Manifold Learning I: ISOMAP and LLE

姚 遠

2017

Python scikit-learn Manifold learning Toolbox

- http://scikit-learn.org/stable/modules/manifold.html
- MDS/PCA, ISOMAP/LLE
- Hessian Eigenmap
- Laplacian Eigenmap
- LTSA

Matlab Dimensionality Reduction Toolbox

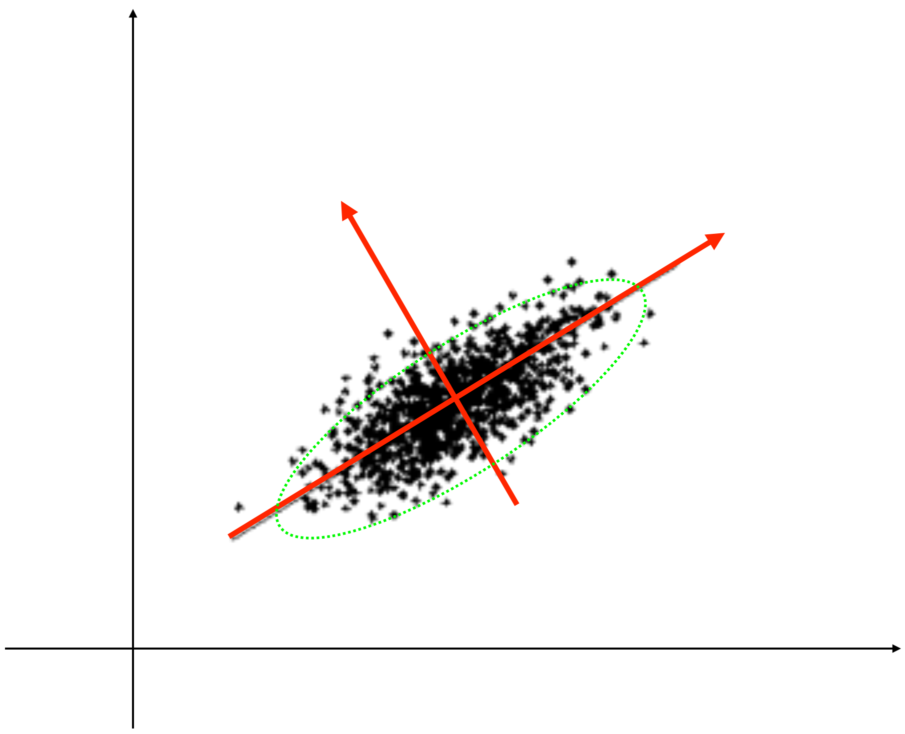
- http://homepage.tudelft.nl/19j49/Matlab_Toolbox_for_Dimensionality_Reduction.html
- drtoolbox contains:
  - Principal Component Analysis (PCA), Probabilistic PC
  - Factor Analysis (FA), Sammon mapping, Linear Discriminant Analysis (LDA)
  - Multidimensional scaling (MDS), Isomap, Landmark Isomap
  - 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
  - …

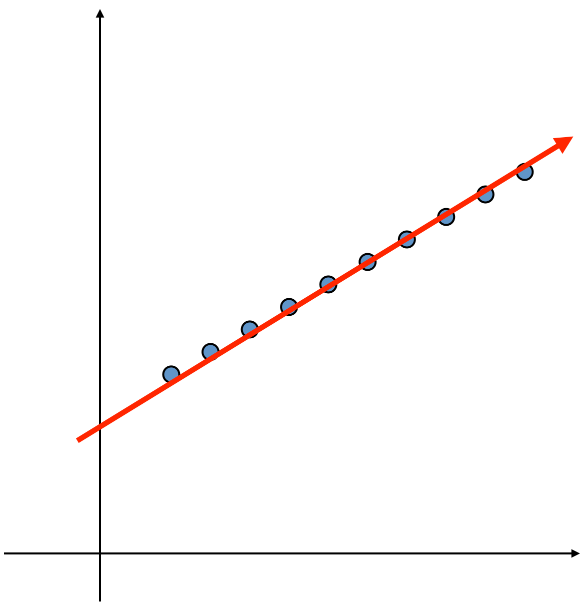
Recall: PCA

- Principal Component Analysis (PCA)

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

EigenValue Decomposition of XX^T

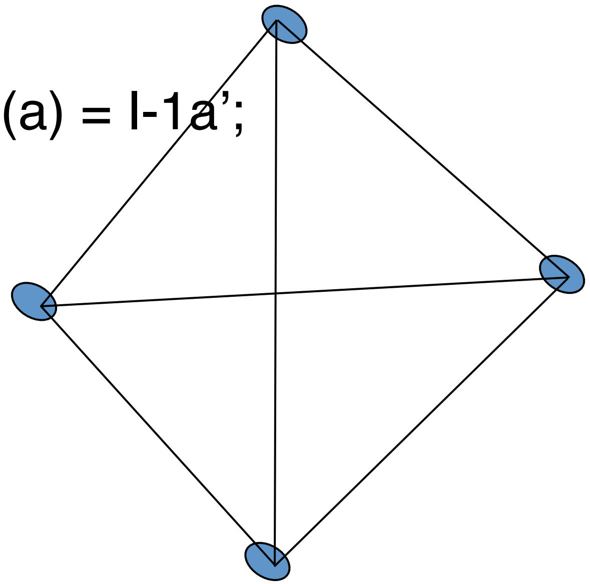




One Dimensional Manifold

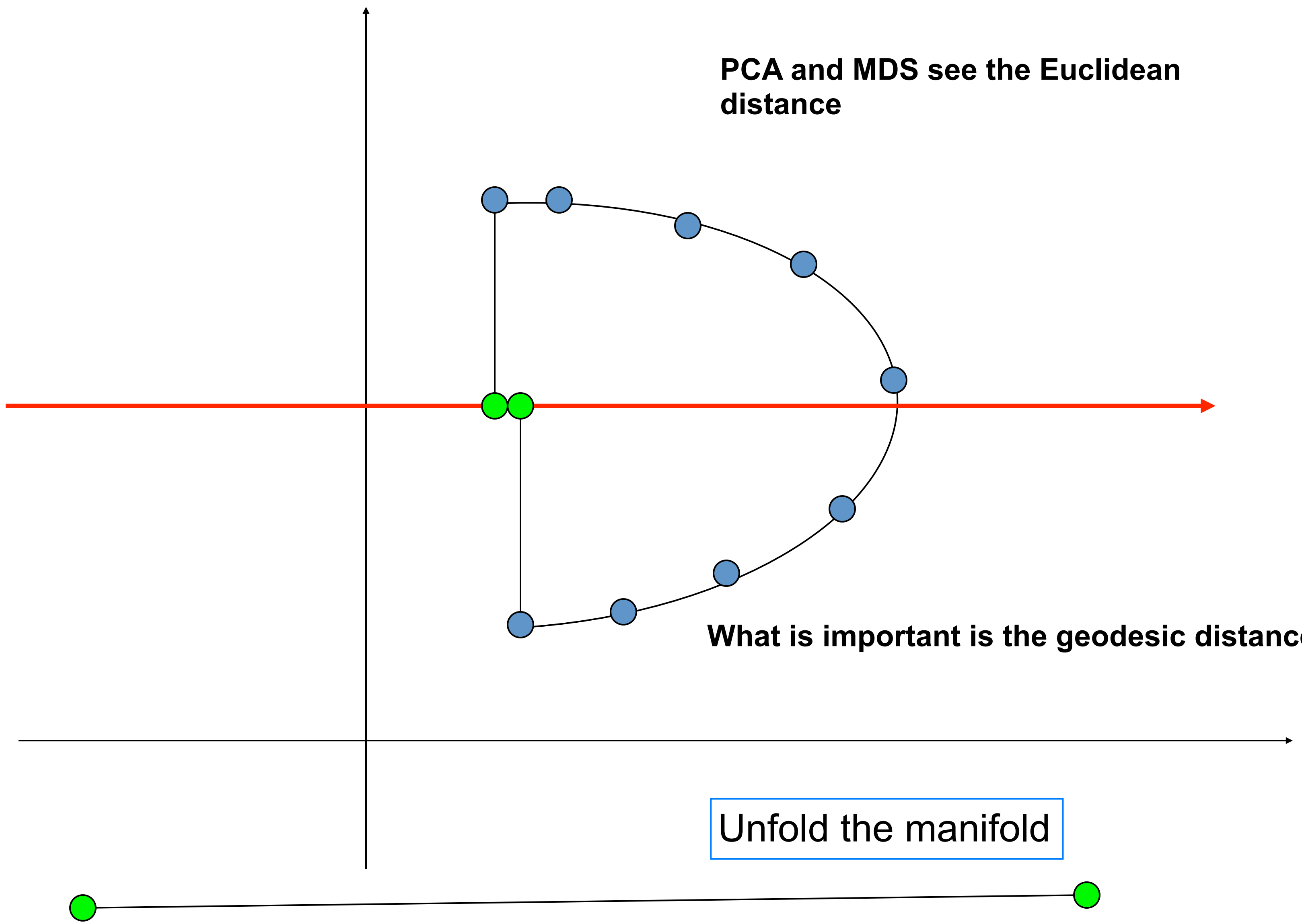
Recall: MDS

- 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-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 $B = YY^T$



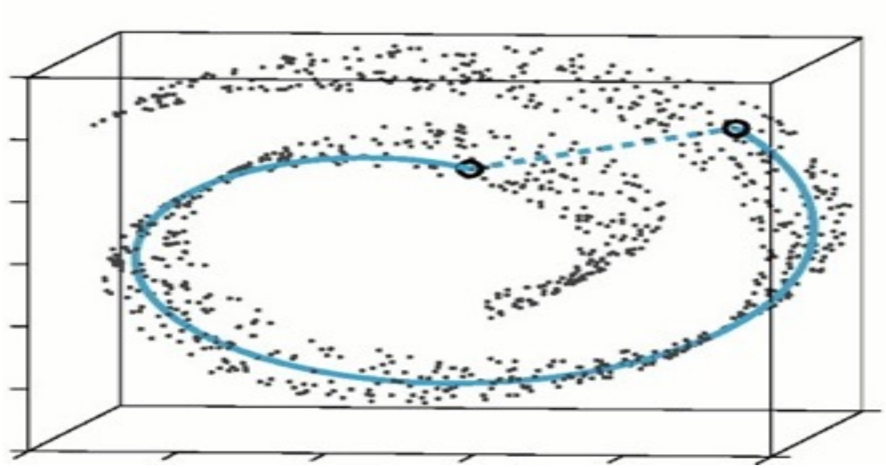
If we preserve the pairwise Euclidean distances do we preserve the structure??

