Introduction to Manifold Learning I: ISOMAP and LLE

姚 遠

2024

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

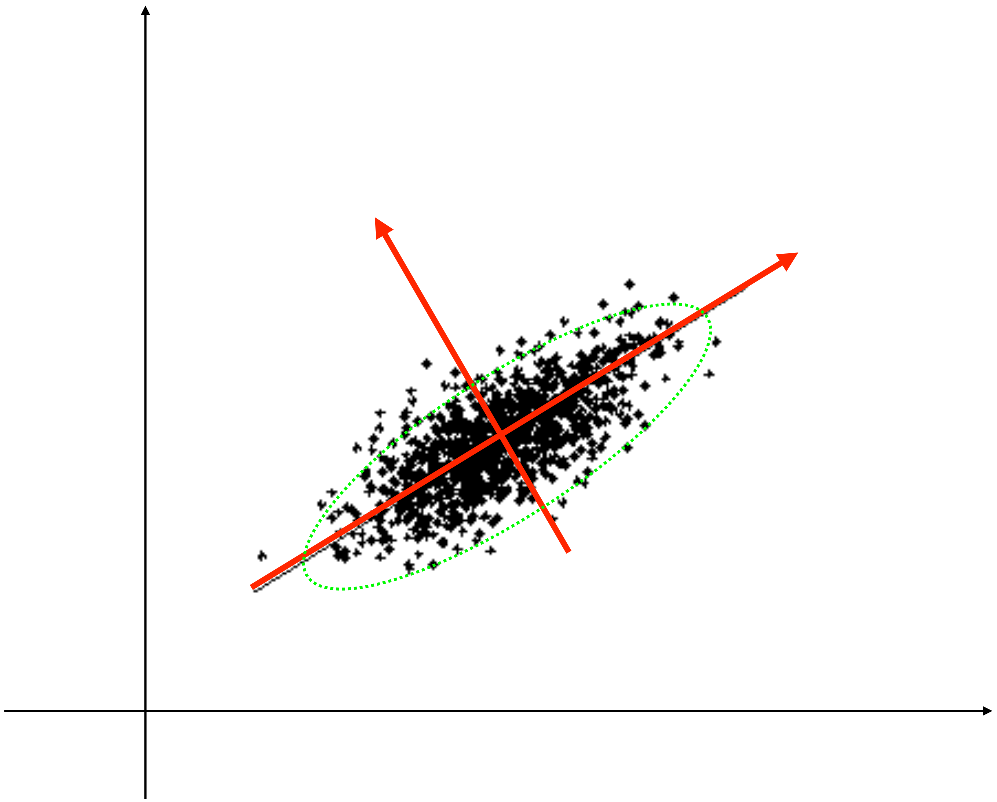
2

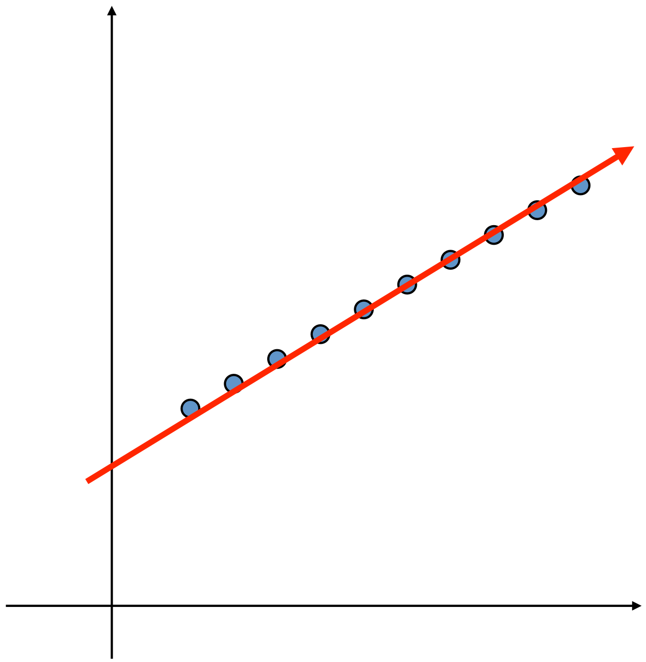
Recall: PCA

- Principal Component Analysis (PCA)

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

EigenValue Decomposition of XX^T

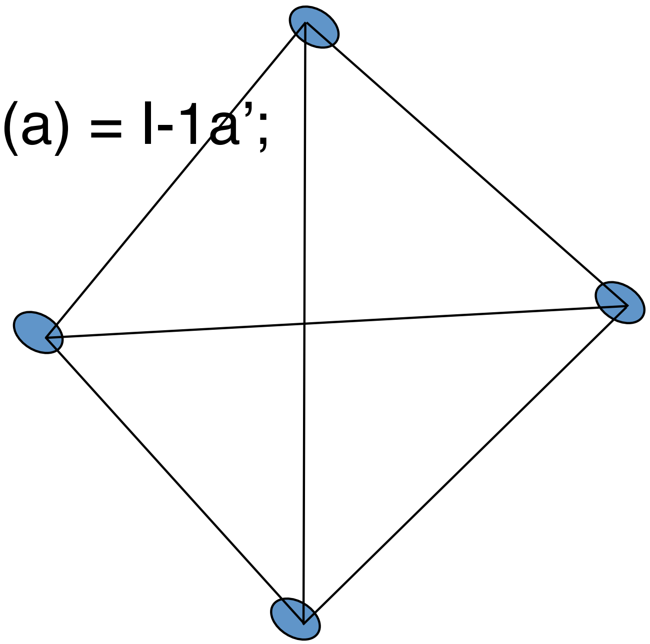




One Dimensional Manifold

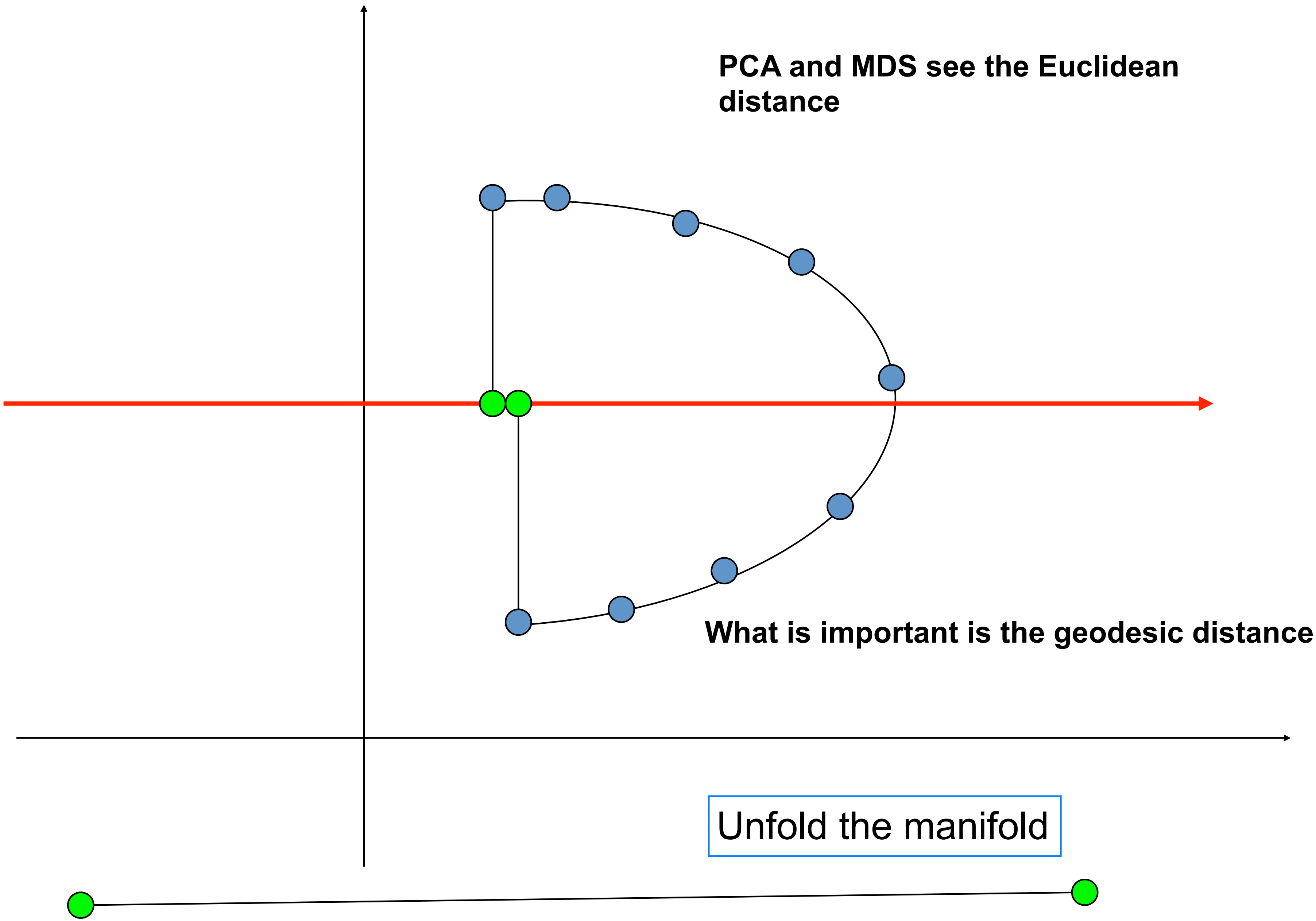
Recall: MDS

- 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-1a';
    - 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 $B = YY^T$



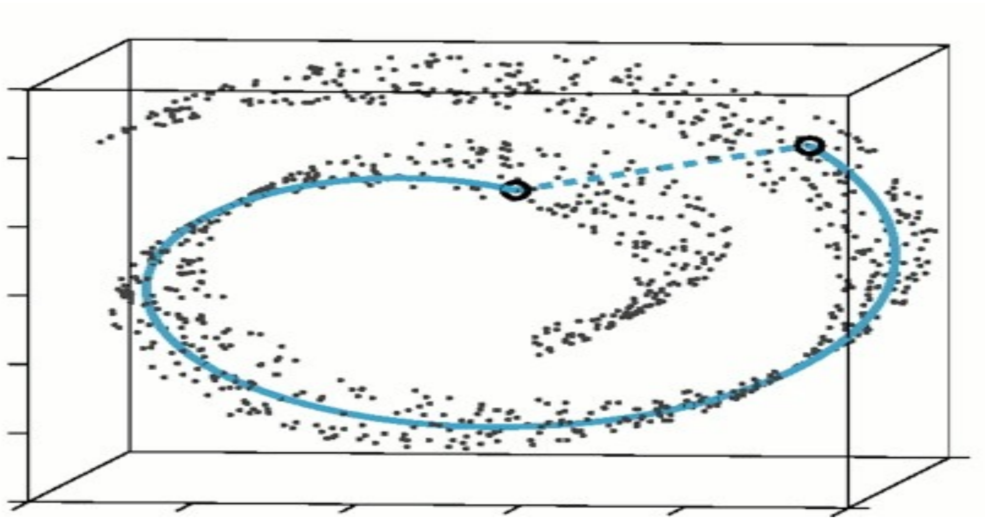
If we preserve the pairwise Euclidean distances do we preserve the structure??

Nonlinear Manifolds..



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

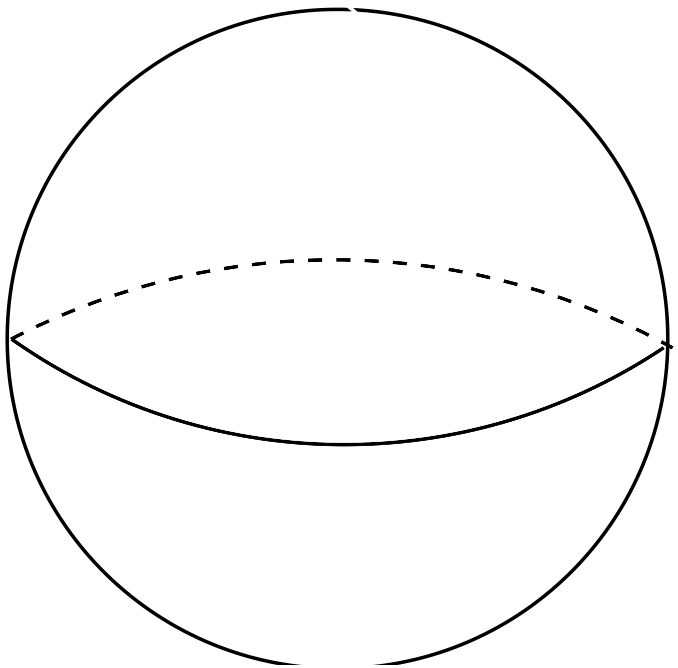
  curvature, geodesics

All you wanna know about differential geometry but were afraid to ask, in 9 easy slides

Embedded (sub-)Manifolds

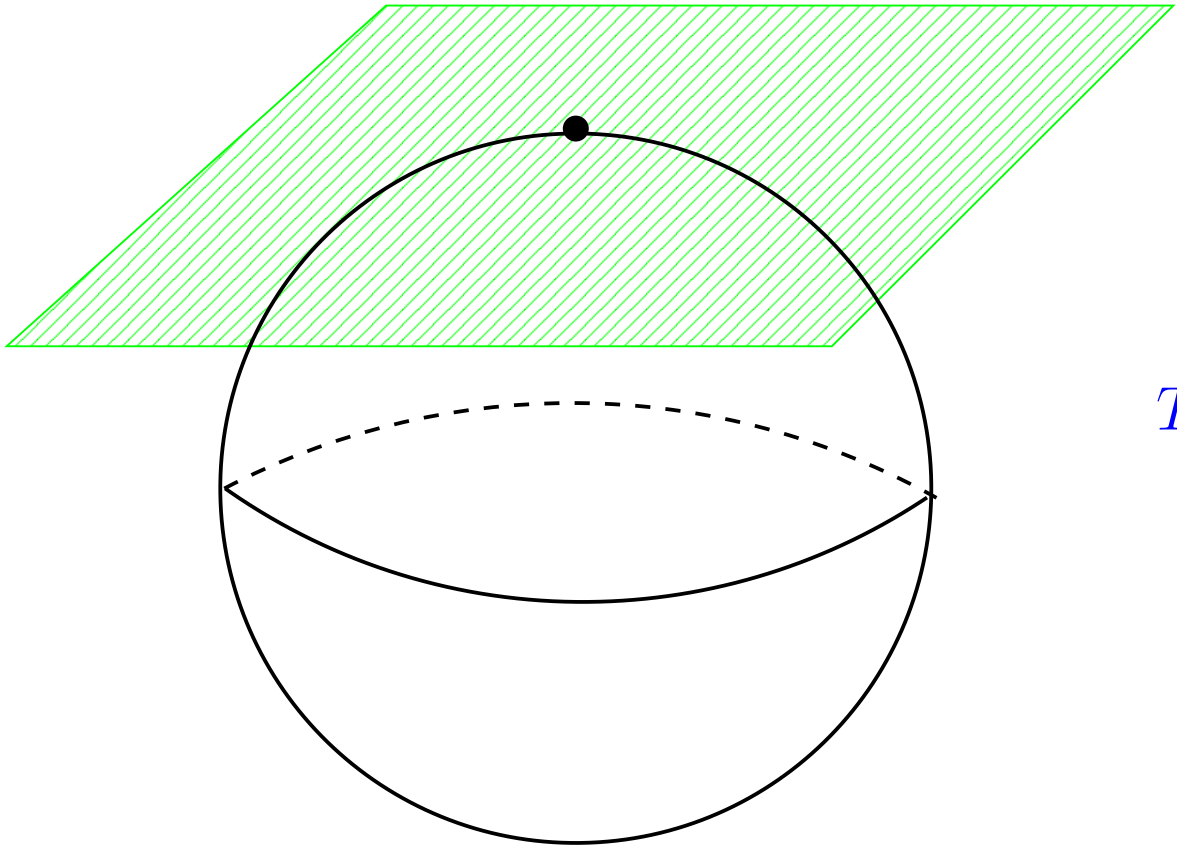
$$\mathcal{M}^k \subset \mathbb{R}^N$$

Locally (not globally) looks like Euclidean space.



