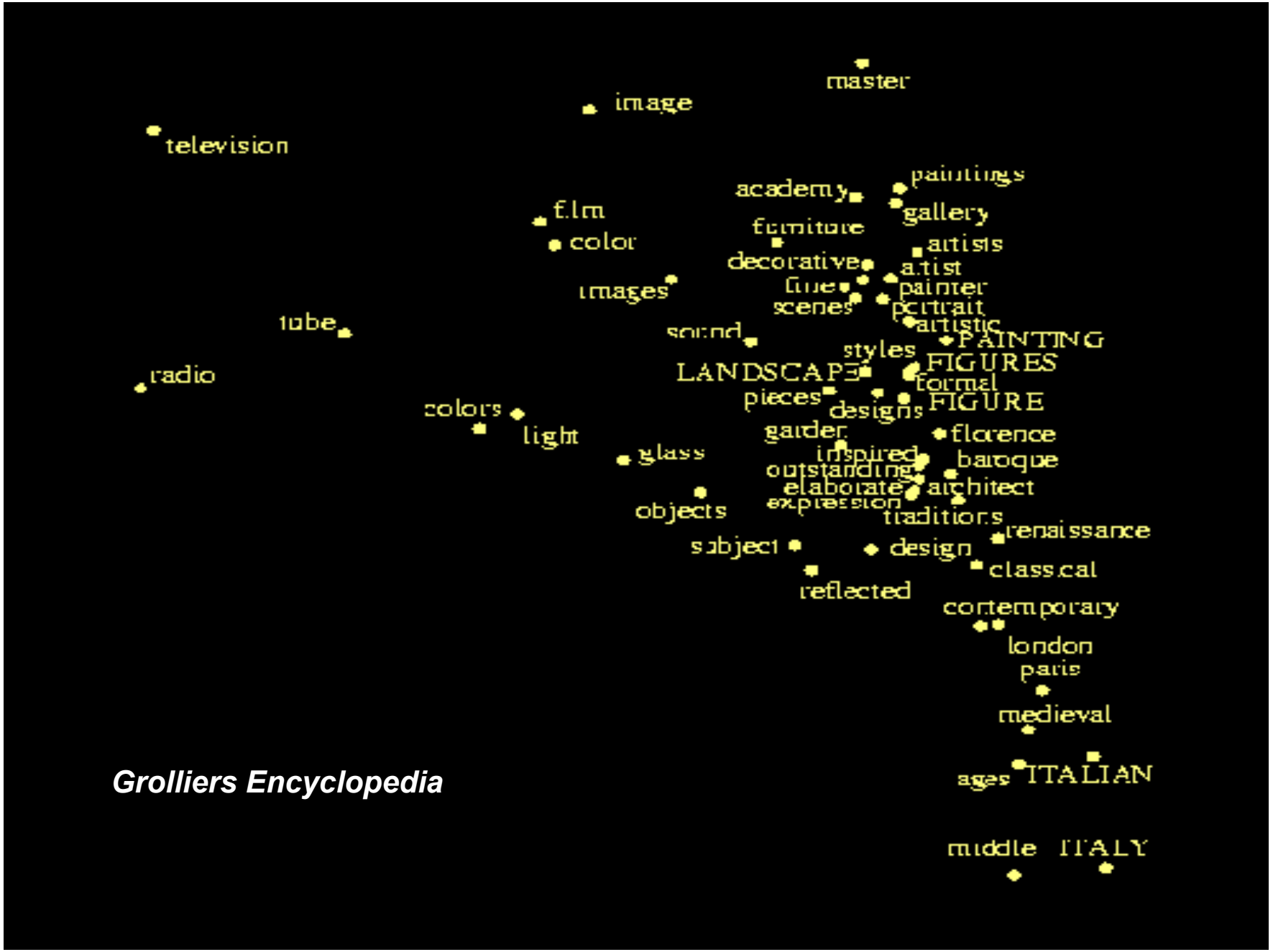
- W is sparse, kN/N^2=k/N nozeros
- W might be negative, additional nonnegative constraint can be imposed
- $B=(I-W)^T(I-W)$ is positive semi-definite (p.s.d.)

- Open Problem: exact reconstruction condition?