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

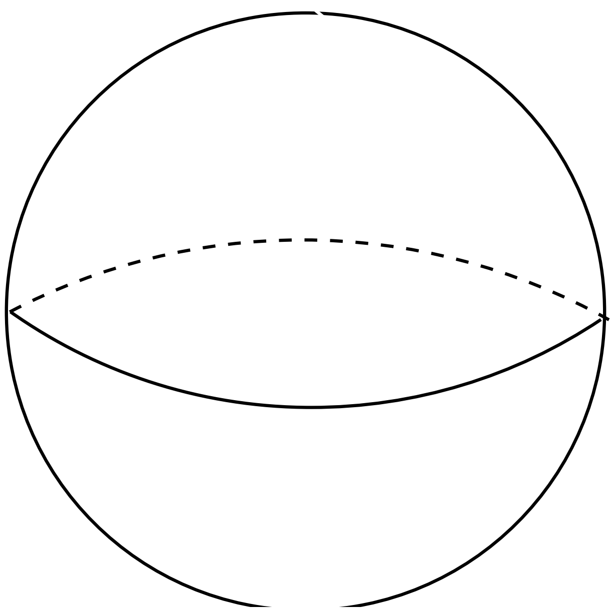
  curvature, geodesics

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

Embedded (sub-)Manifolds

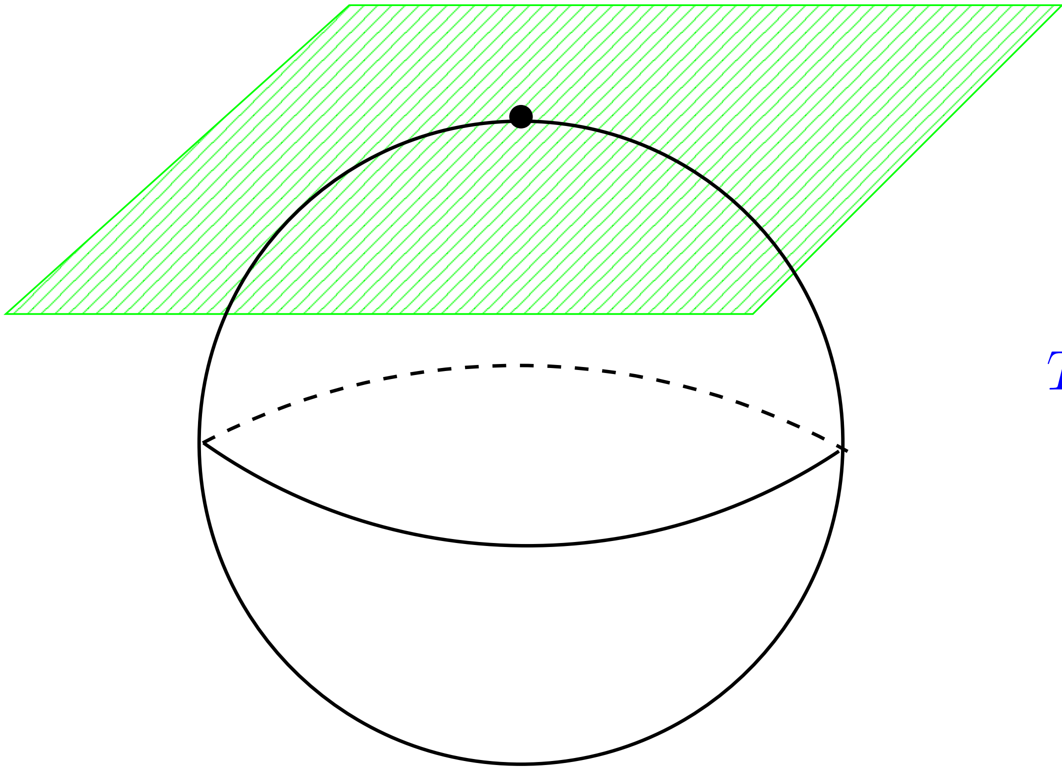
$$\mathcal{M}^k \subset \mathbb{R}^N$$

Locally (not globally) looks like Euclidean space.



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

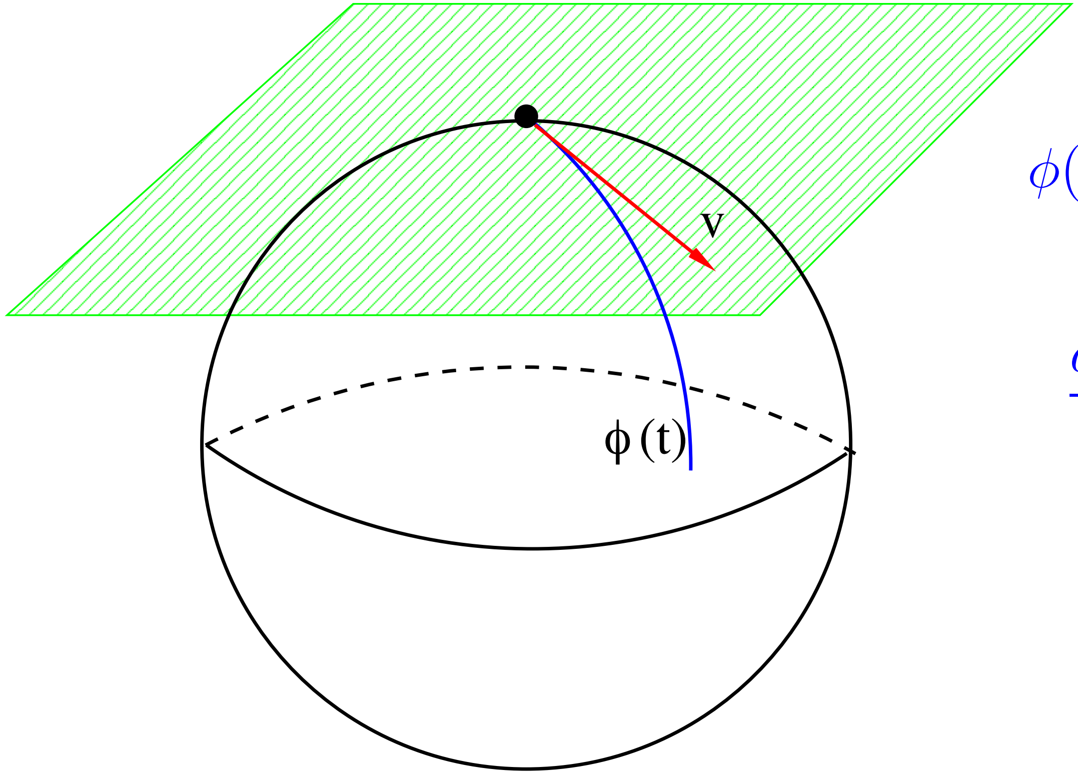
Tangent Space



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

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

Tangent Vectors and Curves



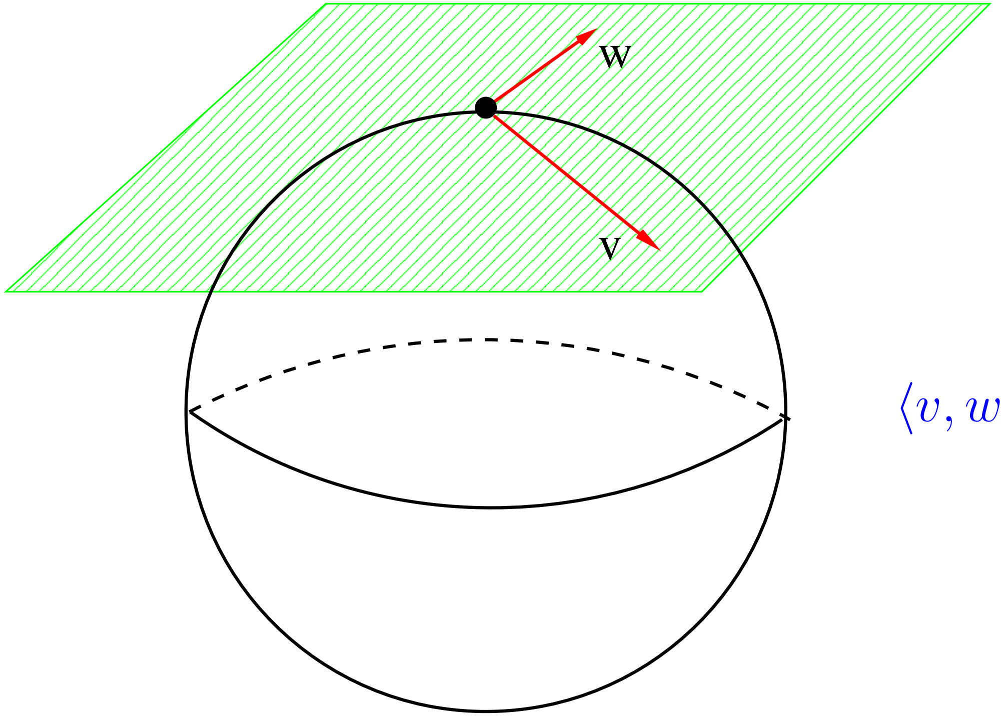
$$\phi(t) : \mathbb{R} \to \mathcal{M}^k$$

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

Tangent vectors <———> curves.