$$S^2 \subset \mathbb{R}^3$$

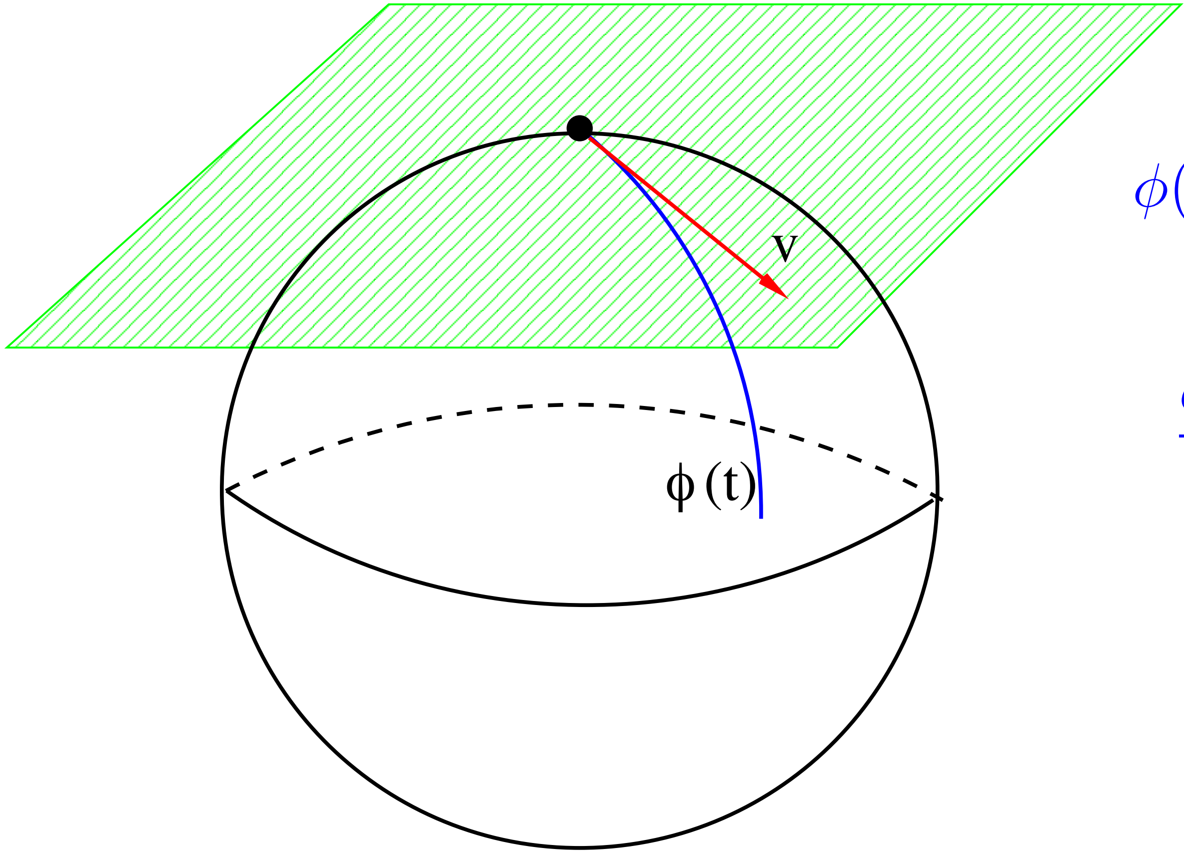
Tangent Space



$T_p\mathcal{M}^k \subset \mathbb{R}^N$

k-dimensional affine subspace of $\mathbb{R}^N$.

Tangent Vectors and Curves



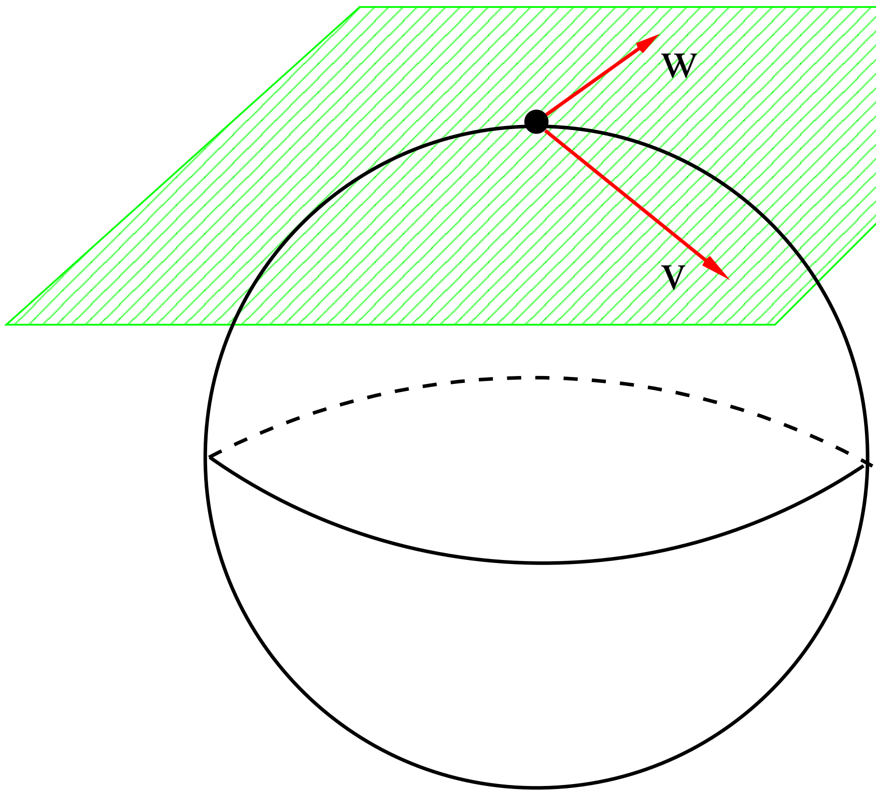
$$\phi(t) : \mathbb{R} \rightarrow \mathcal{M}^k$$

$$\left. \frac{d\phi(t)}{dt} \right|_0 = V$$

Tangent vectors <———> curves.

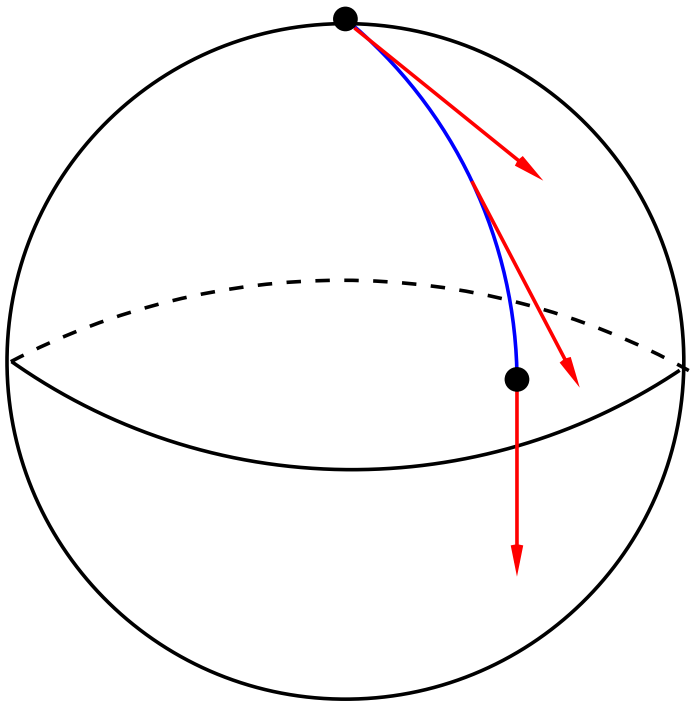
Riemannian Geometry

Norms and angles in tangent space.



$$\langle v, w \rangle \qquad \|v\|,\ \|w\|$$

Geodesics



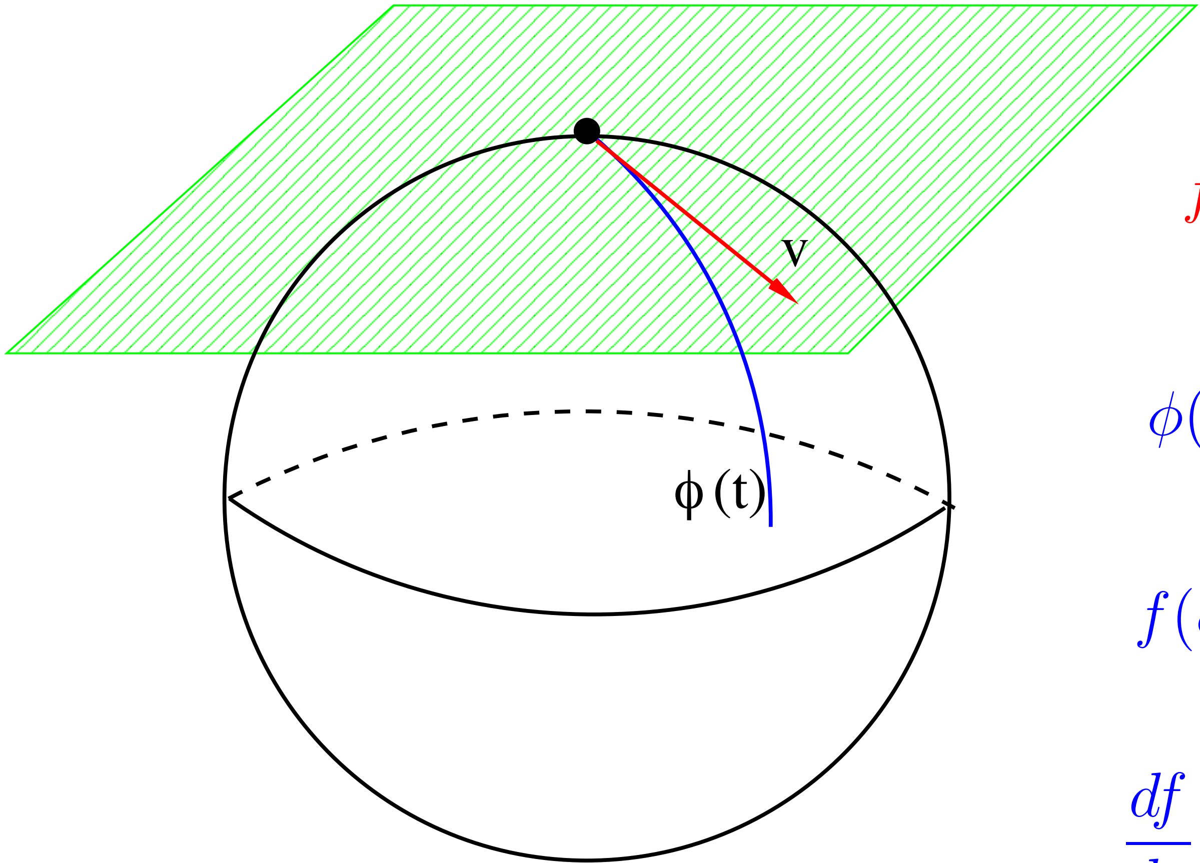
$$\phi(t) : [0,1] \to \mathcal{M}^k$$

$$l(\phi) = \int_0^1 \left\| \frac{d\phi}{dt} \right\| dt$$

Can measure length using norm in tangent space.

Geodesic — shortest curve between two points.

Tangent Vectors vs. Derivatives



$$f : \mathcal{M}^k \to \mathbb{R}$$

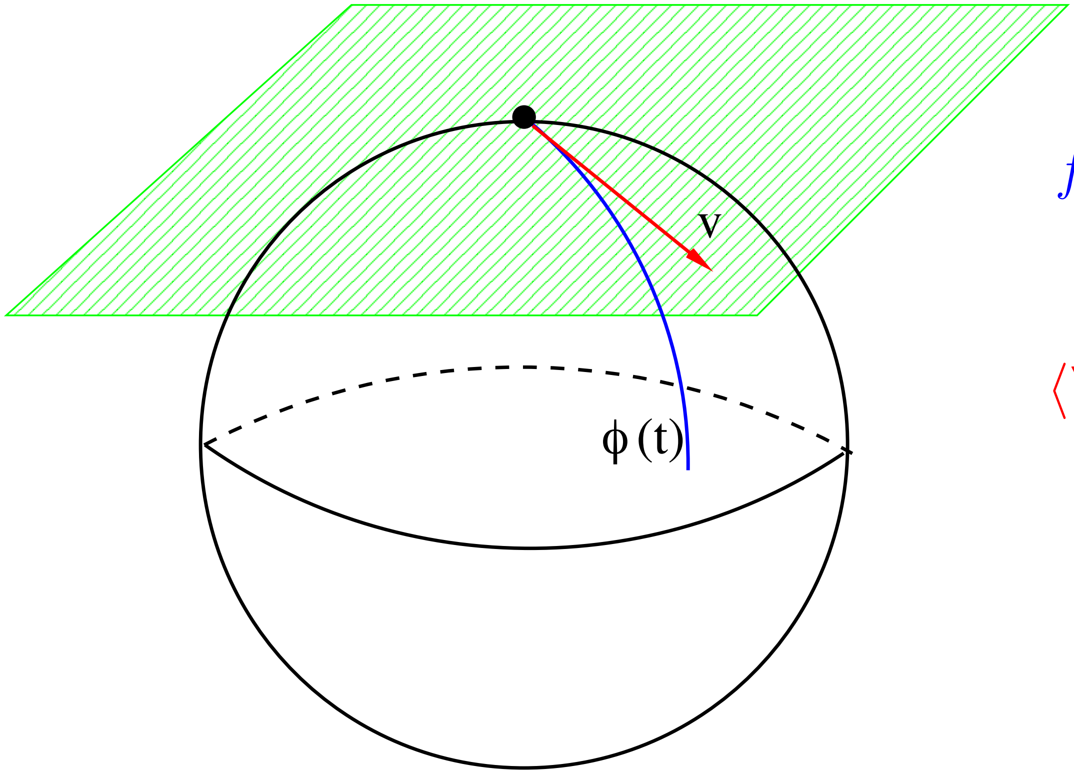
$$\phi(t) : \mathbb{R} \to \mathcal{M}^k$$

$$f(\phi(t)) : \mathbb{R} \to \mathbb{R}$$

$$\frac{df}{dv} = \left.\frac{d\,f(\phi(t))}{d\,t}\right|_0$$

Tangent vectors <———> Directional derivatives.