Grolliers Encyclopedia

Issues of LLE

Pick up a point x_i and its neighbors $\mathcal{N}_i$. Compute the local fitting weights

$$\min_{\sum_{j\in\mathbb{N}_i} w_{ij}=1} \|x_i - \sum_{j\in\mathcal{N}_i} w_{ij}x_j\|^2,$$

which is equivalent to

$$\min_{\sum_{j\in\mathbb{N}_i} w_{ij}=1} \| \sum_{j\in\mathcal{N}_i} w_{ij}(x_j - x_i)\|^2,$$

$$\min_{w_{ij}} \frac{1}{2}\| \sum_{j\in\mathcal{N}_i} w_{ij}(x_j - x_i)\|^2 + \lambda(1 - \sum_{j\in\mathcal{N}_i} w_{ij}).$$

$$w_i = \lambda C_i^{\dagger}\mathbf{1},$$

$$\lambda = \frac{1}{\mathbf{1}^T C_i^{\dagger}\mathbf{1}}, \qquad C_i(j,k) = \langle x_j - x_i, x_k - x_i\rangle$$

ill-posed or ill-conditioned?

Issues of LLE

$$w_i(\mu) = \lambda(C_i + \mu I)^{-1}\mathbf{1} = \sum_j \frac{1}{\lambda_j^{(i)} + \mu} v_j v_j^T \mathbf{1} \quad (82)$$

where the local PCA $C_i = V\Lambda V^T$ ($\Lambda = \mathrm{diag}(\lambda_j^{(i)})$, $V = [v_j]$).

- Low-pass filter of constant 1-vector
  - preserve projections on bottom eigenvectors associated with small eigenvalues $\lambda_j^{(i)} \ll \mu$
  - suppress projections on top eigenvectors associated with large eigenvalues
- If 1-vector is not so well-spread over null eigenspace, instability and missing directions as mu goes down!

Modified LLE (MLLE)

- Use all the null eigenspace!

MLLE replace the weight vector above by a weight matrix $W_i \in \mathbb{R}^{k_i \times s_i}$, a family of s_i weight vectors using bottom s_i eigenvectors of C_i, $V_i = [v_{k_i - s_i + 1}, \ldots, v_{k_i}] \in \mathbb{R}^{k_i \times s_i}$, such that

$$W_i = (1 - \alpha_i) w_i(\mu) \mathbf{1}_{s_i}^T + V_i H_i^T, \tag{83}$$

where $\alpha_i = \|V_i^T \mathbf{1}_{k_i}\|_2 / \sqrt{s_i}$ and $H_i = I_{s_i} - 2uu^T$ ($\|u\|_2 = 1$ or 0) is a Householder matrix ($H_i := I_{s_i}$ if $u = 0$) such that $HV_i^T \mathbf{1}_{k_i} = \alpha_i \mathbf{1}_{s_i}$ (hence $W_i^T \mathbf{1}_{k_i} = \mathbf{1}_{s_i}$, every column of W_i being a legal weight vector). In fact, one can choose u in the direction of $V_i^T \mathbf{1}_{k_i} - \alpha_i \mathbf{1}_{s_i}$. An adaptive choice of s_i is given in [ZW] using the

$$\min_Y \sum_i \sum_{l=1}^{s_i} \|y_i - \sum_{j \in \mathcal{N}_i} W_i(j,l) y_j\|^2 := \sum_i \|Y \widehat{W}_i\|_F^2 = \text{trace}[Y (\sum_i \widehat{W}_i \widehat{W}_i^T) Y^T]$$

where $\widehat{W}_i$ is the embedding of $W_i \in \mathbb{R}^{k_i \times s_i}$ into $\mathbb{R}^{n \times s_i}$,

$$\widehat{W}_i(j,:) = \begin{cases} -1_{s_i}^T, & j = i, \\ W_i, & j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases}$$

MLLE Algorithm (II)

Step 2 (local residue PCA): for each x_i and its neighbors $\mathcal{N}_i$ ($k_i = |\mathcal{N}_i|$), let $C_i = V\Lambda V^T$ be its eigenvalue decomposition where $\Lambda = (\lambda_1, \ldots, \lambda_{k_i})$ with $\lambda_1 \geq \cdots \geq \lambda_{k_i}$. Find the size of almost normal subspace s_i as the maximal size that the ratio of residue eigenvalue sum over principle eigenvalue sum is below a threshold, *i.e.*

$$s_i = \max_l \left\{ l \leq k_i - d, \frac{\sum_{j=k_i-l+1}^{k_i} \lambda_j}{\sum_{j=1}^{k_i-l} \lambda_j} \leq \eta \right\}$$

where η is a parameter, such as the median of ratios of residue eigenvalue sum over principle eigenvalue sum. Construct the normal subspace basis matrix as s_i-bottom eigenvector matrix of C_i, $V_i = [v_{k_i-s_i+1}, \ldots, v_{k_i}] \in \mathbb{R}^{k_i \times s_i}$, define the weight matrix

$$W_i = (1 - \alpha_i) w_i(\mu) \mathbf{1}_{s_i}^T + V_i H_i^T \in \mathbb{R}^{k_i \times s_i},$$

where $\alpha_i = \|V_i^T \mathbf{1}_{k_i}\|_2 / \sqrt{s_i}$ and $H_i = I_{s_i} - 2uu^T / \|u\|^2$ with $u = V_i^T \mathbf{1}_{k_i} - \alpha_i \mathbf{1}_{s_i}$ (or $u = 0$ if it is small).