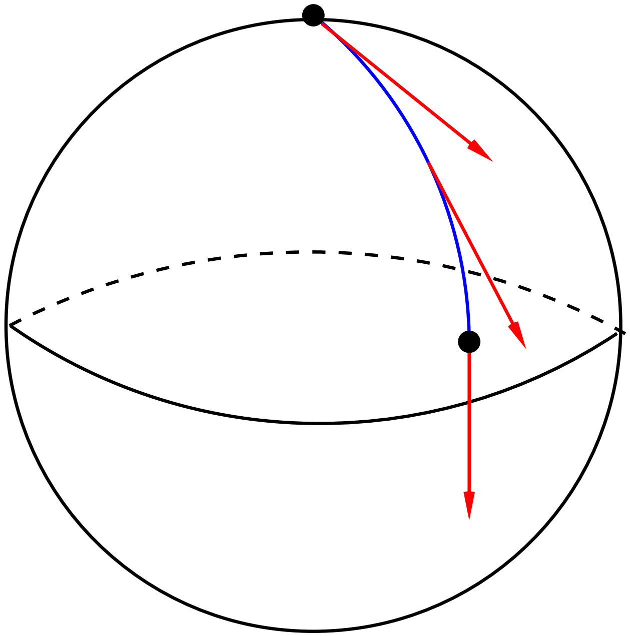
Riemannian Geometry

Norms and angles in tangent space.



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

Geodesics



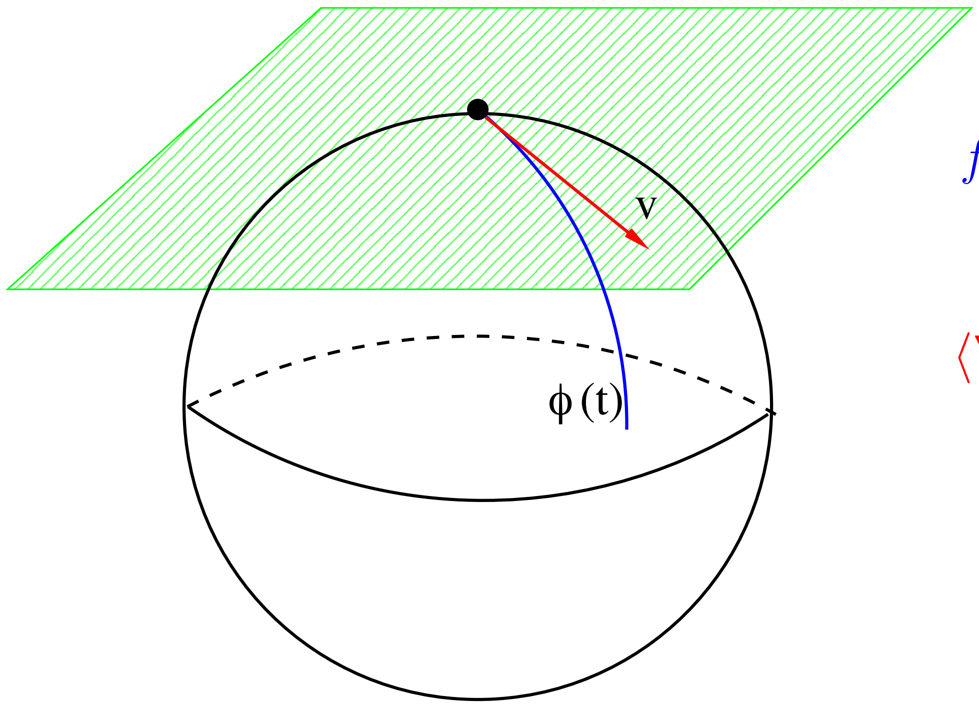
$$\phi(t) : [0, 1] \to \mathcal{M}^k$$

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

Can measure length using norm in tangent space.

Geodesic — shortest curve between two points.

Gradients



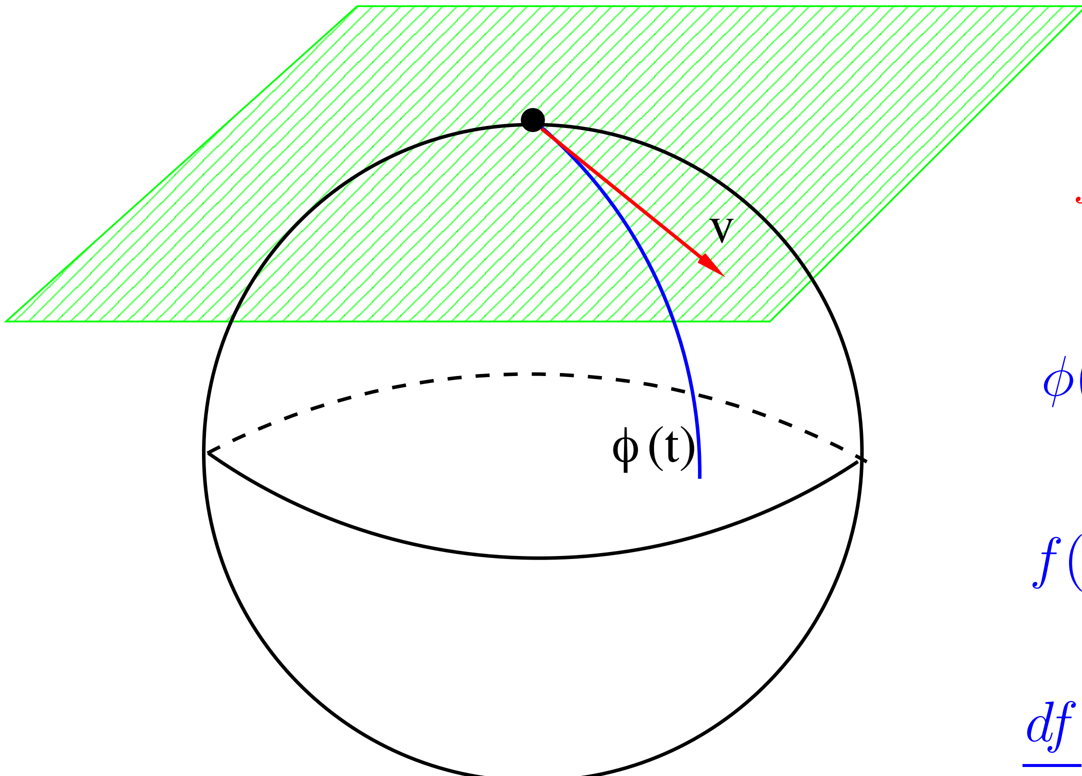
$$f : \mathcal{M}^k \rightarrow \mathbb{R}$$

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

Tangent vectors <——> Directional derivatives.

Gradient points in the direction of maximum change.

Tangent Vectors vs. Derivatives



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

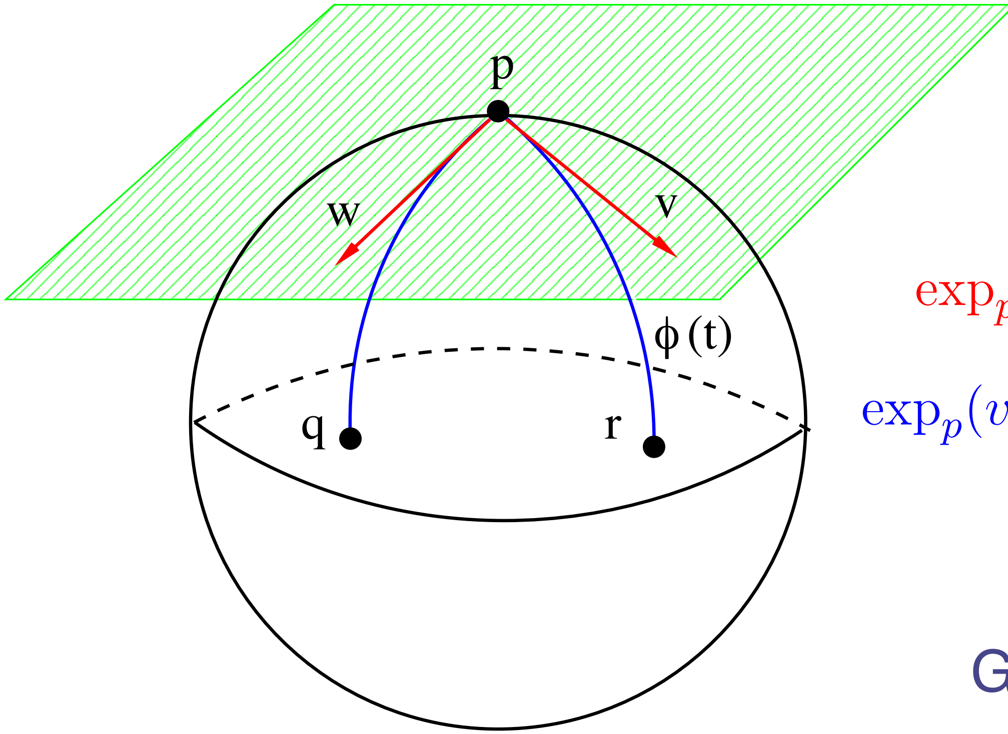
$$\phi(t) : \mathbb{R} \to \mathcal{M}^k$$

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

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

Tangent vectors <———> Directional derivatives.