Gradients



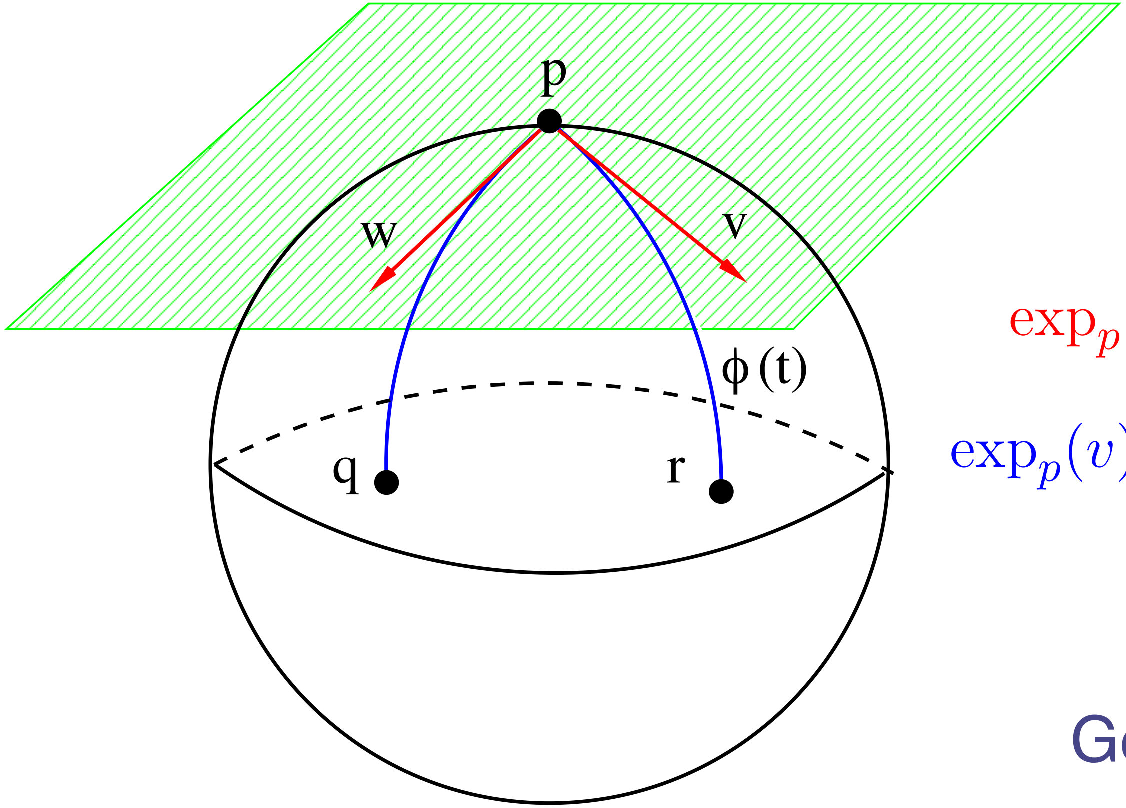
$f : \mathcal{M}^k \rightarrow \mathbb{R}$

$\langle \nabla f, v \rangle \equiv \dfrac{df}{dv}$

Tangent vectors <−−−> Directional derivatives.

Gradient points in the direction of maximum change.

Exponential Maps



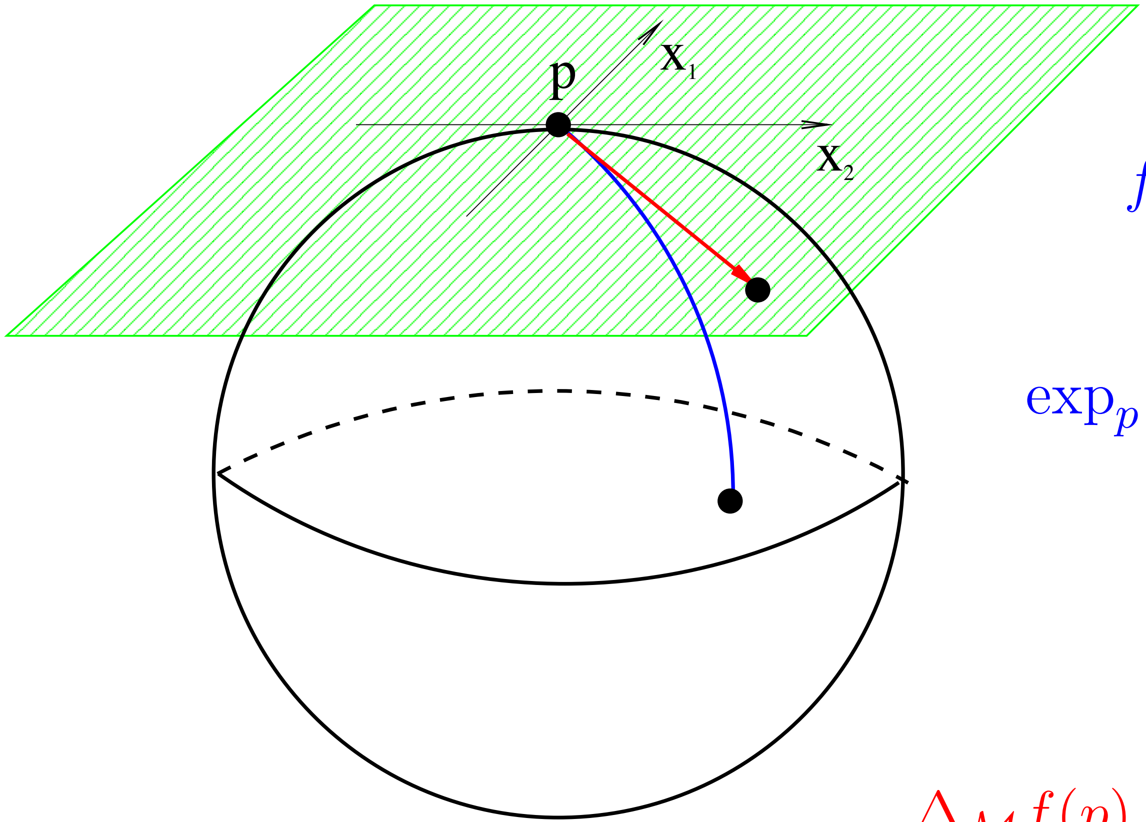
$$\exp_p : T_p\mathcal{M}^k \rightarrow \mathcal{M}^k$$

$$\exp_p(v) = r \quad \exp_p(w) = q$$

Geodesic $\phi(t)$

$$\phi(0) = p, \quad \phi(\|v\|) = q \qquad \left.\frac{d\phi(t)}{dt}\right|_0 = v$$

Laplacian-Beltrami Operator



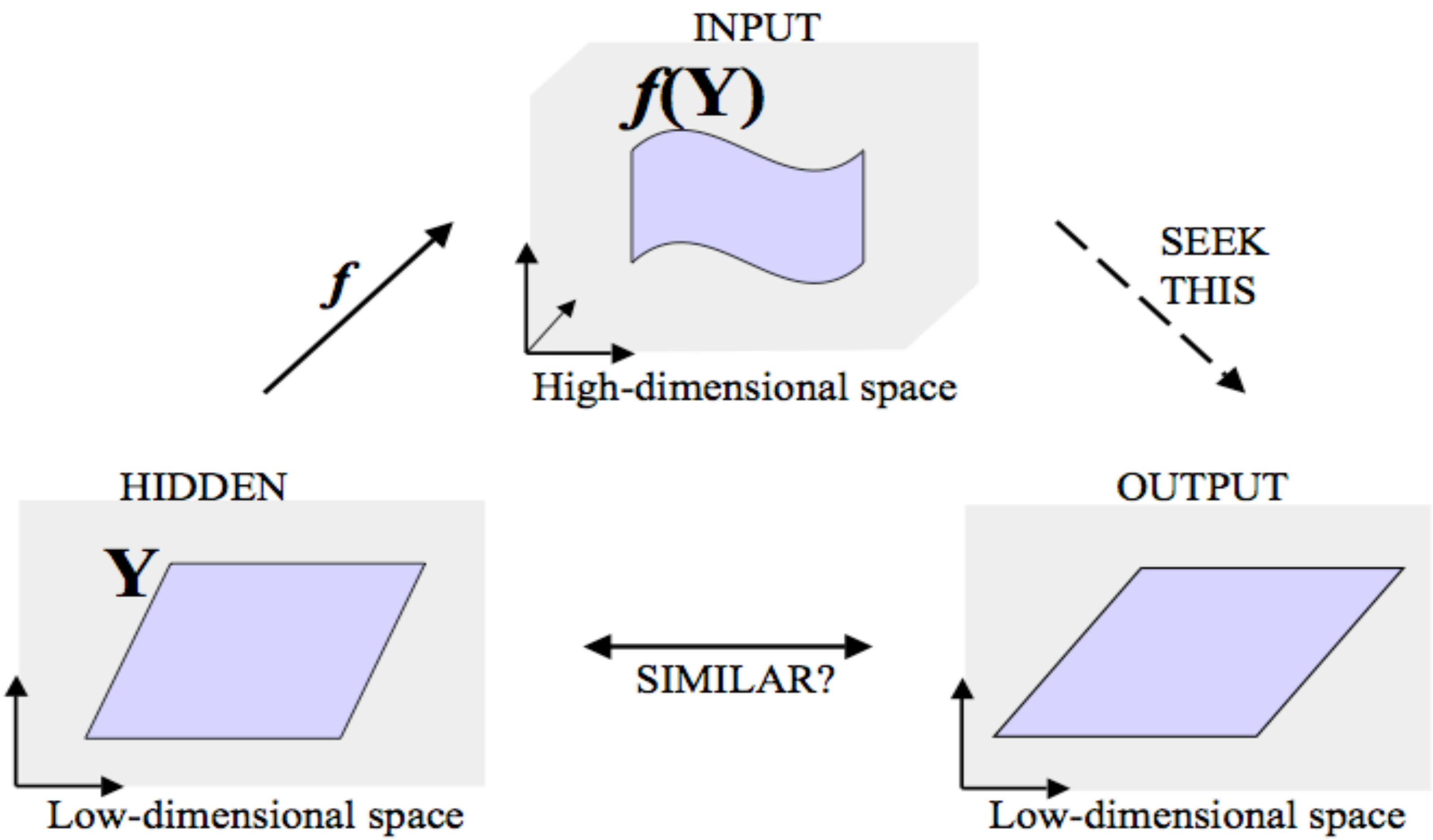
$$f : \mathcal{M}^k \to \mathbb{R}$$

$$\exp_p : T_p\mathcal{M}^k \to \mathcal{M}^k$$

$$\Delta_{\mathcal{M}} f(p) \equiv \sum_i \frac{\partial^2 f(\exp_p(x))}{\partial x_i^2}$$

Orthonormal coordinate system.

Generative Models in Manifold Learning



Spectral Geometric Embedding

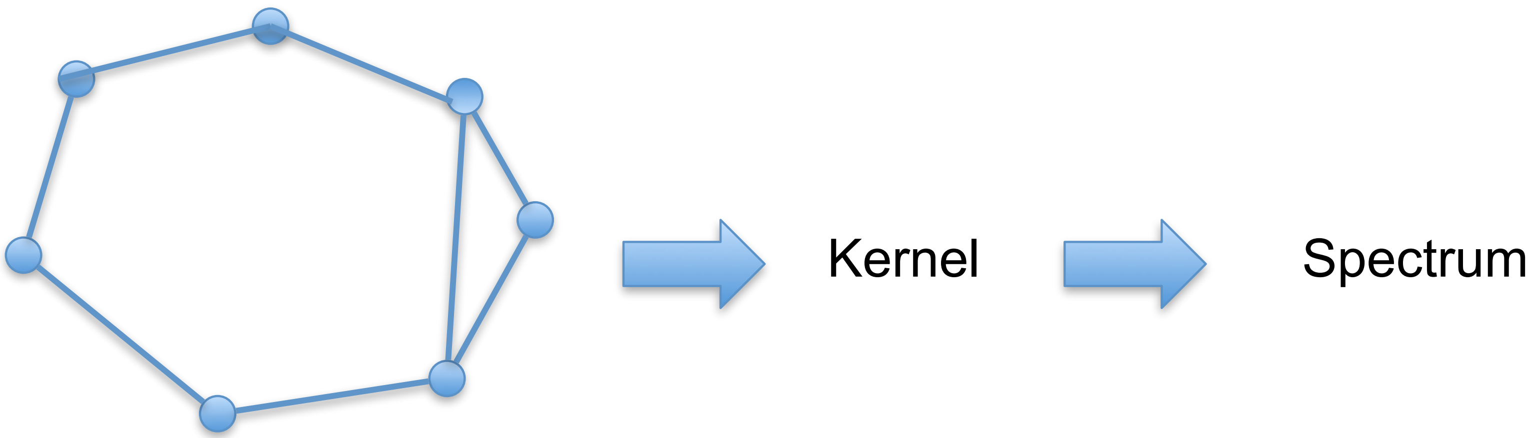
Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$,
Find $y_1, \ldots, y_n \in \mathbb{R}^d$ where $d << N$

- 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

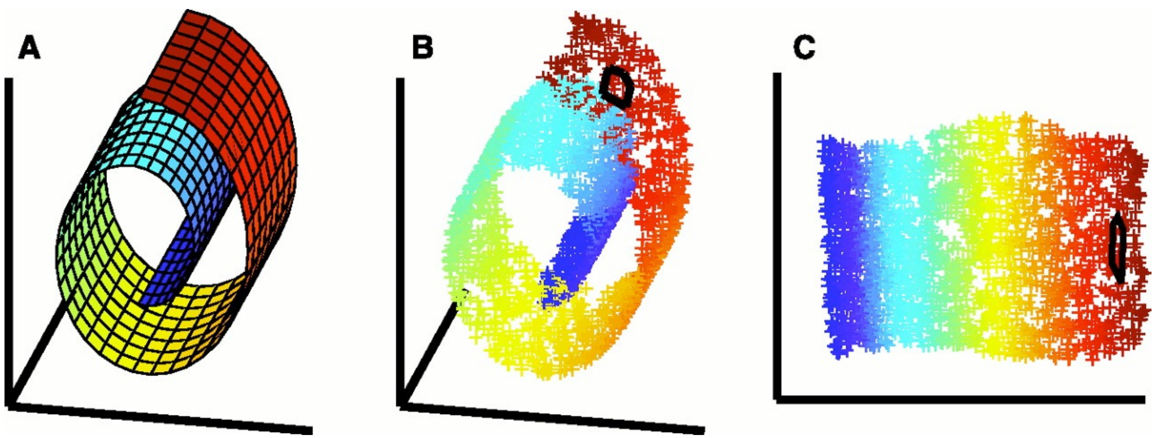
- Tenenbaum-de Silva-Langford Isomap Algorithm
  - Global approach: on a low dimensional embedding
    - Nearby points should be nearby.
    - Faraway points should be faraway.