MLLE Algorithm (III)

Step 3 (global alignment): define the weight embedding matrix

$$\widehat{W}_i(j,:) = \begin{cases} -1_{s_i}^T, & j = i, \\ W_i, & j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases}$$

Compute $K = \widehat{W}^T \widehat{W}$ which is a positive semi-definite kernel matrix;
Step 4 (Eigenmap): Compute Eigenvalue decomposition $K = U\Lambda U^T$ with $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{n-1} > \lambda_n = 0$; choose bottom $d+1$ nonzero eigenvalues and corresponding eigenvectors and drop the smallest eigenvalue-eigenvector (0-constant) pair, such that

$$U_d = [u_{n-d}, \ldots, u_{n-1}], \quad u_j \in \mathbb{R}^n,$$

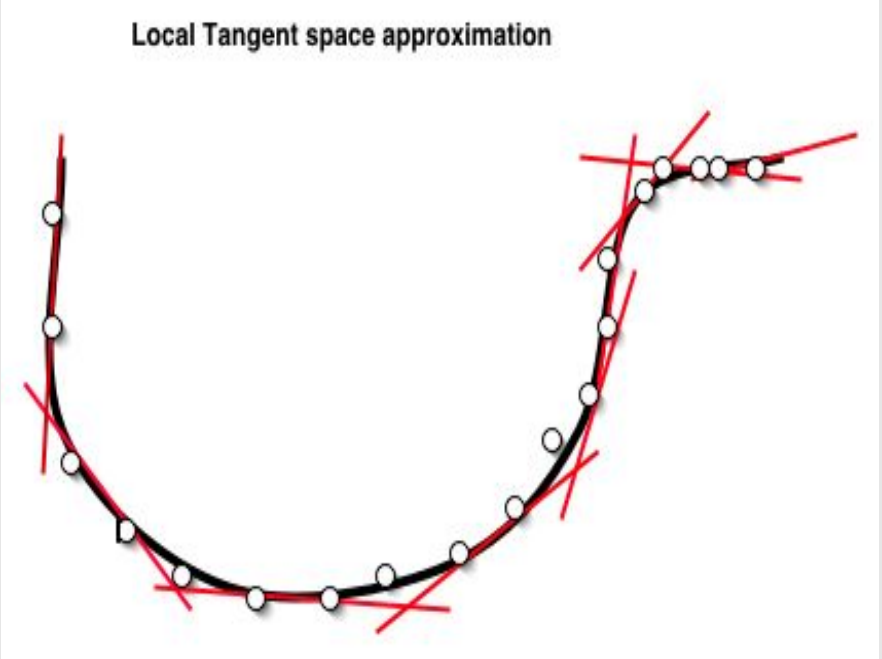
$$\Lambda_d = \text{diag}(\lambda_{n-d}, \ldots, \lambda_{n-1}).$$

Define $Y_d = U_d \Lambda_d{}^{\frac{1}{2}}$.

Issues of MLLE

- MLLE computes bottom eigenvectors of local Gram (Covariance) matrix, expensive in computation and sensitive to noise
- How about only using top eigenvectors in local PCA?
  - LTSA
  - Hessian LLE

Local Tangent Space Alignment



Find a good approximation of tangent space of curve using discrete samples.
— Principal curve/manifold (Hastie-Stuetzle'89, Zha-Zhang'02)

Local SVD

For each x_i in $\mathbb{R}^d$ with neighbor $\mathcal{N}_i$ of size $|\mathcal{N}_i| = k_i - 1$, let $X^{(i)} = [x_{j_1}, x_{j_2}, \ldots, x_{j_{k_i}}] \in \mathbb{R}^{p \times k_i}$ be the coordinate matrix. Consider the local SVD (PCA)

$$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_{k_i}} - \mu_i]^{p \times k_i} = X^{(i)} H = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T,$$

where $H = I - \frac{1}{k_i} \mathbf{1}_{k_i} \mathbf{1}_{k_i}^T$. Left singular vectors $\{\tilde{U}_1^{(i)}, ..., \tilde{U}_d^{(i)}\}$ give an orthonormal basis of the approximate d-dimensional tangent space at x_i. Right singular vectors $(\tilde{V}_1^{(i)}, \ldots, \tilde{V}_d^{(i)}) \cdot \tilde{\Sigma} \in \mathbb{R}^{k_i \times d}$ present the d-coordinates of k_i samples with respect to the tangent space basis.