Exponential Maps



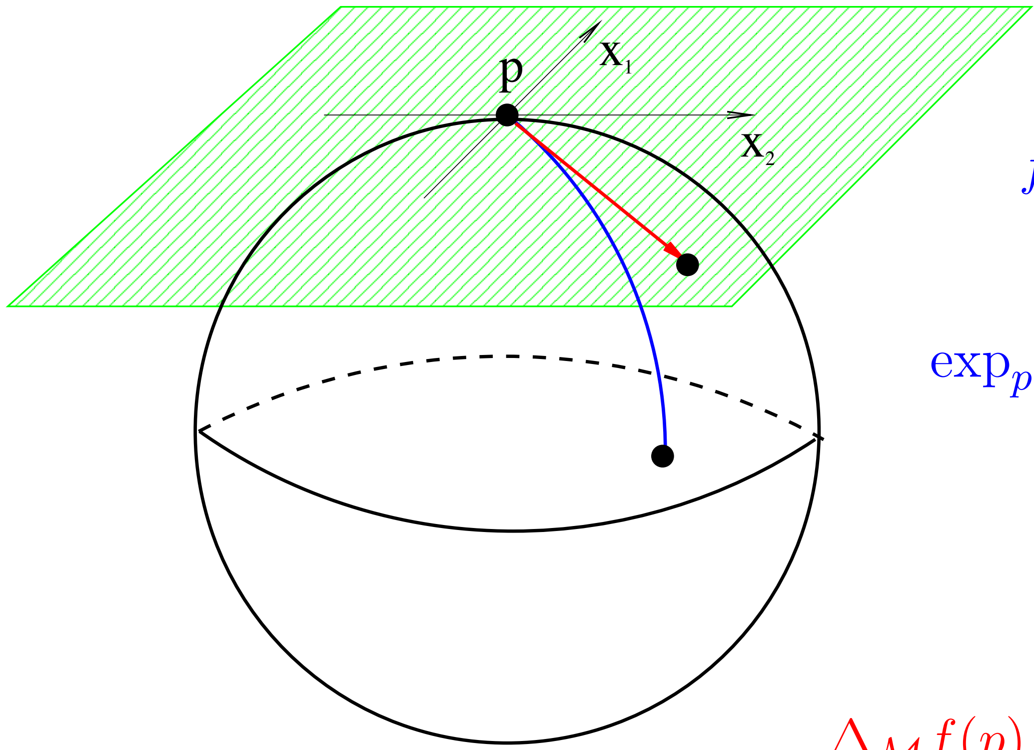
$$\exp_p : T_p\mathcal{M}^k \to \mathcal{M}^k$$

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

Geodesic $\phi(t)$

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

Laplacian-Beltrami Operator



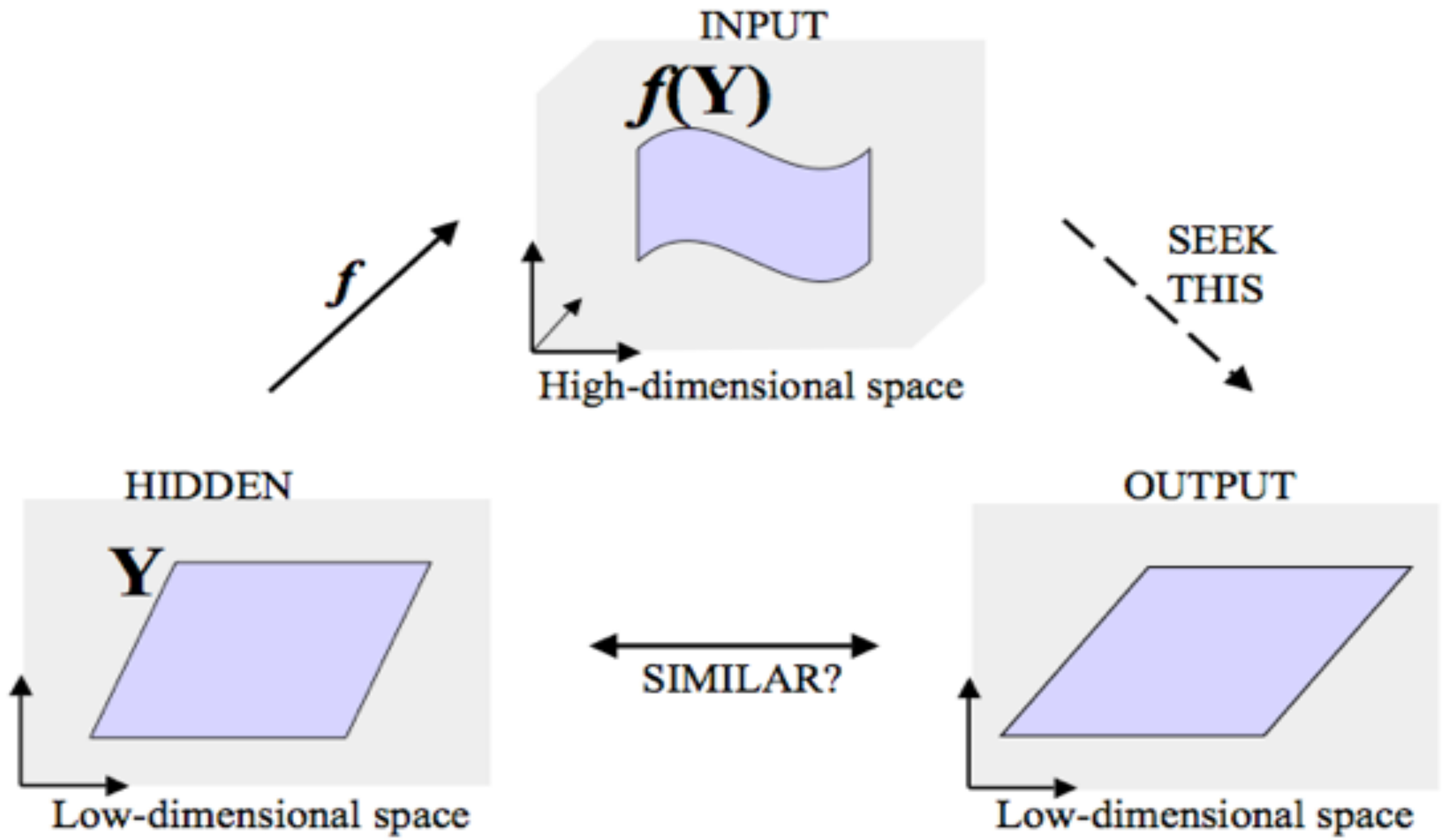
$$f : \mathcal{M}^k \to \mathbb{R}$$

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

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

Orthonormal coordinate system.

Generative Models in Manifold Learning



Spectral Geometric Embedding

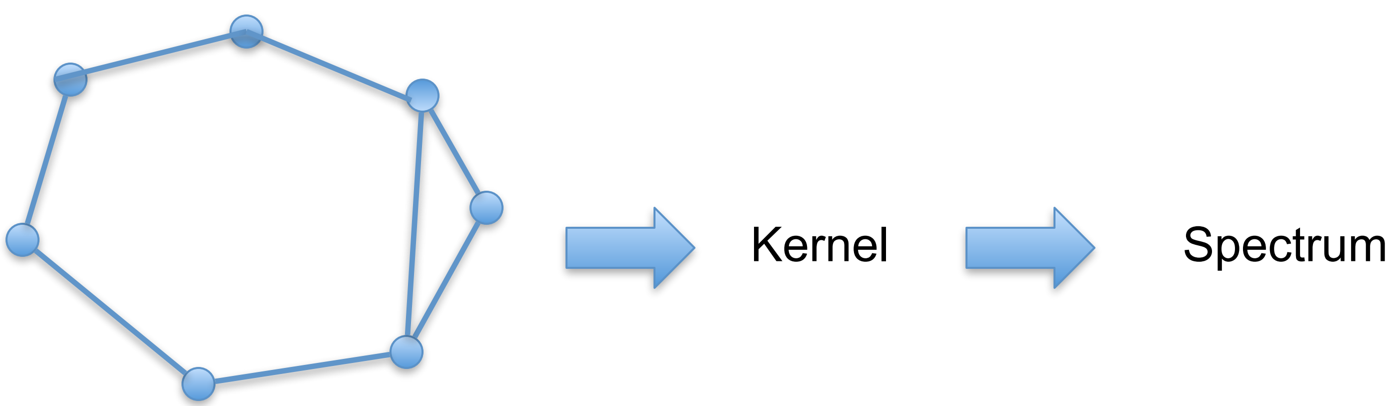
Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$,
Find $y_1, \ldots, y_n \in \mathbb{R}^d$ where $d << N$

- 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

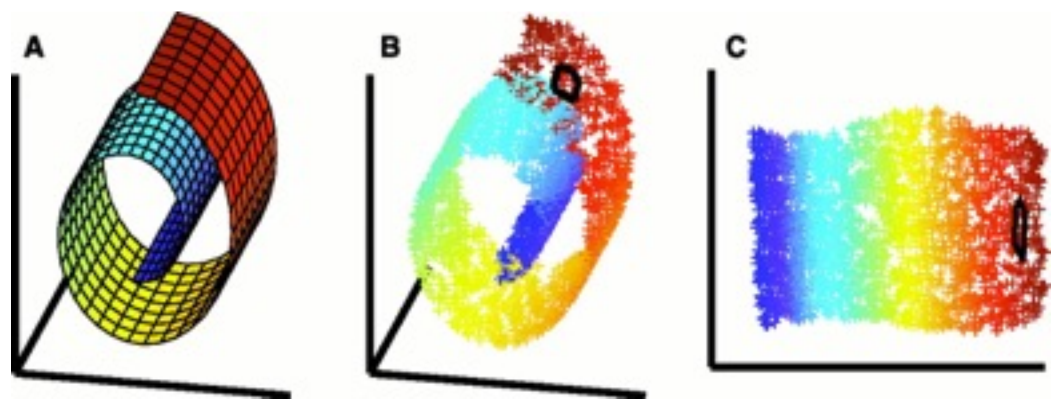
- Tenenbaum-de Silva-Langford Isomap Algorithm
  - Global approach.
  - On a low dimensional embedding
    - Nearby points should be nearby.
    - Faraway points should be faraway.