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

Isomap

- **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

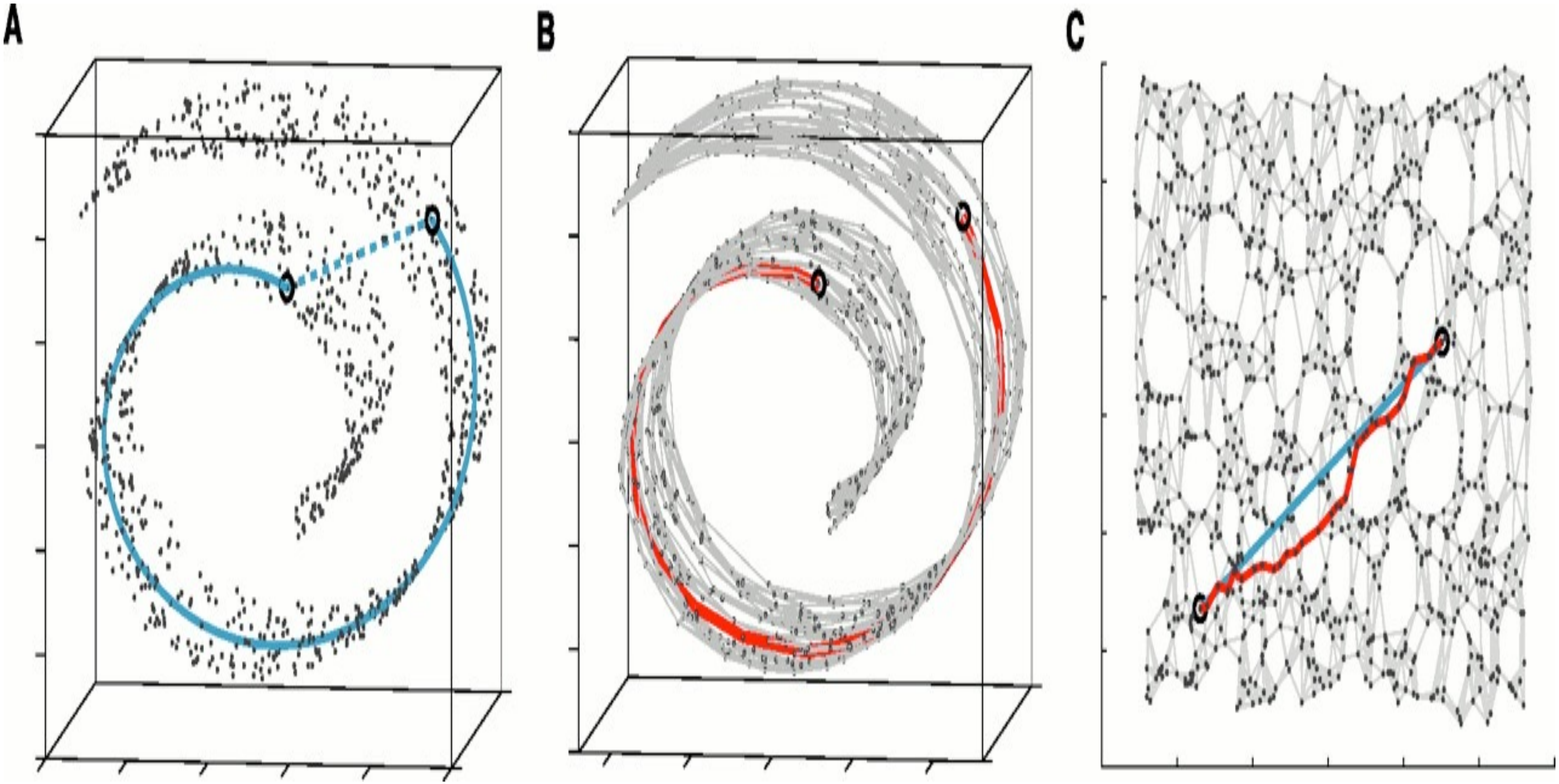
Once we have all pairwise geodesic distances use classical metric MDS

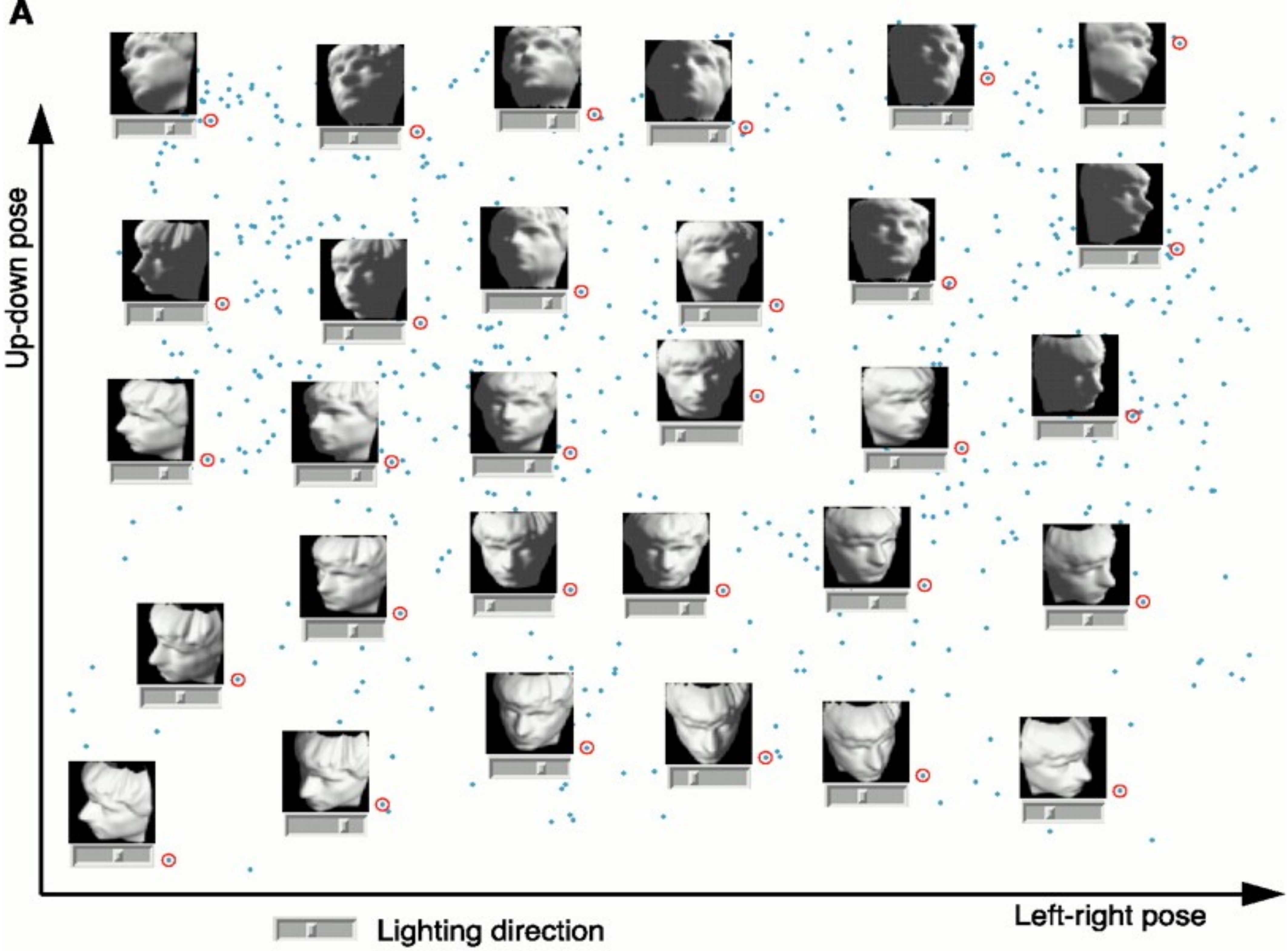


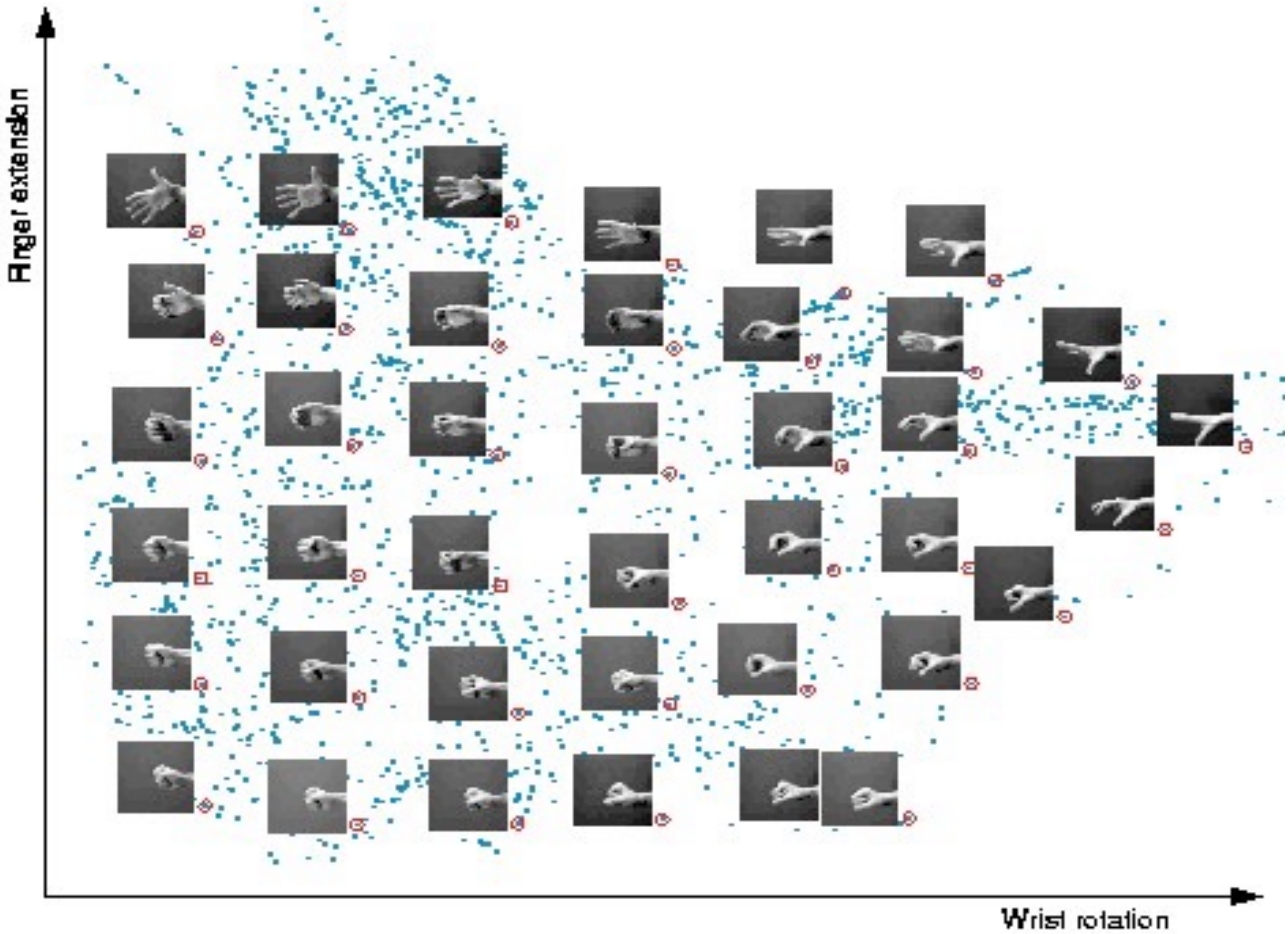
Isomap - Algorithm

- 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 ($O(kN^2 \log N)$)
- Construct a lower dimensional embedding.
  - Classical MDS ($K = -0.5\, H\, D\, H' = U\, S\, U'$)