LTSA

Let $Y_i \in \mathbb{R}^{d \times k_i}$ be the embedding coordinates of the samples in $\mathbb{R}^d$ and $L_i : \mathbb{R}^{p \times d}$ be an estimated basis of the tangent space at x_i in $\mathbb{R}^p$. Let $\Theta_i = \tilde{U}_d^{(i)} \tilde{\Sigma}_d (\tilde{V}_d^{(i)})^T \in \mathbb{R}^{p \times k_i}$ be the truncated SVD using top d components. LTSA looks for the minimizer of the following problem

$$\min_{Y,L} \sum_i \|E_i\|^2 = \sum_i \left\| Y_i (I - \frac{1}{n} 1 1^T) - L_i^T \Theta_i \right\|^2 . \tag{84}$$

One can estimate $L_i^T = Y_i (1 - \frac{1}{n} 1 1^T) \Theta_i^\dagger$. Hence it reduces to

$$\min_{Y} \sum_i \|E_i\|^2 = \sum_i \left\| Y_i (I - \frac{1}{n} 1 1^T)(I - \Theta_i^\dagger \Theta_i) \right\|^2 \tag{85}$$

where $I - \Theta_i^\dagger \Theta_i$ is the projection to the normal space at x_i. This is equivalent to

LTSA Kernel

$$G_i = [1/\sqrt{k_i}, \tilde{V}_1^{(i)}, ..., \tilde{V}_d^{(i)}]^{k_i \times (d+1)},$$

$$W_i^{k_i \times k_i} = I - G_i G_i^T,$$

$$K^{n \times n} = \Phi = \sum_{i=1}^{n} S_i W_i W_i^T S_i^T$$

where the selection matrix $S_i^{n \times k} : [x_{i_1}, ..., x_{i_k}] = [x_1, ..., x_n] S_i^{n \times k}$.

1) Constant eigenvector is of 0-eigenvalue

2) So choose d+1 smallest eigenvectors for embedding

LTSA Algorithm (Zha-Zhang'02)

Algorithm 6: LTSA Algorithm

Input: A weighted undirected graph $G = (V, E)$ such that

1 $V = \{x_i \in \mathbb{R}^p : i = 1, \ldots, n\}$

2 $E = \{(i, j) :$ if j is a neighbor of i, i.e. $j \in \mathcal{N}_i\}$, e.g. k-nearest neighbors

Output: Euclidean d-dimensional coordinates $Y = [y_i] \in \mathbb{R}^{k \times n}$ of data.

3 ***Step 1*** (local PCA): Compute local SVD on neighborhood of x_i, $x_{i_j} \in \mathcal{N}(x_i)$,

$$\tilde{X}^{(i)} = [x_{i_1} - \mu_i, ..., x_{i_k} - \mu_i]^{p \times k} = \tilde{U}^{(i)} \tilde{\Sigma} (\tilde{V}^{(i)})^T,$$

where $\mu_i = \sum_{j=1}^{k} x_{i_j}$. Define

$$G_i = [1/\sqrt{k}, \tilde{V_1}^{(i)}, ..., \tilde{V_d}^{(i)}]^{k \times (d+1)};$$

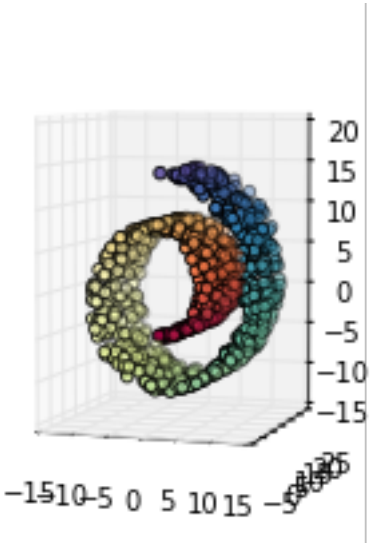
4 ***Step 2*** (tangent space alignment): Alignment (kernel) matrix