- 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

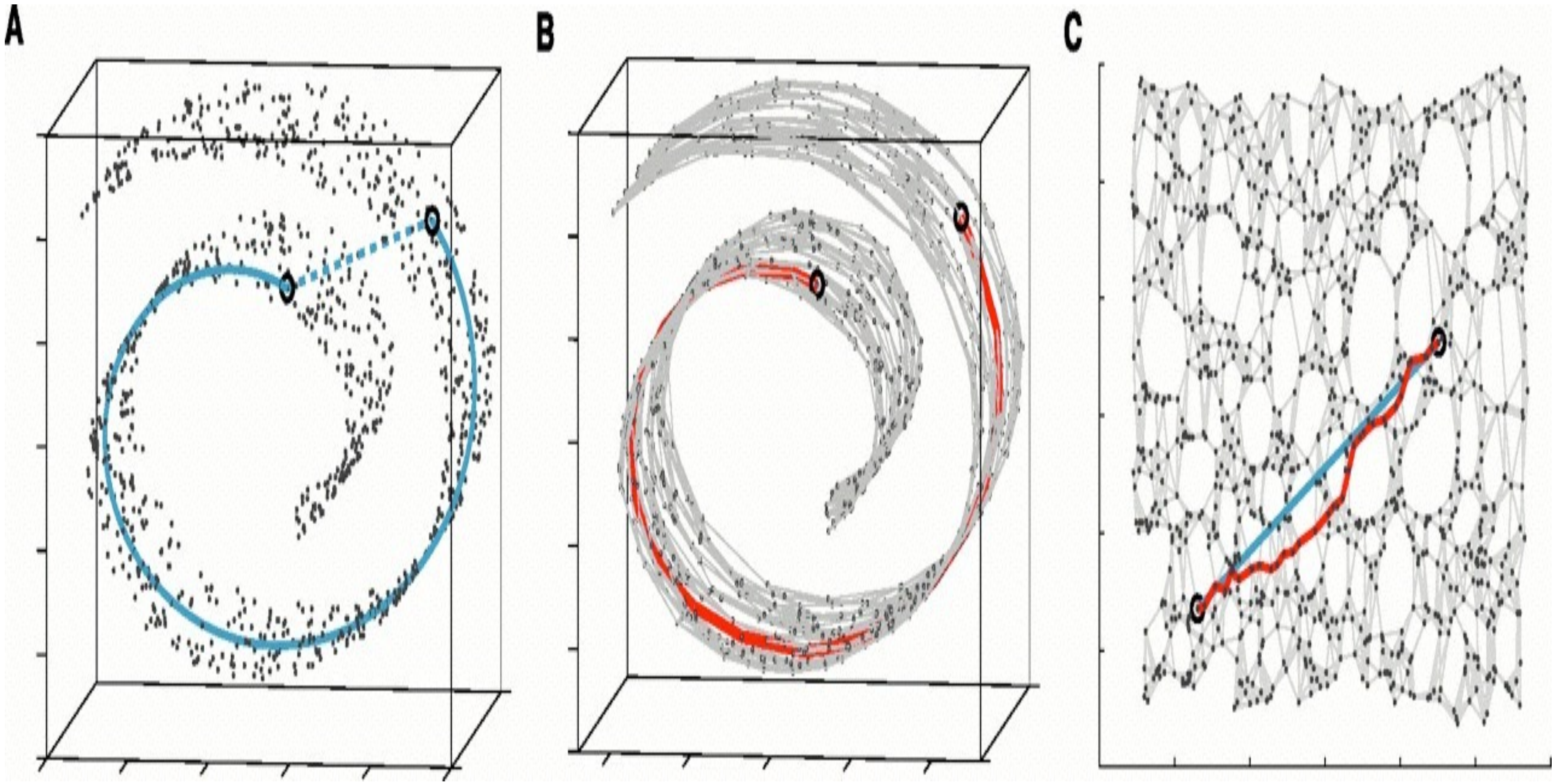
Once we have all pairwise geodesic distances use classical metric MDS