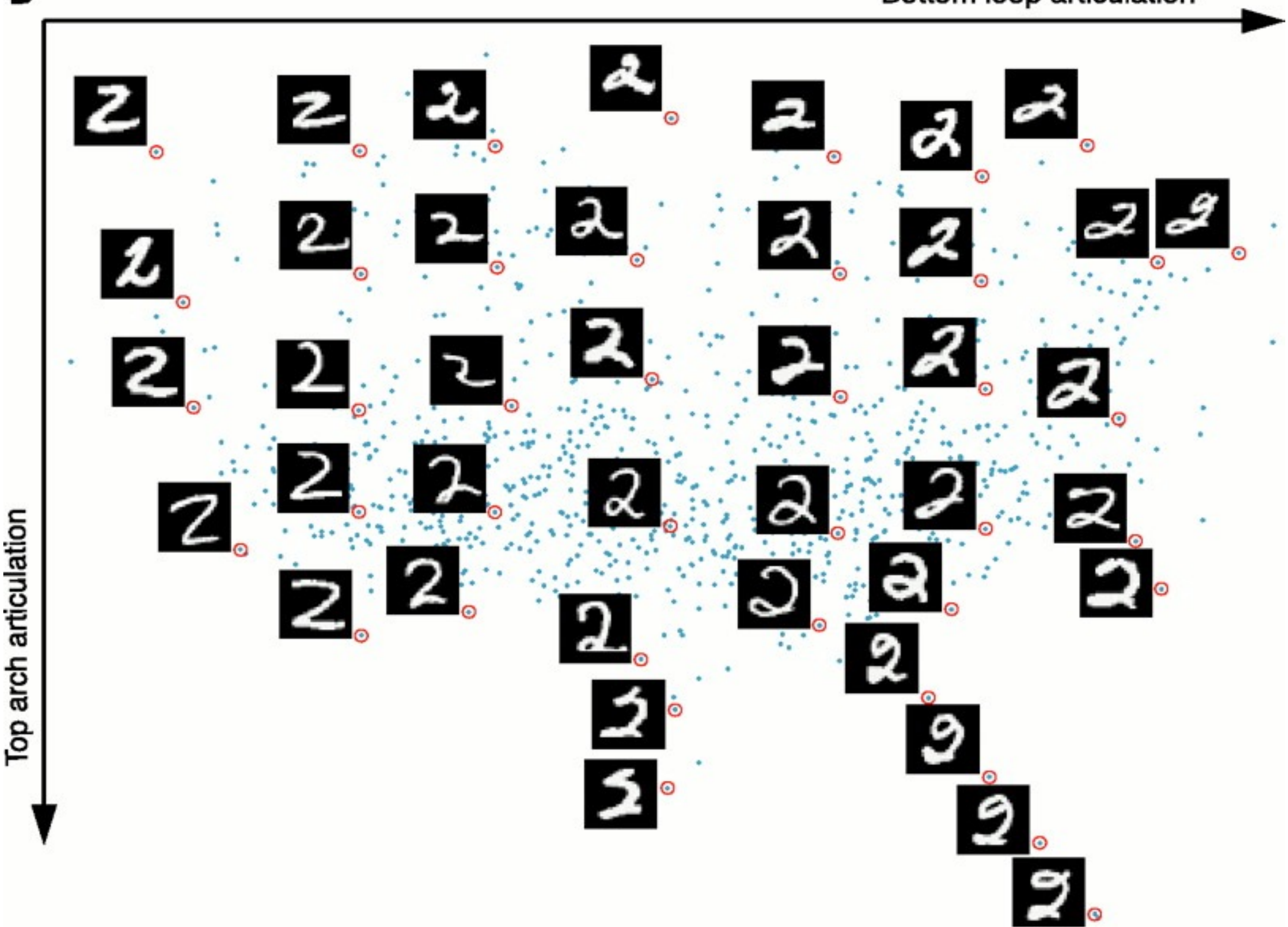
Isomap




A
Up-down pose
Left-right pose
Lighting direction


Finger extension
Wrist rotation

B



Residual Variance vs. Intrinsic Dimension

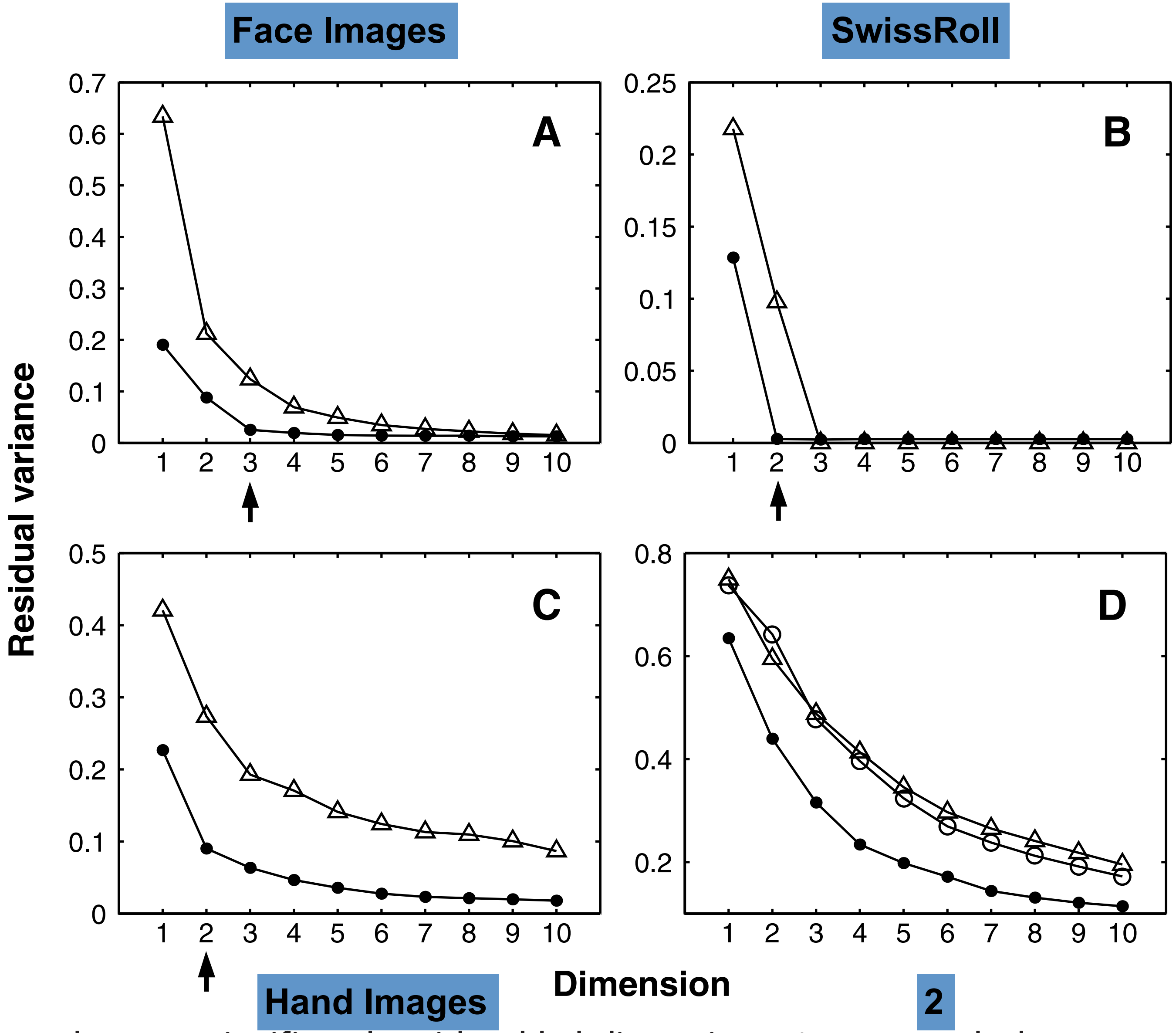
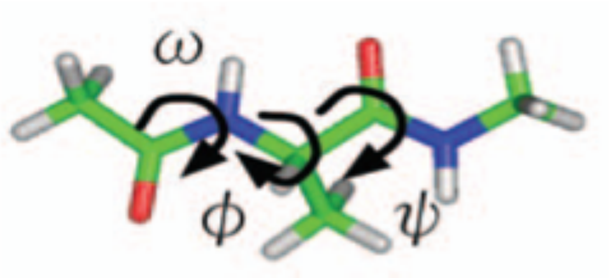
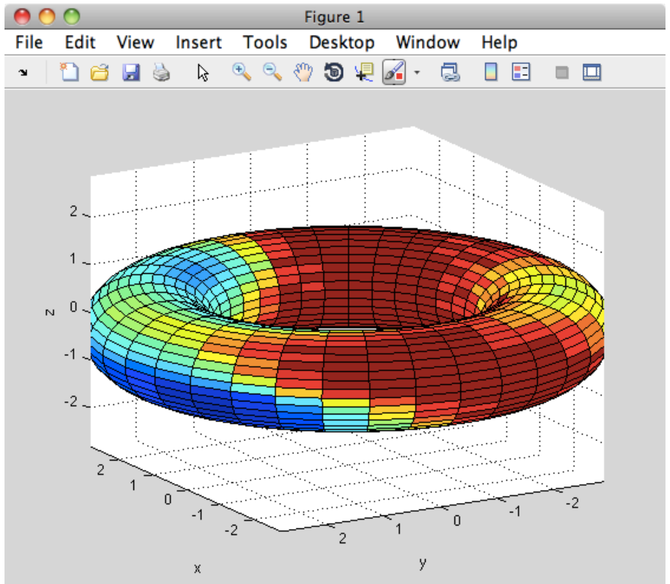
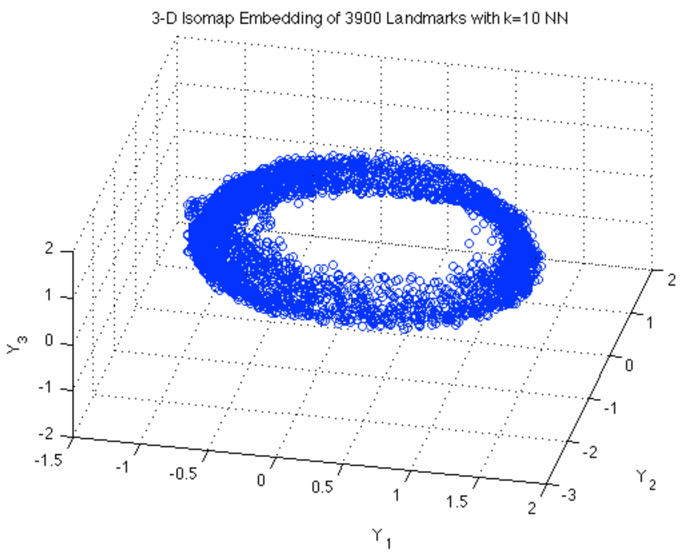


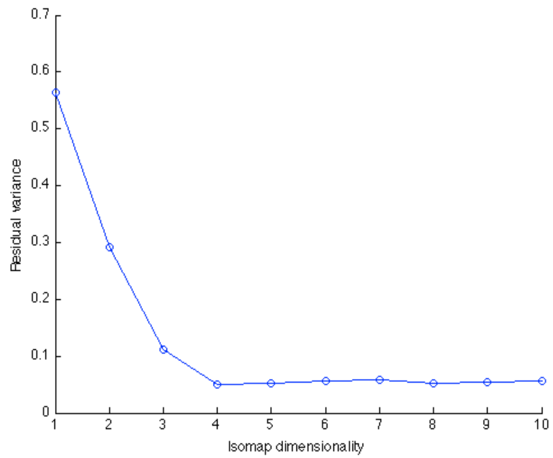
Fig. 2. The residual variance of PCA (open triangles), MDS [open triangles in (A) through (C); open circles in (D)], and Isomap (filled circles) on four data sets (*42*). (**A**) Face images varying in pose and illumination (Fig. 1A). (**B**) Swiss roll data (Fig. 3). (**C**) Hand images varying in finger extension and wrist rotation (*20*). (**D**) Handwritten "2"s (Fig. 1B). In all cases, residual variance decreases as the dimensionality *d* is increased. The intrinsic dimensionality of the data can be estimated by looking for the "elbow" at which this curve ceases to decrease significantly with added dimensions. Arrows mark the true or approximate dimensionality, when known. Note the tendency of PCA and MDS to overestimate the dimensionality, in contrast to Isomap.

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 Shortcomings 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

Nyström Method

- We are going to find top-k eigenvector decomposition of K:

- Let

$$\mathbf{K} = \left[\begin{array}{c|c} \mathbf{A} & \mathbf{B} \\ \hline \mathbf{B}^T & \mathbf{C} \end{array}\right] \succeq 0$$

$$\Rightarrow \mathbf{K} = \begin{bmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Y} \\ \mathbf{Y}^T\mathbf{X} & \mathbf{Y}^T\mathbf{Y} \end{bmatrix}$$

where

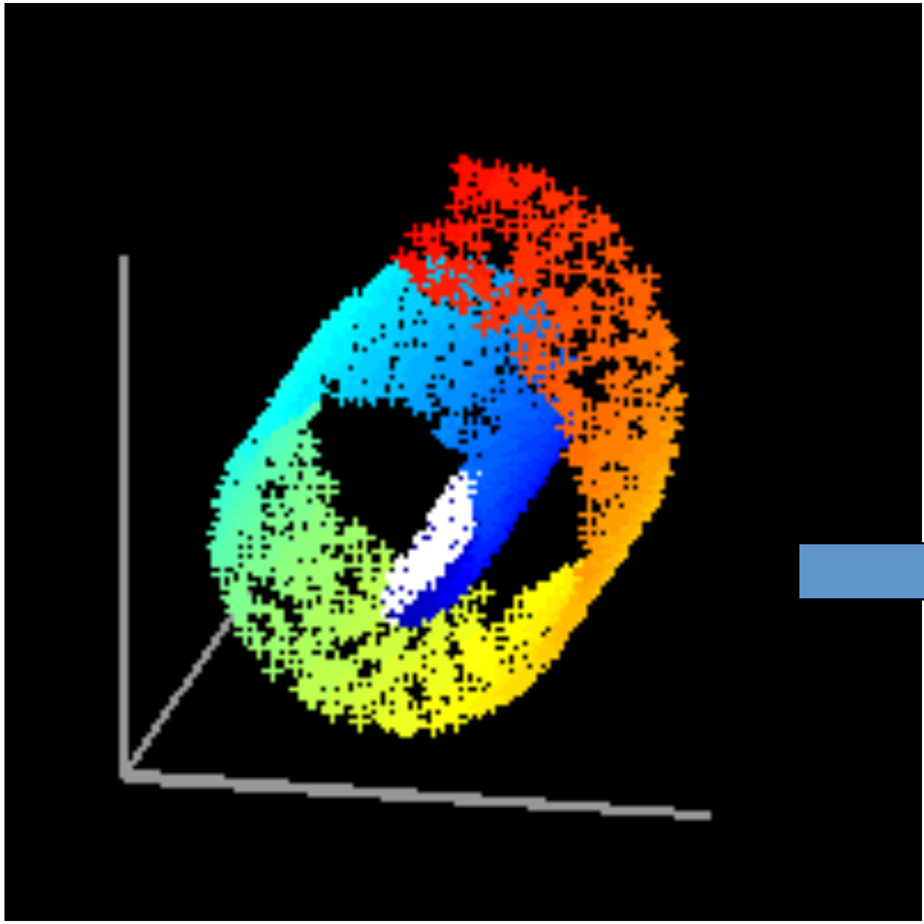
$$\mathbf{A} = \mathbf{X}^T\mathbf{X}$$
$$\mathbf{B} = \mathbf{X}^T\mathbf{Y}$$

Nyström Approximation

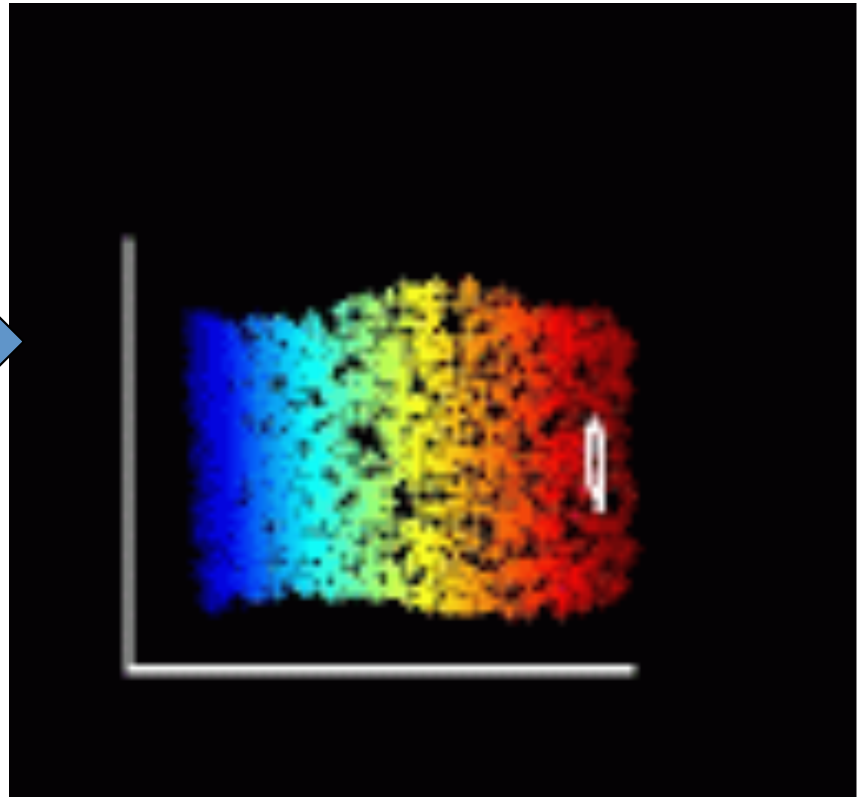
- Take
$$\mathbf{A} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{U}^T$$
and
$$\mathbf{X} = \boldsymbol{\Gamma}_{[k]}^{1/2}\mathbf{U}_{[k]}^T$$
where the subscript $[k]$ indicates the submatrices corresponding to the eigenvectors with the k largest positive eigenvalues. The coordinates corresponding to $\mathbf{B}$ can be derived as
$$\mathbf{Y} = \mathbf{X}^{-T}\mathbf{B} = \boldsymbol{\Gamma}_{[k]}^{-1/2}\mathbf{U}_{[k]}^T\mathbf{B}$$
- Nyström approximates $\mathbf{K}$ by
$$\tilde{\mathbf{K}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B} \end{bmatrix}$$
with approximation error $\|\mathbf{C} - \mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\|$.

Locally Linear Embedding

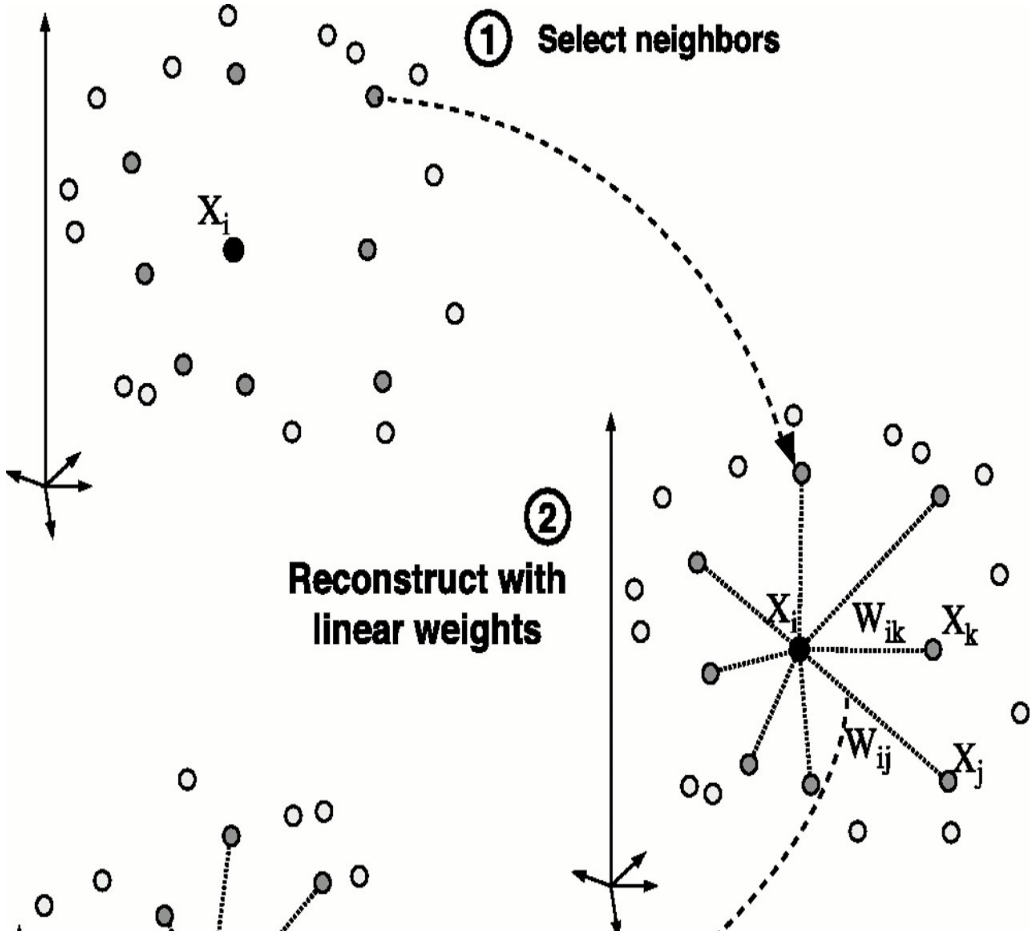
manifold is a topological space which is locally Euclidean."