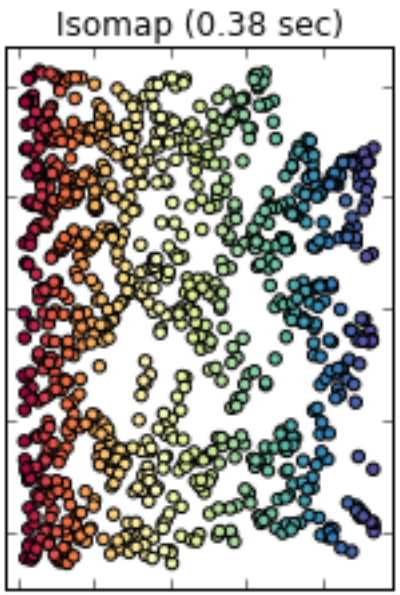
$$K^{n \times n} = \sum_{i=1}^{n} S_i W_i W_i^T S_i^T, \quad W_i^{k \times k} = I - G_i G_i^T,$$

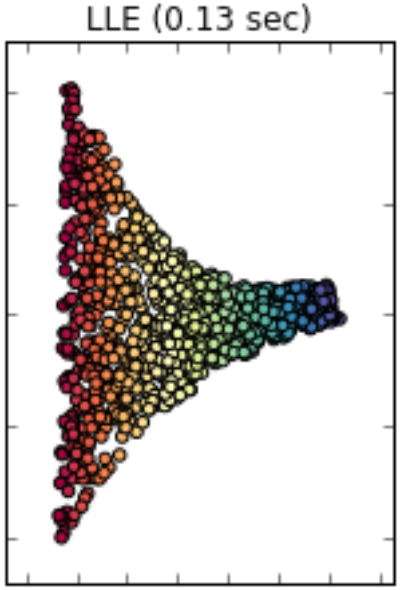
where selection matrix $S_i^{n \times k} : [x_{i_1}, ..., x_{i_k}] = [x_1, ..., x_n] S_i^{n \times k}$;

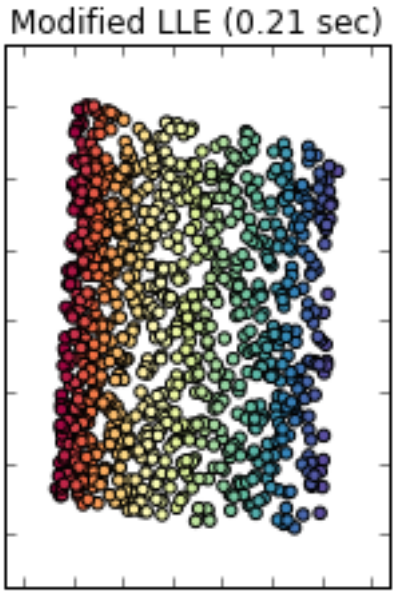
5 ***Step 3***: Find smallest $d + 1$ eigenvectors of K and drop the smallest eigenvector, the remaining d eigenvectors will give rise to a d-embedding.

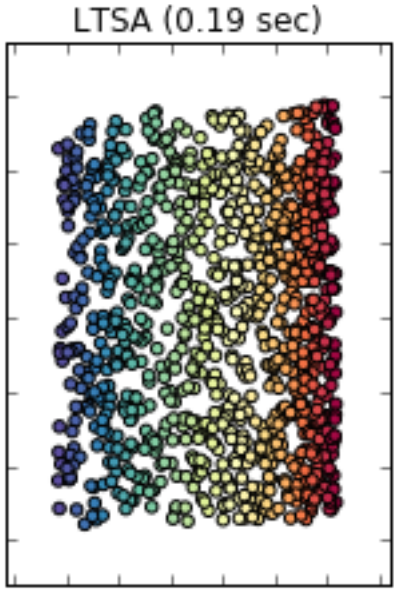
Comparisons on Swiss Roll











https://nbviewer.jupyter.org/url/math.stanford.edu/~yuany/course/data/plot_compare_methods.ipynb

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 O(N^3, but L-ISOMAP)	Local approach O(N^2)
Might not work for nonconvex manifolds with holes	Nonconvex manifolds with holes
Extensions: Landmark, Conformal & Isometric ISOMAP	Extensions: MLLE, LTSA, 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
- Zhenyue Zhang, Jing Wang. MLLE: Modified Locally Linear Embedding Using Multiple Weights, NIPS 2016.
- Zhenyue Zhang and Hongyuan Zha, Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment, SIAM Journal of Scientific Computing, 2002

Acknowledgement

- Slides stolen from Ettinger, Vikas C. Raykar, Vin de Silva.