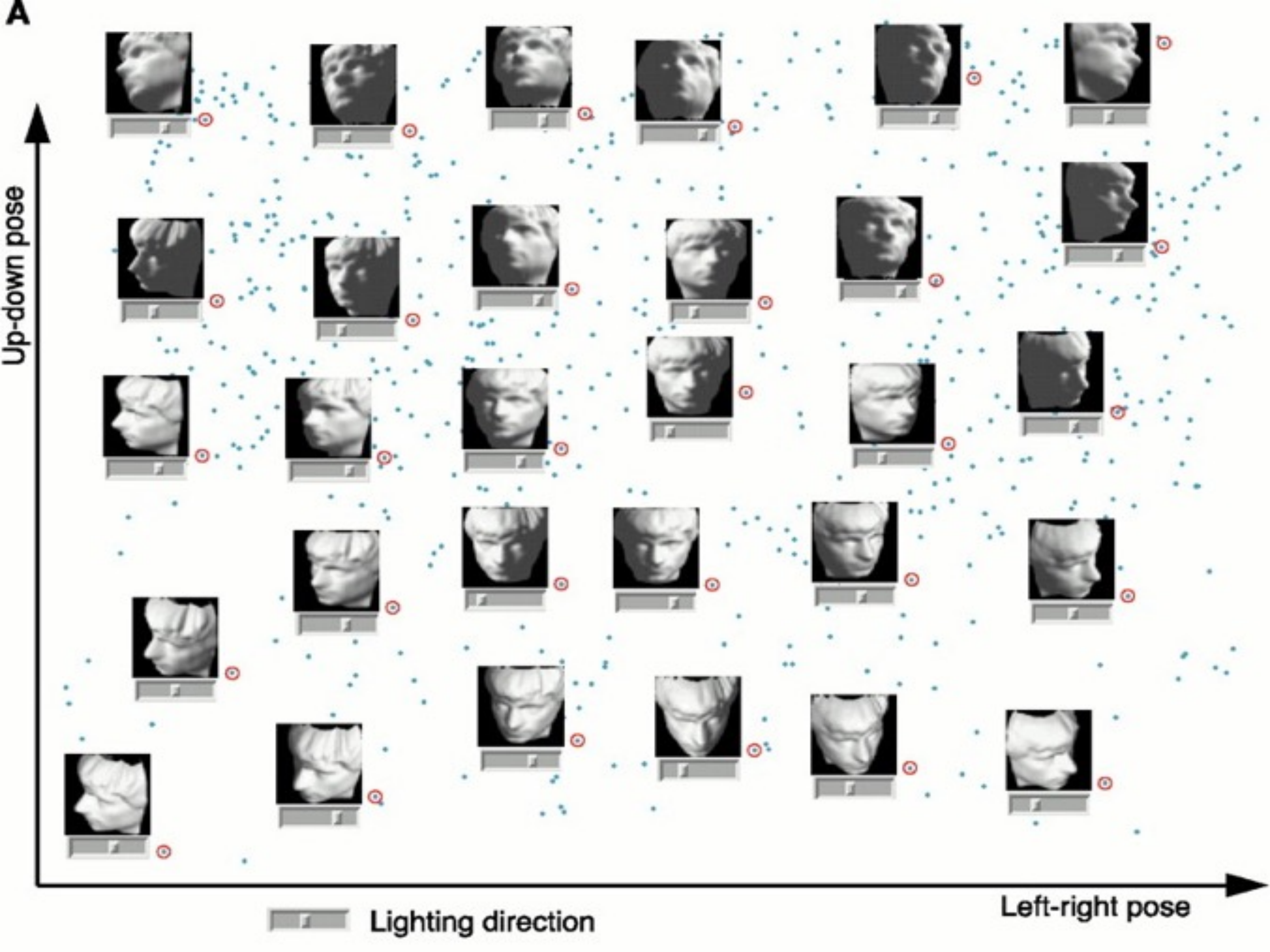


Isomap - Algorithm

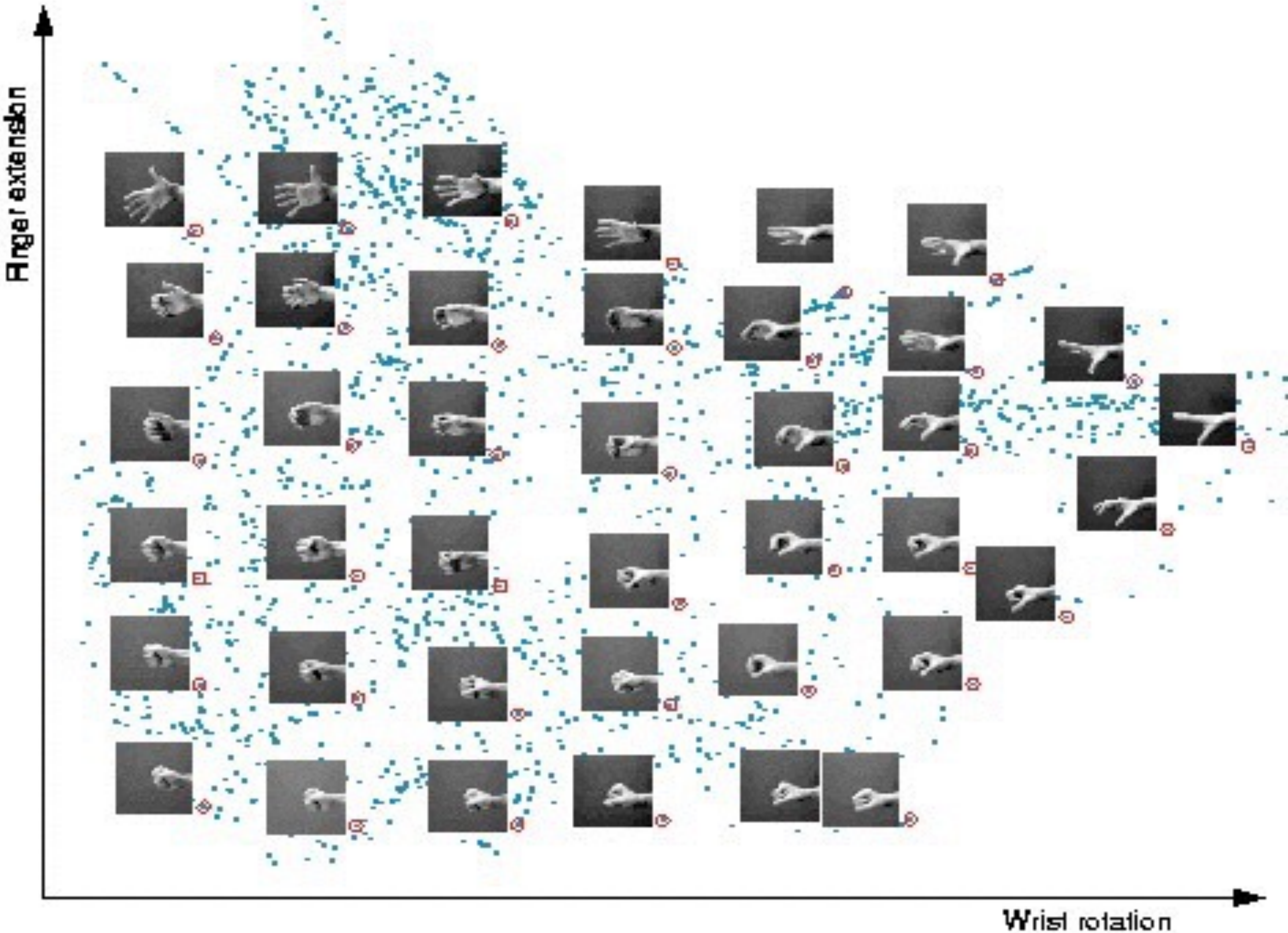
- Construct an n-by-n neighborhood graph
  - connecting points whose distances are within a fixed radius.
  - K nearest neighbor graph
- Compute the shortest path (geodesic) distances between nodes: D
  - Floyd's Algorithm (O(N^3))
  - Dijkstra's Algorithm (O($kN^2 \log N$))
- Construct a lower dimensional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

Isomap

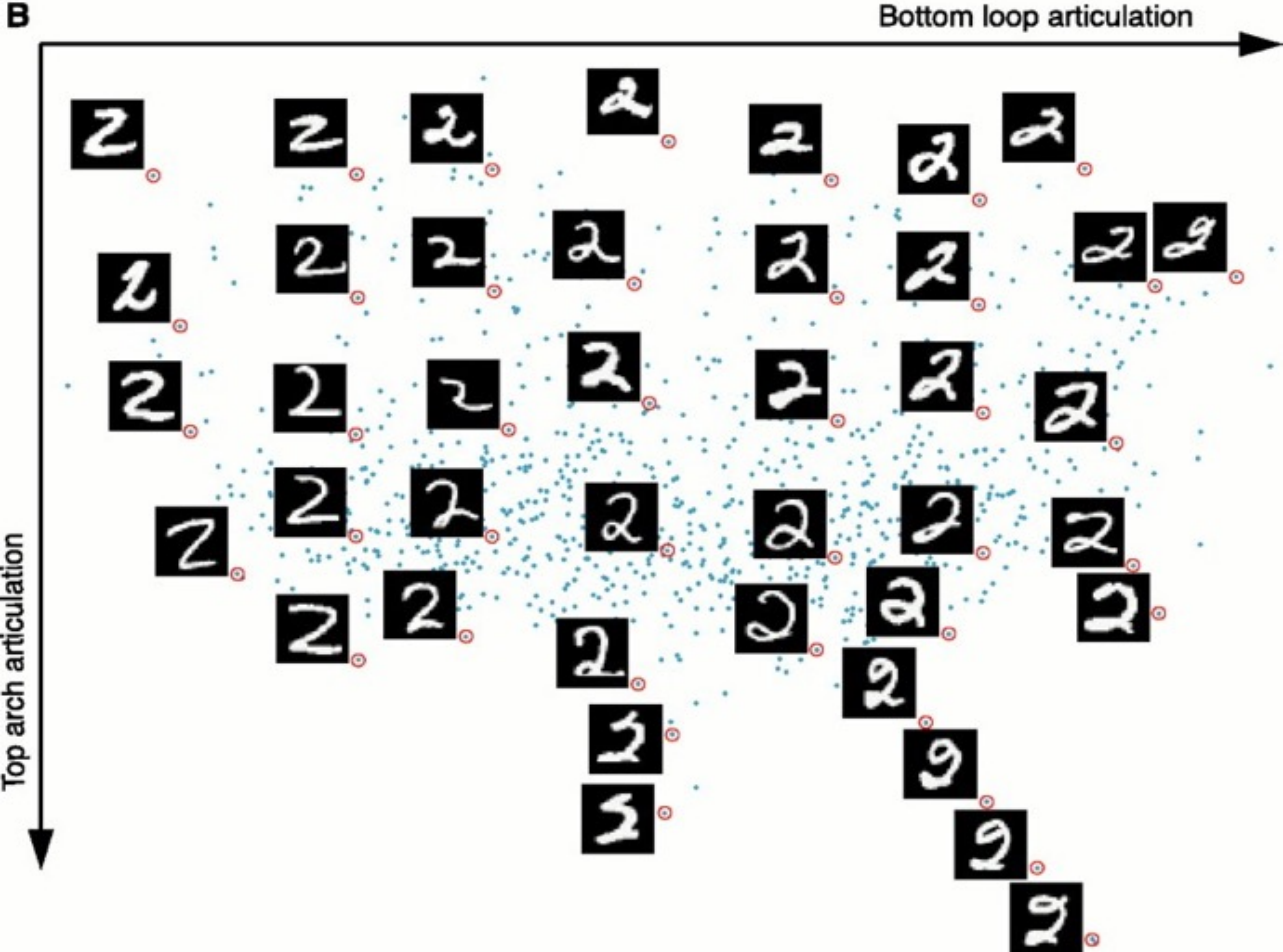




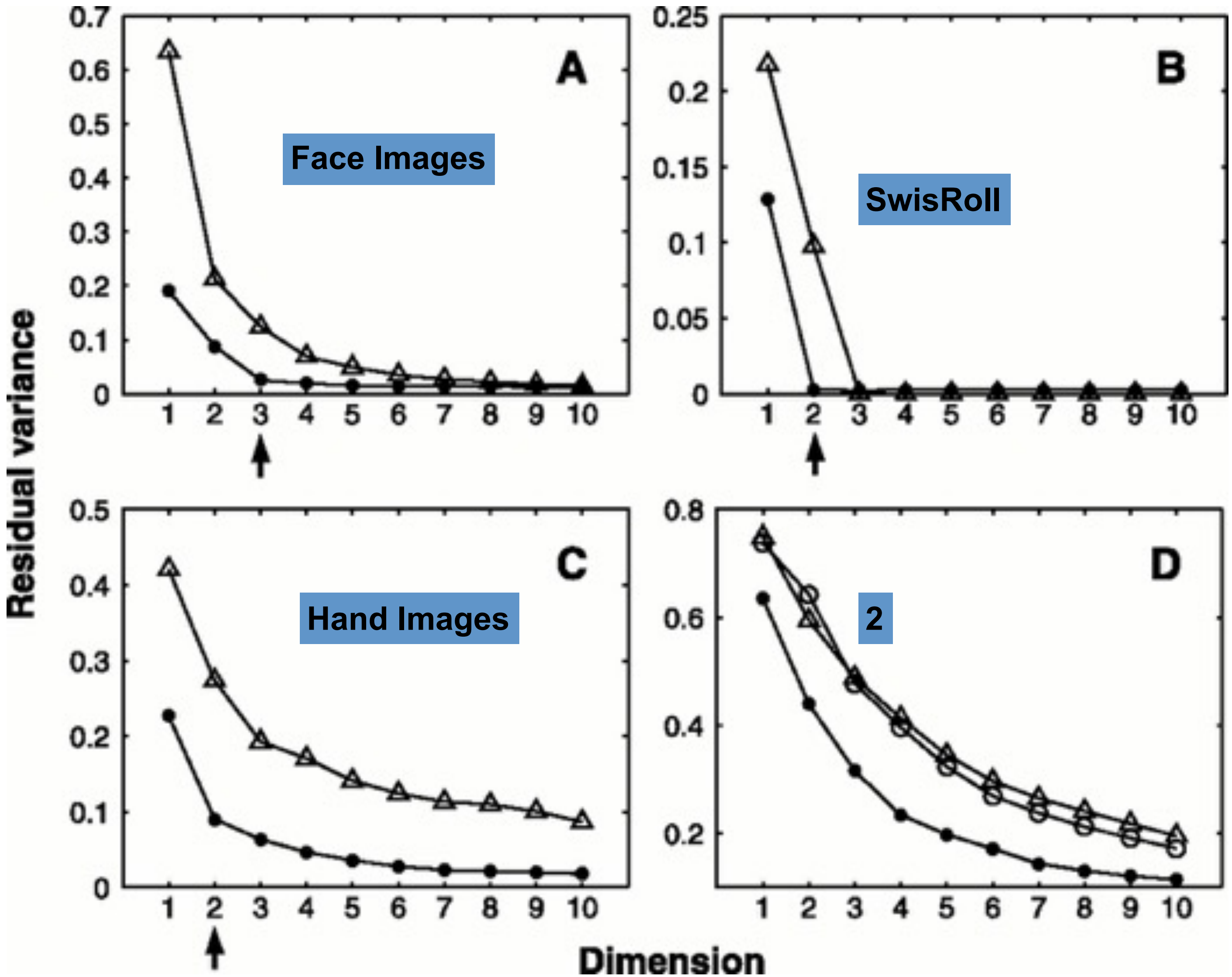
A
Up-down pose
Left-right pose
Lighting direction