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.

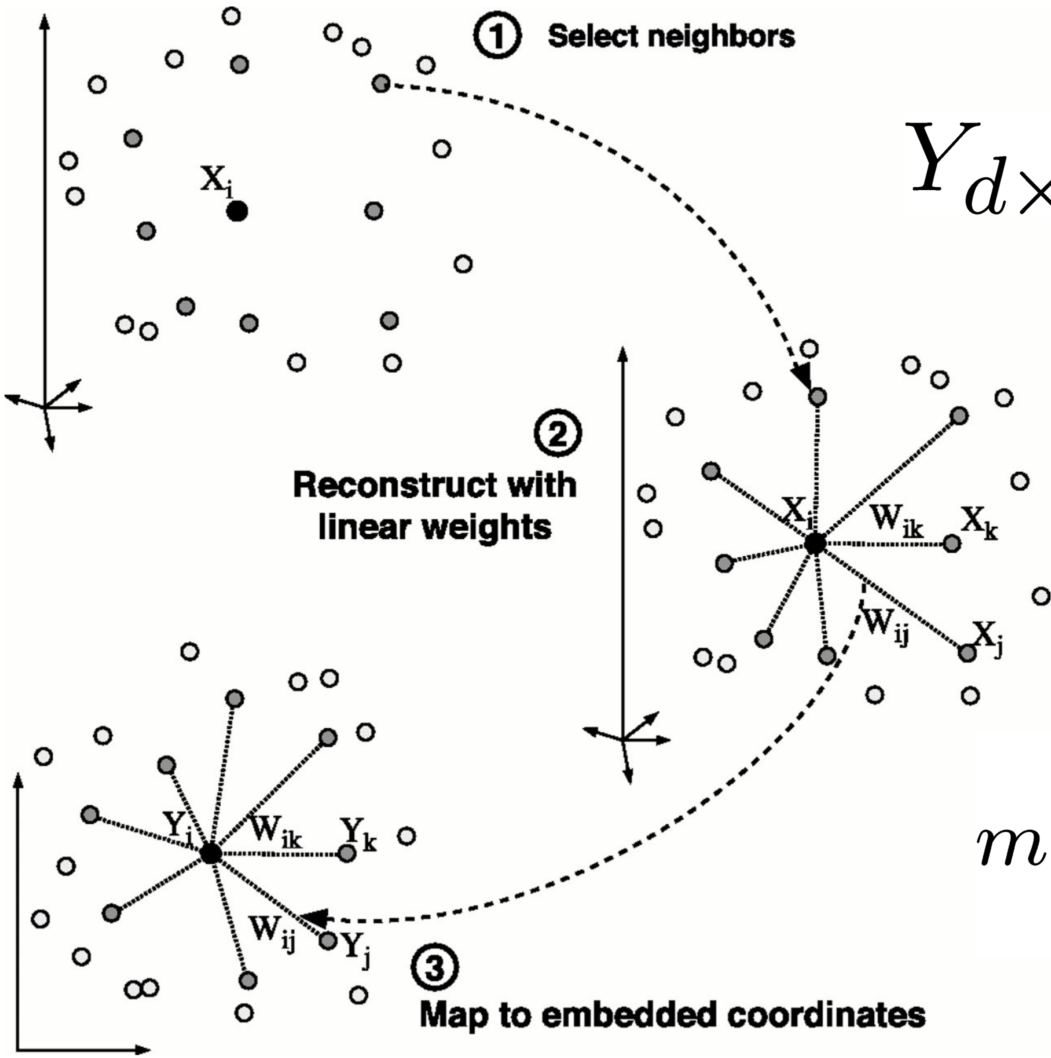
$$min_W \parallel X_i - \sum_{j=1}^{K} W_{ij} X_j \parallel^2 \qquad (1)$$

Derivation on board

Important property…

- 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…



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

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

LLE Algorithm: Local Fit

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

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

which is equivalent to

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

that is, finding a linear combination (possibly *not unique*!) for the subspace spanned by $\{(x_j - x_i) : j \in \mathcal{N}_i\}$.

LLE Algorithm: Local Fit (II)

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

$$w_i = \lambda C_i^{\dagger}\mathbf{1}, \tag{5}$$

where the Lagrange multiplier equals to the following normalization parameter

$$\lambda = \frac{1}{\mathbf{1}^T C_i^{\dagger}\mathbf{1}}, \tag{6}$$

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: Global Alignent

- 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,

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

  – (Kernel) Construct a positive semi-definite matrix $K = (I-W)^T(I-W)$ and find $d+1$ smallest eigenvectors of K, $v_0, v_1, \ldots, v_d$ 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

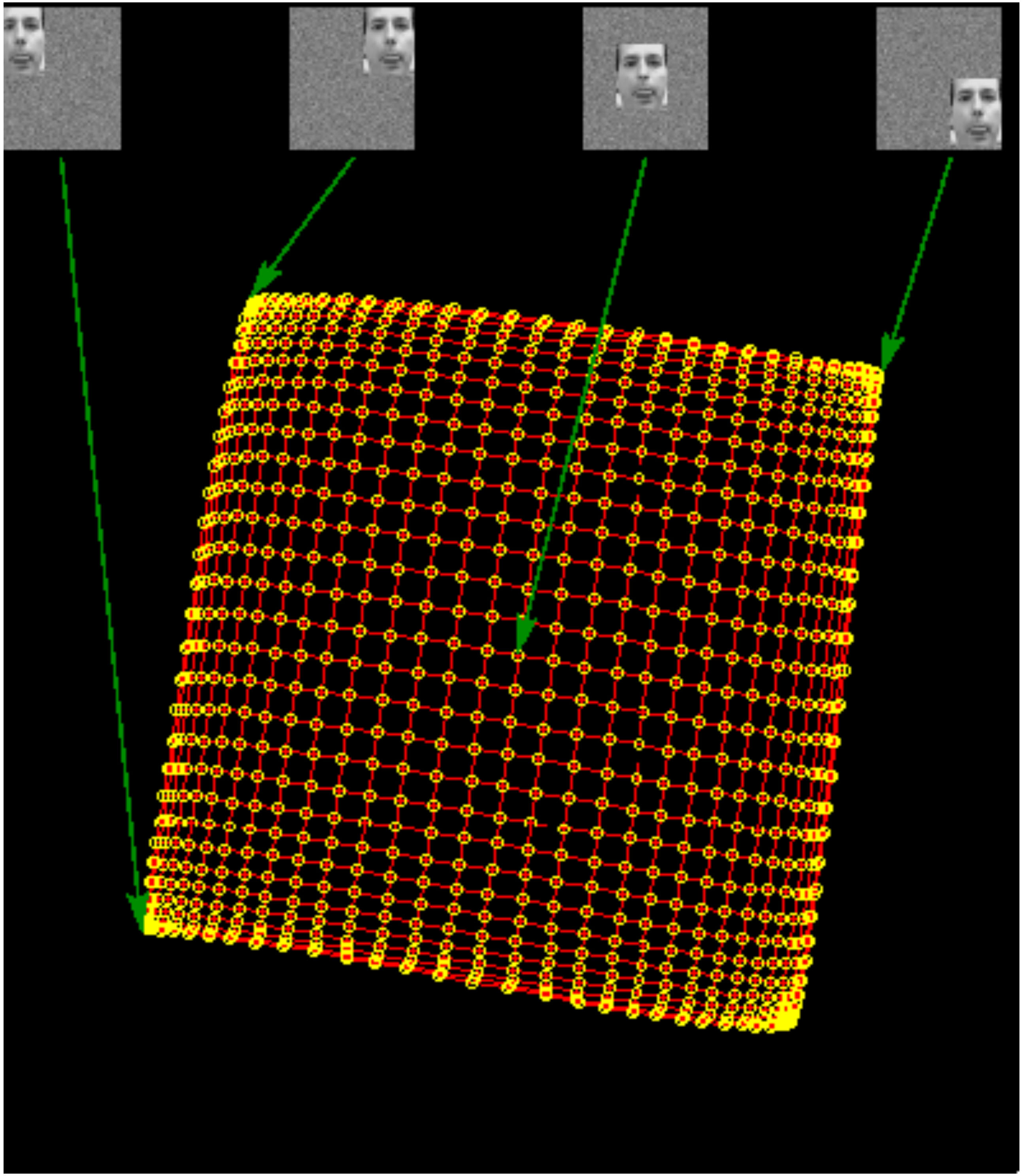
$$Y = [v_1/\sqrt{\lambda_1}, \ldots, v_d/\sqrt{\lambda_d}]^T.$$

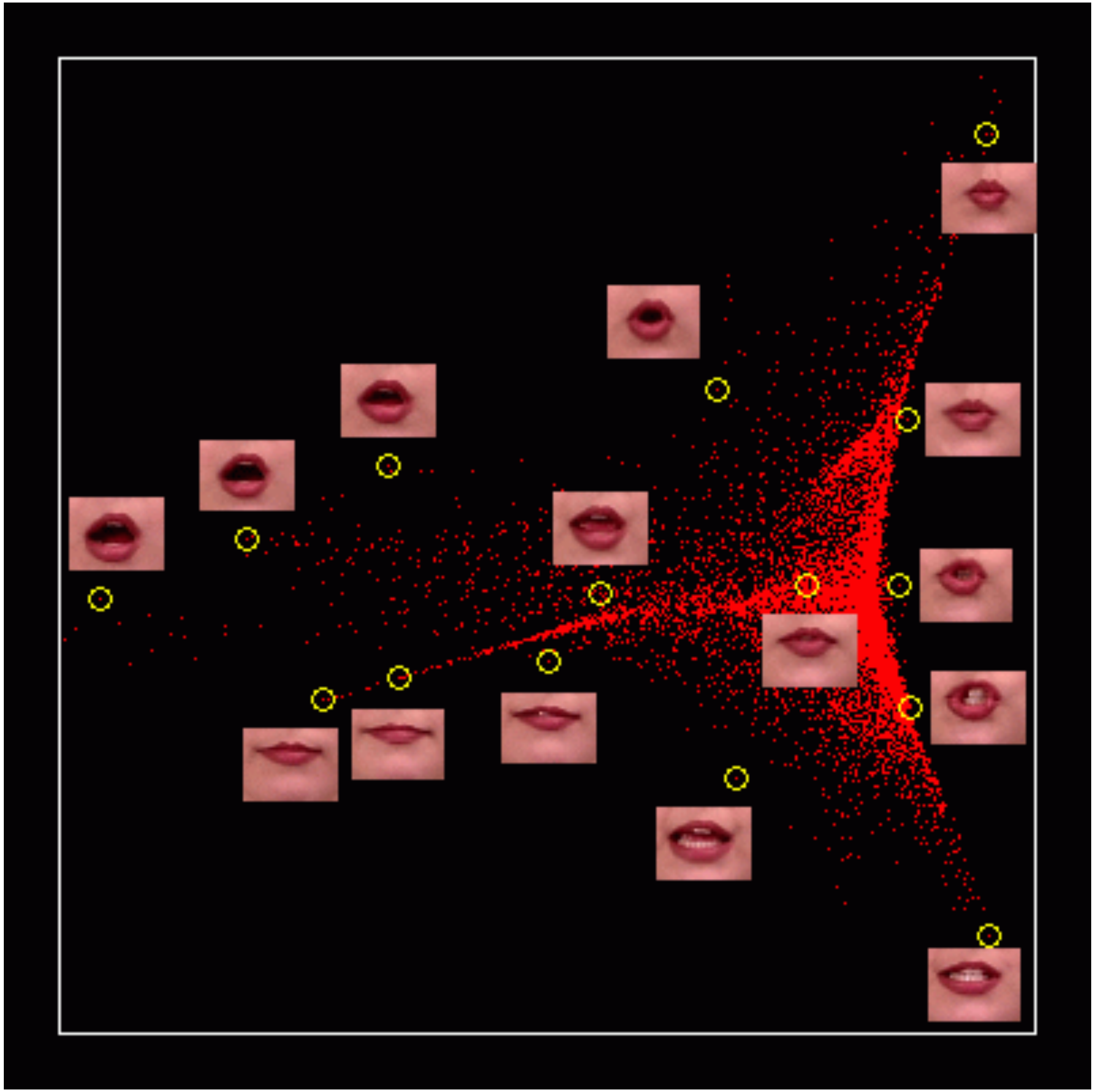
Remarks on LLE

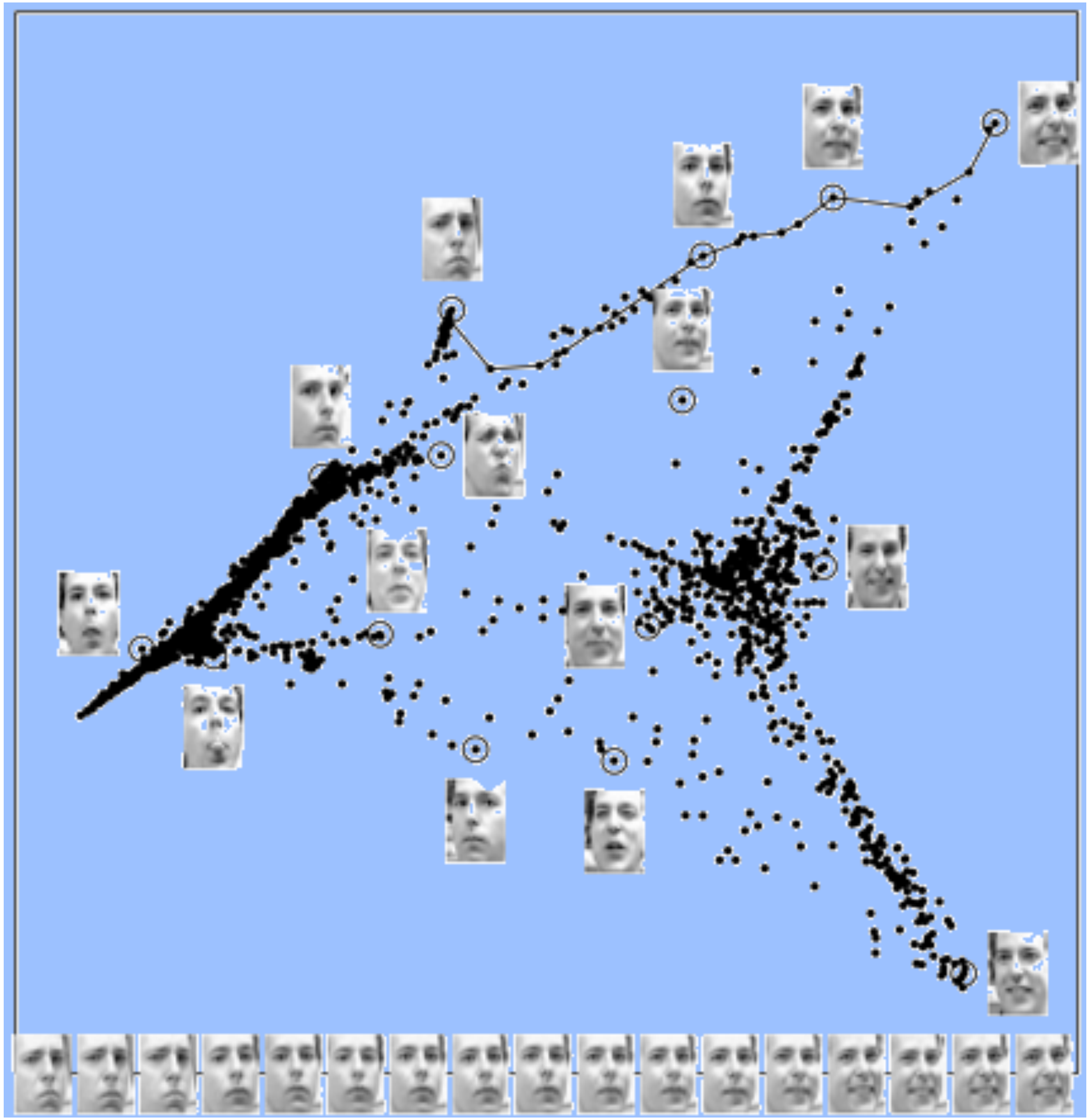
- Searching k-nearest neighbors is of O(kN)