Finger extension
Wrist rotation

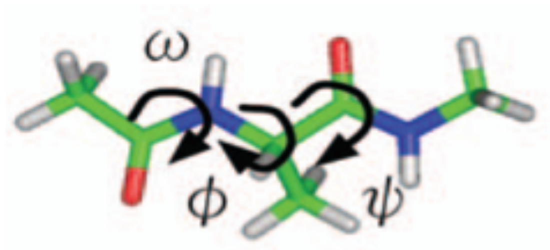
B

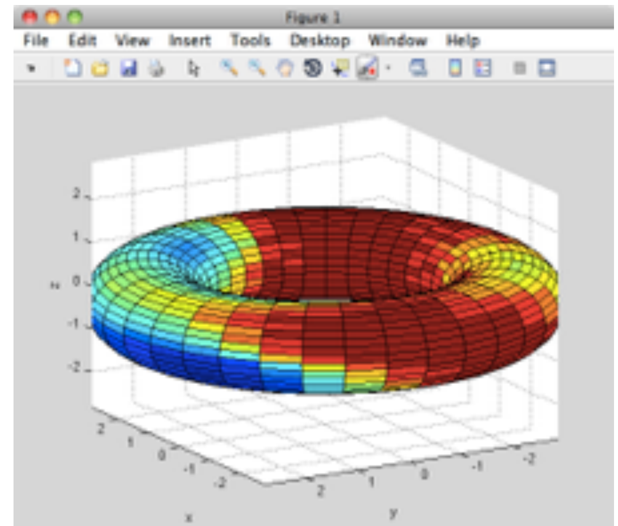


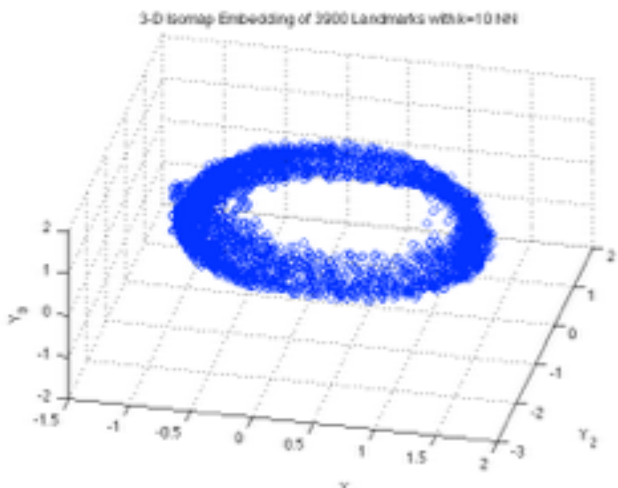
Residual Variance vs. Intrinsic Dimension

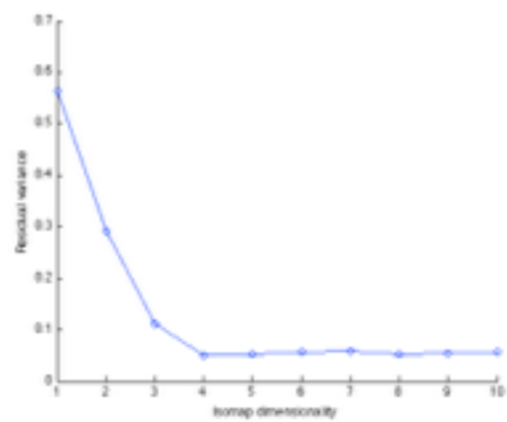


ISOMAP on Alanine-dipeptide









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 $d_G \approx d_M$.

Main 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.

Large Scale ISOMAP: Landmark and Nystrom

- 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

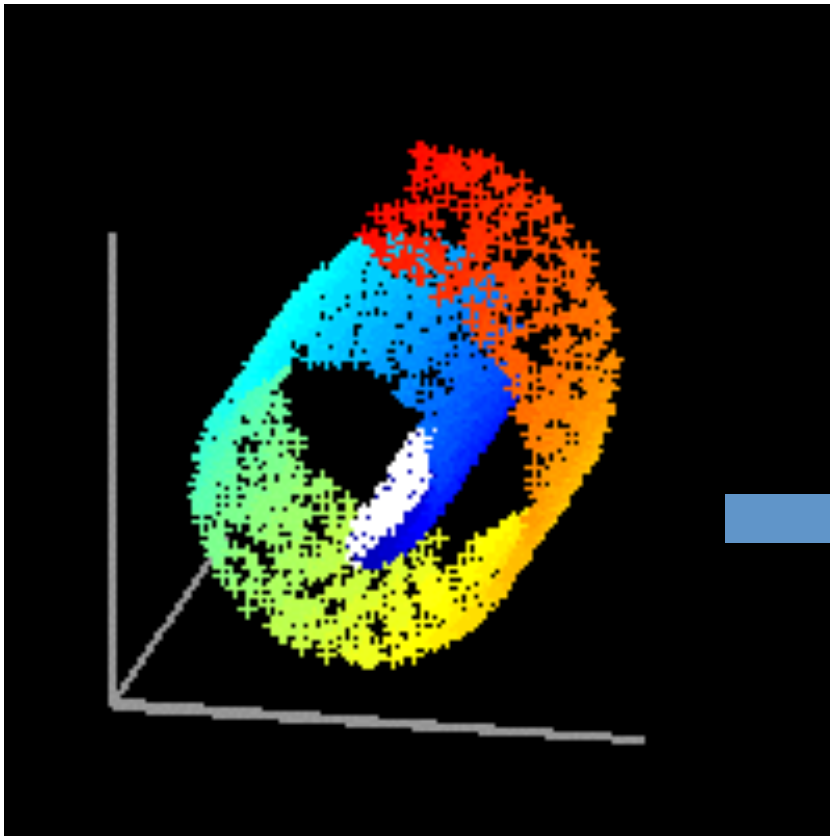
- Random
- MiniMax: k-center
- Hierarchical landmarks: cover-tree
- Nyström extension method

A Shortcoming of ISOMAP

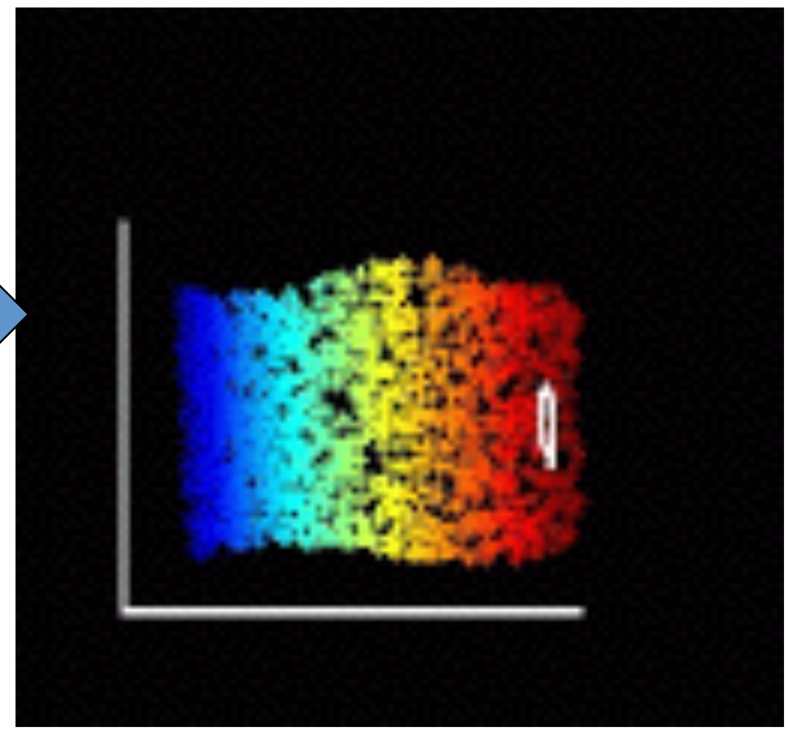
- One need to compute pairwise shortest path between all sample pairs (i,j)
  - Global
  - Non-sparse
  - Cubic complexity $O(N^3)$