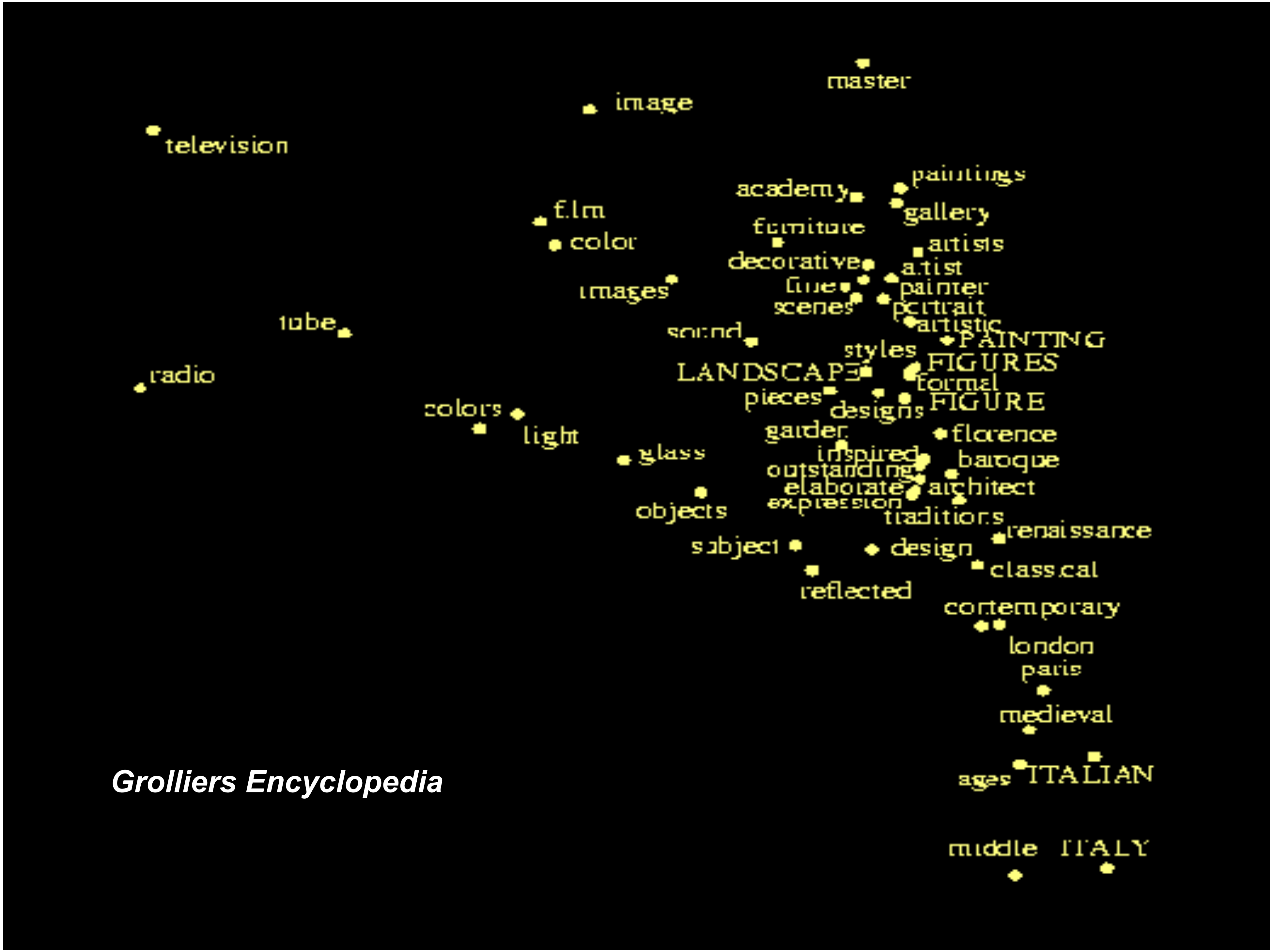
- W is sparse, kN/N^2=k/N nozeros
- W might be negative, additional nonnegative constraint can be imposed
- $B=(I-W)^T(I-W)$ is positive semi-definite (p.s.d.)

- Open Problem: exact reconstruction condition?








master
image
television
paintings
academy
gallery
film
color
furniture
artists
decorative
artist
images
fine
painter
scenes
portrait
tube
sound
artistic
PAINTING
styles
FIGURES
LANDSCAPE
formal
radio
pieces
FIGURE
designs
colors
light
garden
florence
glass
inspired
baroque
outstanding
elaborate
architect
expression
objects
traditions
renaissance
subject
design
classical
reflected
contemporary
london
paris
medieval
ages
ITALIAN
middle
ITALY
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

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

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)

- Modified Locally Linear Embedding (MLLE) remedies the issue using multiple weight vectors projected from orthogonal complement of local PCA.

- MLLE replace the weight vector w_i $(w^T \mathbf{1}_{k_i} = 1)$ 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

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

  where $\alpha_i = \|V_i^T \mathbf{1}_{k_i}\|_2 / \sqrt{s_i}$ and $H_i = I_{s_i} - 2uu^T$ $(\|u\|_2 = 1 \text{ or } 0)$ 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 is a legal weight vector.

MLLE (II)

- u: one can choose u in the direction of $V_i^T \mathbf{1}_{k_i} - \alpha_i \mathbf{1}_{s_i}$.

- s_i: an adaptive choice of s_i is based on the trade-off between residual variation and explained variation.
  - 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 dimension 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.

MLLE (III)

- Equipped with this weight matrix, one can set the objective function by simultaneously minimizing the residue over all reconstruction weights:

$$\begin{aligned}\min_Y \sum_i \sum_{l=1}^{s_i} \|y_i - \sum_{j\in\mathcal{N}_i} W_i(j,l) y_j\|^2 \quad &:= \quad \sum_i \|Y\widehat{W}_i\|_F^2 \\ &= \quad \mathrm{tr}[Y(\sum_i \widehat{W}_i \widehat{W}_i^T) Y^T]\end{aligned}$$

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

$$\widehat{W}_i(j,:) = \begin{cases} -\mathbf{1}_{s_i}^T, & j = i, \\ W_i, & j \in \mathcal{N}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

MLLE Algorithm

- **Step 1** (local fitting): for each x_i and its neighbors $\mathcal{N}_i$, solve

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

  by $\hat{w}_i(\mu) = (C_i + \mu I)^{-1}\mathbf{1}$ for some regularization parameter $\mu > 0$ and $w_i = \hat{w}_i/\hat{w}_i^T\mathbf{1}$. This is the same as LLE.

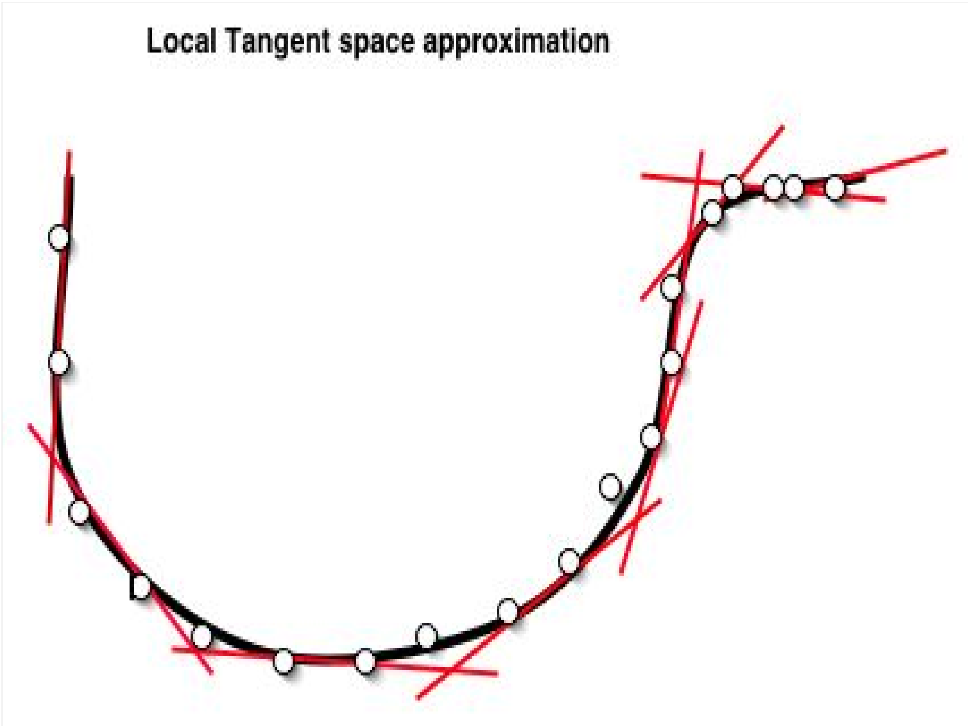
- **Step 2** (local residue PCA): get W_i as above.

- **Step 3** (global alignment): define the kernel matrix $K = \widehat{W}^T\widehat{W} = U\Lambda U^T$ with $\Lambda = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{n-1} > \lambda_n = 0$; choose bottom $d+1$ eigenvalues and drop the smallest one (0-constant), such that $U_d = [u_{n-d}, \ldots, u_{n-1}]$ and $\Lambda_d = \mathbf{diag}(\lambda_{n-d}, \ldots, \lambda_{n-1})$. Return the embedding $Y_d = U_d\Lambda_d^{\frac{1}{2}}$.

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 PCA

- For each x_i in $\mathbb{R}^p$ 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 $(\tilde{V}_1^{(i)}, \ldots, \tilde{V}_d^{(i)}) \in \mathbb{R}^{k_i \times d}$ 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 $\mathbb{R}^d$.
- Let $L_i : \mathbb{R}^{p \times d}$ 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

$$\min_{Y,L} \sum_i \|E_i\|^2 = \sum_i \left\| Y_i (I - \frac{1}{k_i} \mathbf{1}\mathbf{1}^T) - L_i^T \Theta_i \right\|^2 . \qquad (3)$$

LTSA

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

$$\min_Y \sum_i \|E_i\|^2 = \sum_i \left\| Y_i(I - \frac{1}{k_i}\mathbf{1}\mathbf{1}^T)(I - \Theta_i^{\dagger}\Theta_i) \right\|^2 \qquad (4)$$

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

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

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

$E = \{(i, j) :$ 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.

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

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

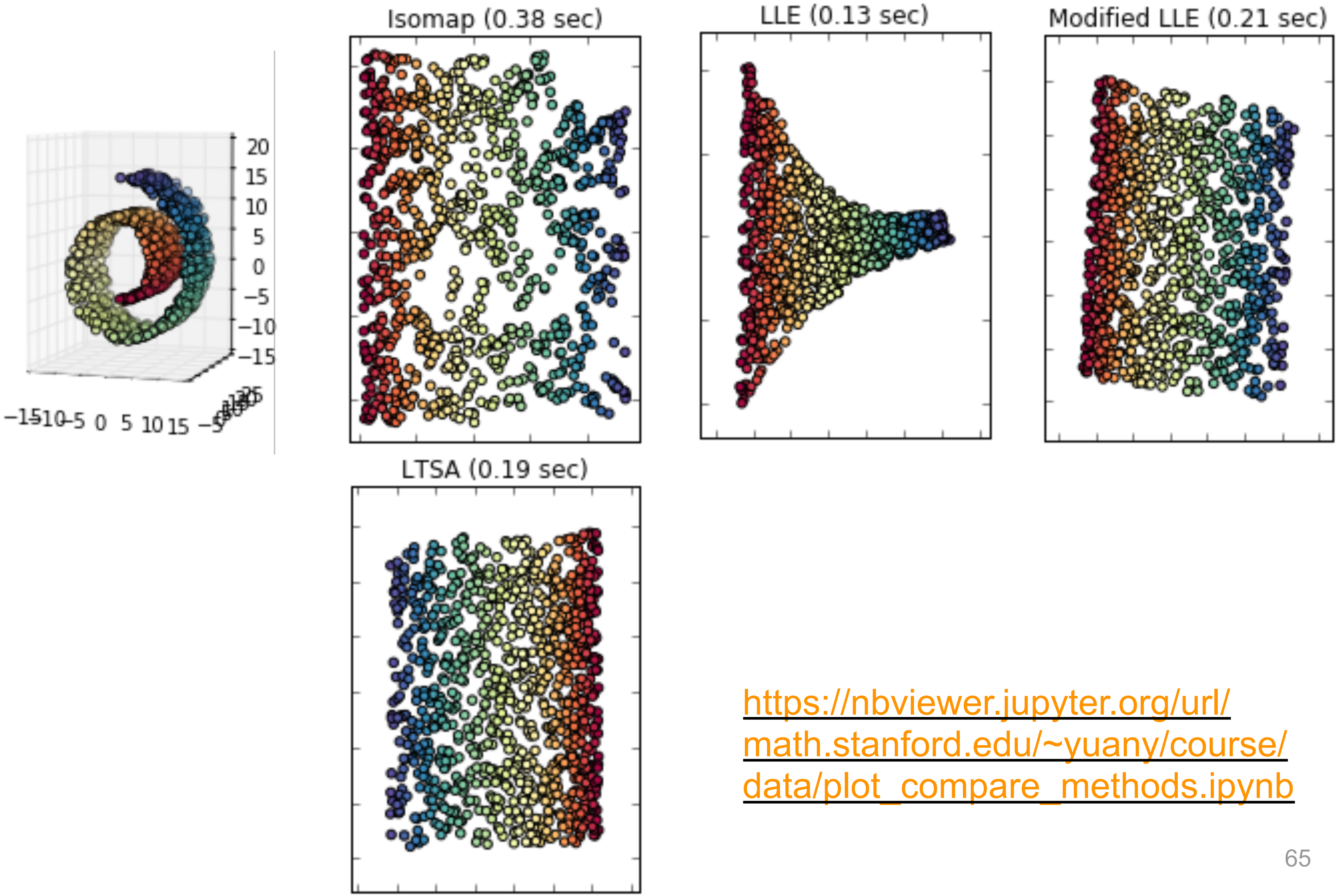
Step 2 (tangent space alignment): Alignment (kernel) matrix

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

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

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

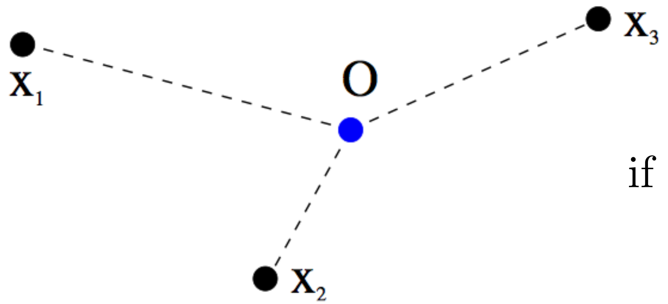
Summary..

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.

Hessian LLE (Eigenmap)

Hessian LLE



In LLE, one chooses the weights w_{ij} to minimize the following energy

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

if the points $\tilde{x}_j = x_j - x_i$ are linearly dependent

$$0 = \sum_{j\in\mathcal{N}_i} w_{ij}\tilde{x}_j, \quad \text{and} \quad 1 = \sum_{j\in\mathcal{N}_i} w_{ij}.$$

For any smooth function $y(x)$, consider its Taylor expansion up to the second order

$$y(x) = y(0) + x^T \nabla y(0) + \frac{1}{2} x^T (\mathcal{H}y)(0)x + o(\|x\|^2).$$

$$\begin{aligned}
(I - W)y(0) &:= y(0) - \sum_{j\in\mathcal{N}_i} w_{ij} y(\tilde{x}_j) \\
&\approx y(0) - \sum_{j\in\mathcal{N}_i} w_{ij} y(0) - \sum_{j\in\mathcal{N}_i} w_{ij} \tilde{x}_j^T \nabla y(0) - \frac{1}{2} \sum_{j\in\mathcal{N}_i} \tilde{x}_j^T (\mathcal{H}y)(0)\tilde{x}_j \\
&= -\frac{1}{2} \sum_{j\in\mathcal{N}_i} \tilde{x}_j^T (\mathcal{H}y)(0)\tilde{x}_j.
\end{aligned}$$

Hessian Null

The Hessian matrix

$$(\mathcal{H}y)(0) := \left[\frac{\partial^2 y(x)}{\partial x(i)\partial x(j)}\right]_{x=0} = 0,$$

if function $y(x)$ is a linear transform of the coordinates $x \in \mathbb{R}^p$ in the tangent space at x_i. In this case $(I - W)y(0) = 0$ and y reaches a minimizer.

In other words, the kernel of $(\mathcal{H}y)$ has dimension $d+1$, consisting the constant function and d linearly independent coordinates. Inspired by such an observation, Donoho and Grimes [**DG03b**] proposed Hessian LLE (Eigenmap) in search of

$$\min_{y \perp \mathbf{1}} \int \|\mathcal{H}y\|^2, \quad \|y\| = 1.$$

Hessian LLE Algorithm (I)

Algorithm 7: Hessian LLE Algorithm

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

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

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

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

Step 1: Compute local PCA on neighborhood of x_i, for,

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

where $\mu_i = \sum_{j=1}^k x_{i_j} = \frac{1}{k} X_i \mathbf{1}$;

- Left top singular vectors $\{\tilde{U}_1^{(i)}, ..., \tilde{U}_d^{(i)}\}$ give an orthonormal basis of the approximate tangent space at x_i,
- Right top singular vectors $[\tilde{V}_1^{(i)}, ..., \tilde{V}_d^{(i)}]$ are representation coordinates in the tangent space of local sample points around x_i.

Continued...

Hessian LLE Algorithm (II)

Step 2: Null Hessian estimation: define

$$M = [1, \tilde{V}_1, ..., \tilde{V}_d, \tilde{V}_1^2, \tilde{V}_1 \odot \tilde{V}_2, ..., \tilde{V}_{d-1} \odot \tilde{V}_d, \tilde{V}_d^2] \in \mathbb{R}^{k\times(1+d+\binom{d+1}{2})}$$

where $\tilde{V}_i \odot \tilde{V}_j = [\tilde{V}_{ik}\tilde{V}_{jk}]^T \in \mathbb{R}^k$ denotes the elementwise product (Hadamard product) between vector $\tilde{V}_i$ and $\tilde{V}_j$. Now we perform a Gram-Schmidt Orthogonalization procedure on M, get

$$\tilde{M} = [1, \hat{v}_1, ..., \hat{v}_d, \hat{w}_1, \hat{w}_2, ..., \hat{w}_{\binom{d+1}{2}}] \in \mathbb{R}^{k\times(1+d+\binom{d+1}{2})}$$

Define

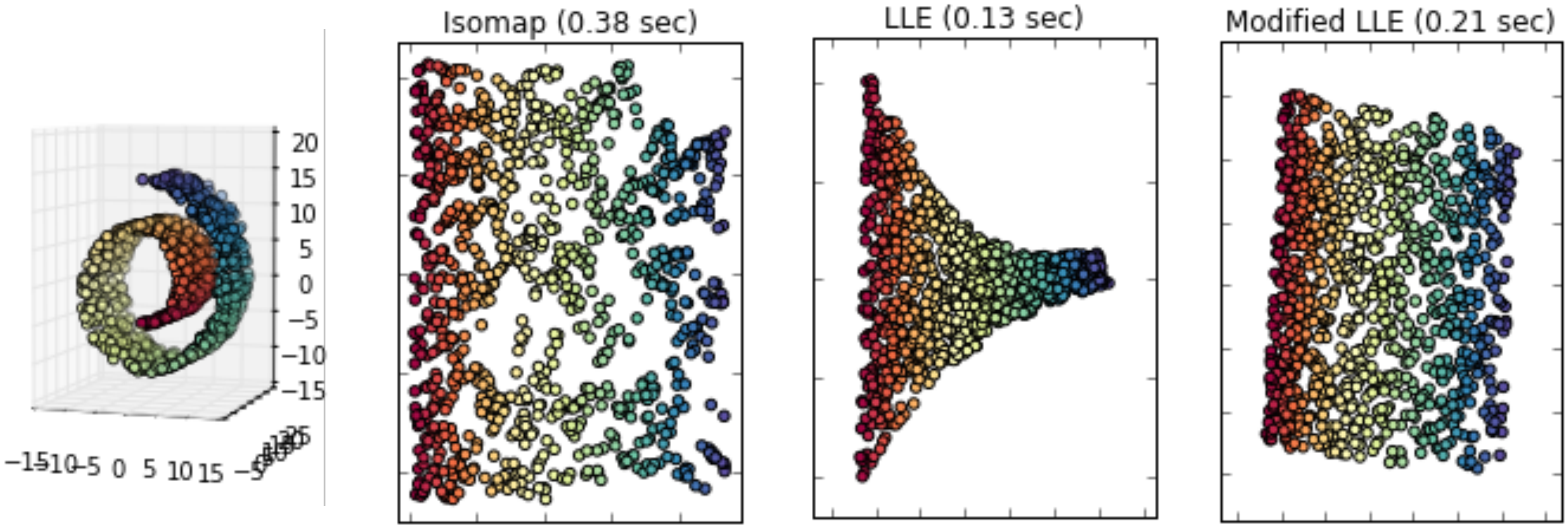
$$[H^{(i)}]^T = [last \quad \binom{d+1}{2} \quad columns \quad of \quad \tilde{M}]_{k\times\binom{d+1}{2}}.$$

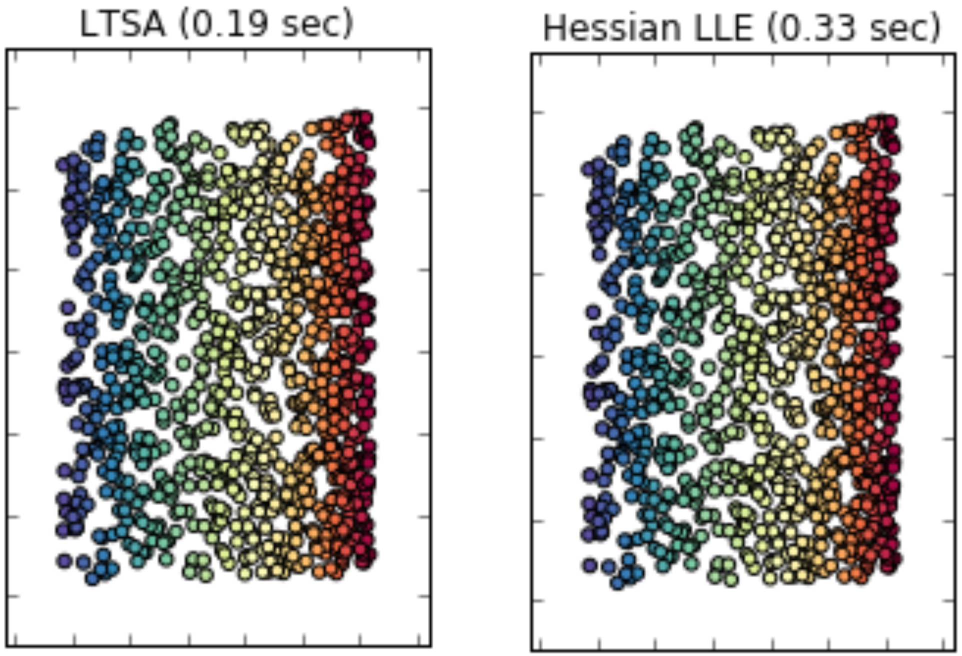
Step 3: Define

$$K = \sum_{i=1}^{n} S^{(i)} H^{(i)T} H^{(i)} S^{(i)T} \in \mathbb{R}^{n\times n}, \quad [x_1, .., x_n] S^{(i)} = [x_{i_1}, ..., x_{i_k}],$$

find smallest $d+1$ eigenvectors of K and drop the smallest eigenvector, and the remaining d eigenvectors will give rise to a d-embedding.

Comparisons on Swiss Roll

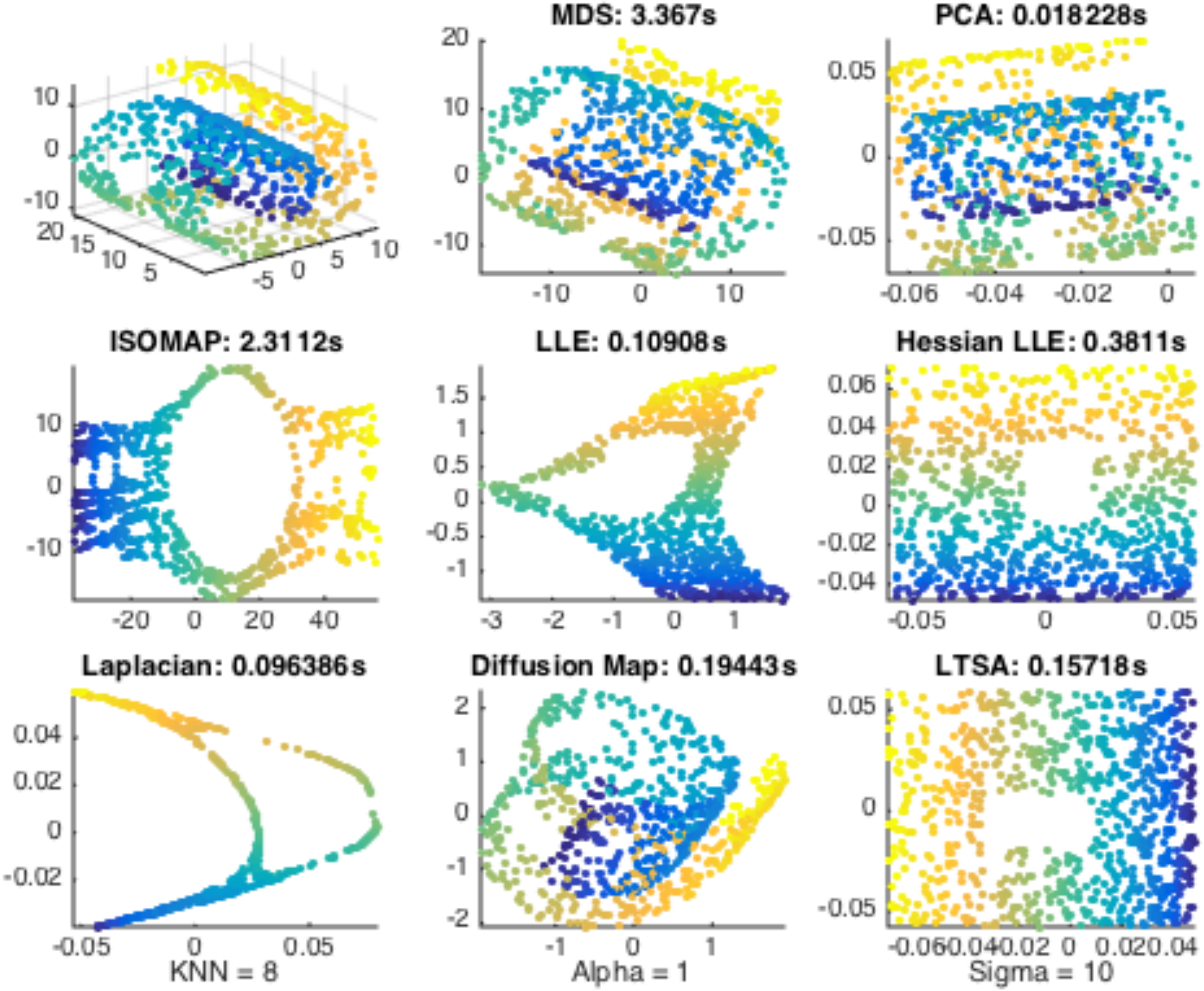




https://nbviewer.jupyter.org/url/math.stanford.edu/~yuany/course/data/plot_compare_methods.ipynb

Comparisons on Swiss Roll with a Hole

- mani.m



Two Assumptions on ISOMAP

(ISO1) *Isometry.* The mapping ψ preserves geodesic distances. That is, define a distance between two points m and m' on the manifold according to the distance travelled by a bug walking along the manifold M according to the shortest path between m and m'. Then the isometry assumption says that

$$G(m, m') = |\theta - \theta'|, \qquad \forall m \leftrightarrow \theta, m' \leftrightarrow \theta',$$

where $|\cdot|$ denotes Euclidean distance in $\mathbb{R}^d$.

(ISO2) *Convexity.* The parameter space Θ is a convex subset of $\mathbb{R}^d$. That is, if θ, θ' is a pair of points in Θ, then the entire line segment $\{(1-t)\theta + t\theta' : t \in (0,1)\}$ lies in Θ.

Convexity is hard to meet: consider two balls in an image which never intersect, whose center coordinate space (x_1,y_1,x_2,y_2) must have a hole.

Relaxations (Donoho-Grimes'2003)

(LocISO1) *Local Isometry.* In a small enough neighborhood of each point m, geodesic distances to nearby points m' in M are identical to Euclidean distances between the corresponding parameter points θ and θ'.

(LocISO2) *Connectedness.* The parameter space Θ is a open connected subset of $\mathbb{R}^d$.

Convergence of Hessian LLE (Donoho-Grimes)

Theorem 1 *Suppose $M = \psi(\Theta)$ where Θ is an open connected subset of $\mathbb{R}^d$, and ψ is a locally isometric embedding of Θ into $\mathbb{R}^n$. Then $\mathcal{H}(f)$ has a $d+1$ dimensional nullspace, consisting of the constant function and a d-dimensional space of functions spanned by the original isometric coordinates.*

We give the proof in Appendix A.

Corollary 2 *Under the same assumptions as Theorem 1, the original isometric coordinates θ can be recovered, up to a rigid motion, by identifying a suitable basis for the null space of $\mathcal{H}(f)$.*

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.