Locally Linear Embedding

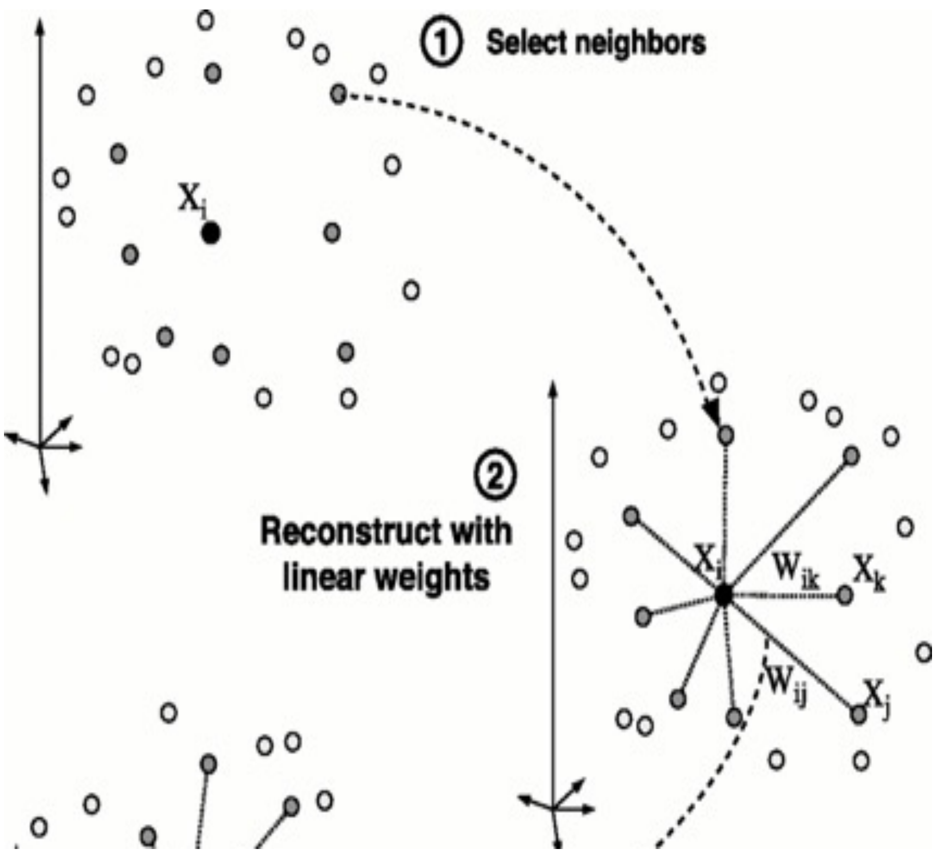
***manifold is a topological space which is locally Euclidean**.”*



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 are chosen to minimize the reconstruction Error.

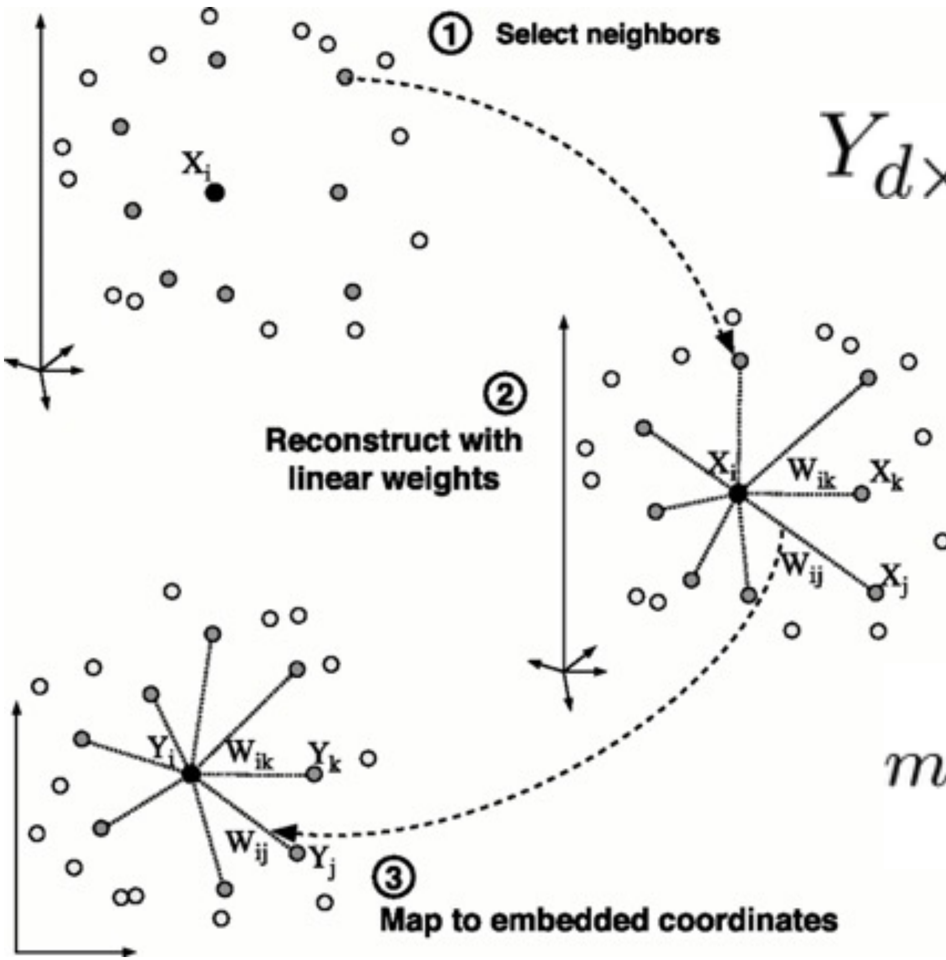
$$min_W \parallel X_i - \sum_{j=1}^{K} W_{ij} X_j \parallel^2 \qquad (1)$$

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…



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

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

LLE Algorithm (I)

(1) Construct a neighbor hood graph

(2) Local fitting:

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

Compute the local fitting weights

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

This can be done by Lagrange multiplier method, *i.e.* solving

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

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 $C^{(i)}_{jk} = \langle x_j - x_i, x_k - x_i\rangle$, whence the weights are

$$w_i = C_i^{\dagger}(\bar{X}_i^T x_i + \lambda \mathbf{1}),$$

where the Lagrange multiplier equals to

$$\lambda = \frac{1}{\mathbf{1}^T C_i^{\dagger} \mathbf{1}} \left(1 - \mathbf{1}^T C_i^{\dagger} \bar{X}_i^T x_i\right),$$

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 (II)

(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,

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

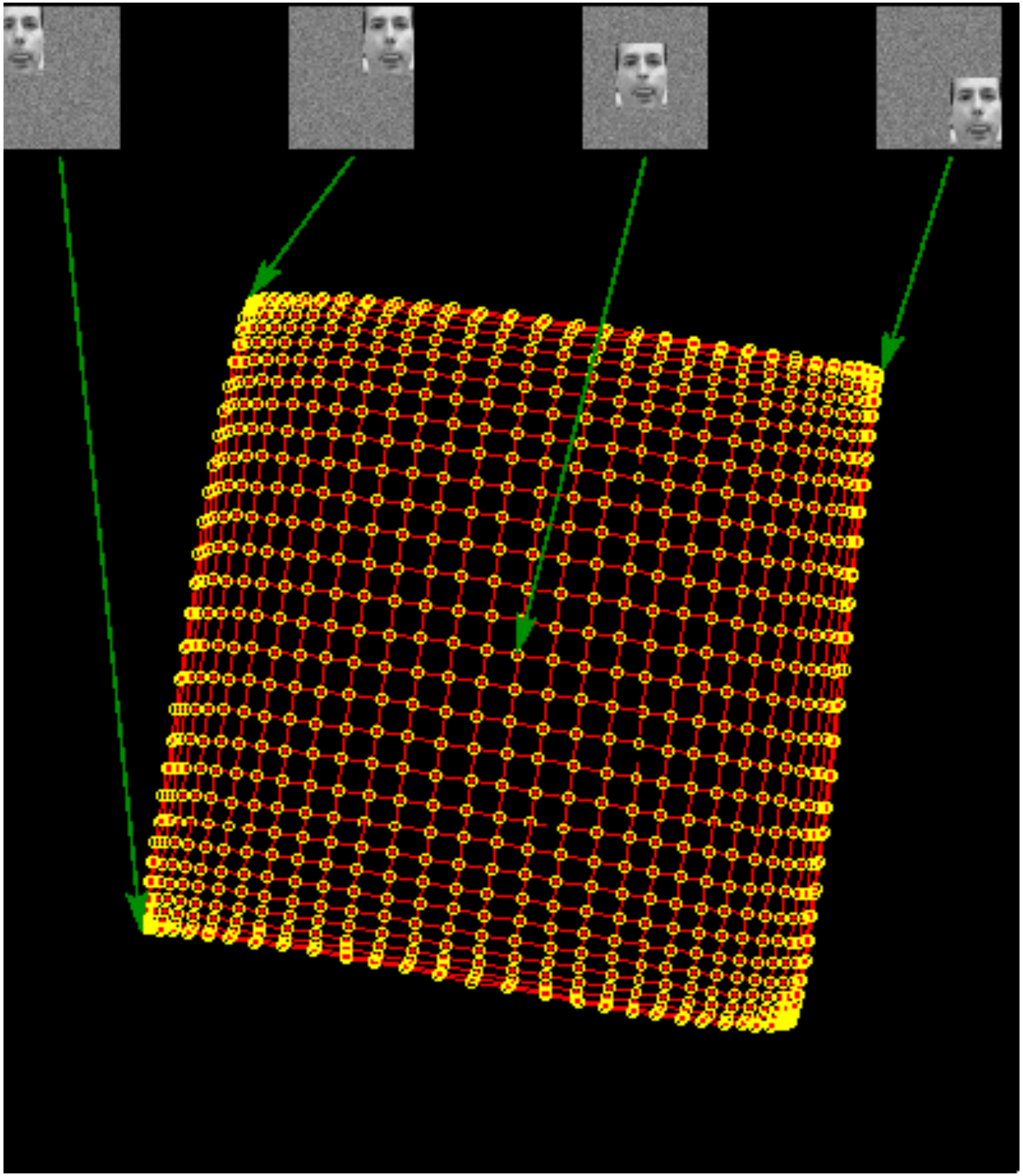
In other words, construct a positive semi-definite matrix $B = (I - W)^T(I-W)$ and find $d+1$ smallest eigenvectors of B, $v_0, v_1, \ldots, v_d$ associated smallest eigenvalues $\lambda_0, \ldots, \lambda_d$. Drop the smallest eigenvector which is the constant vector explaining the degree of freedom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

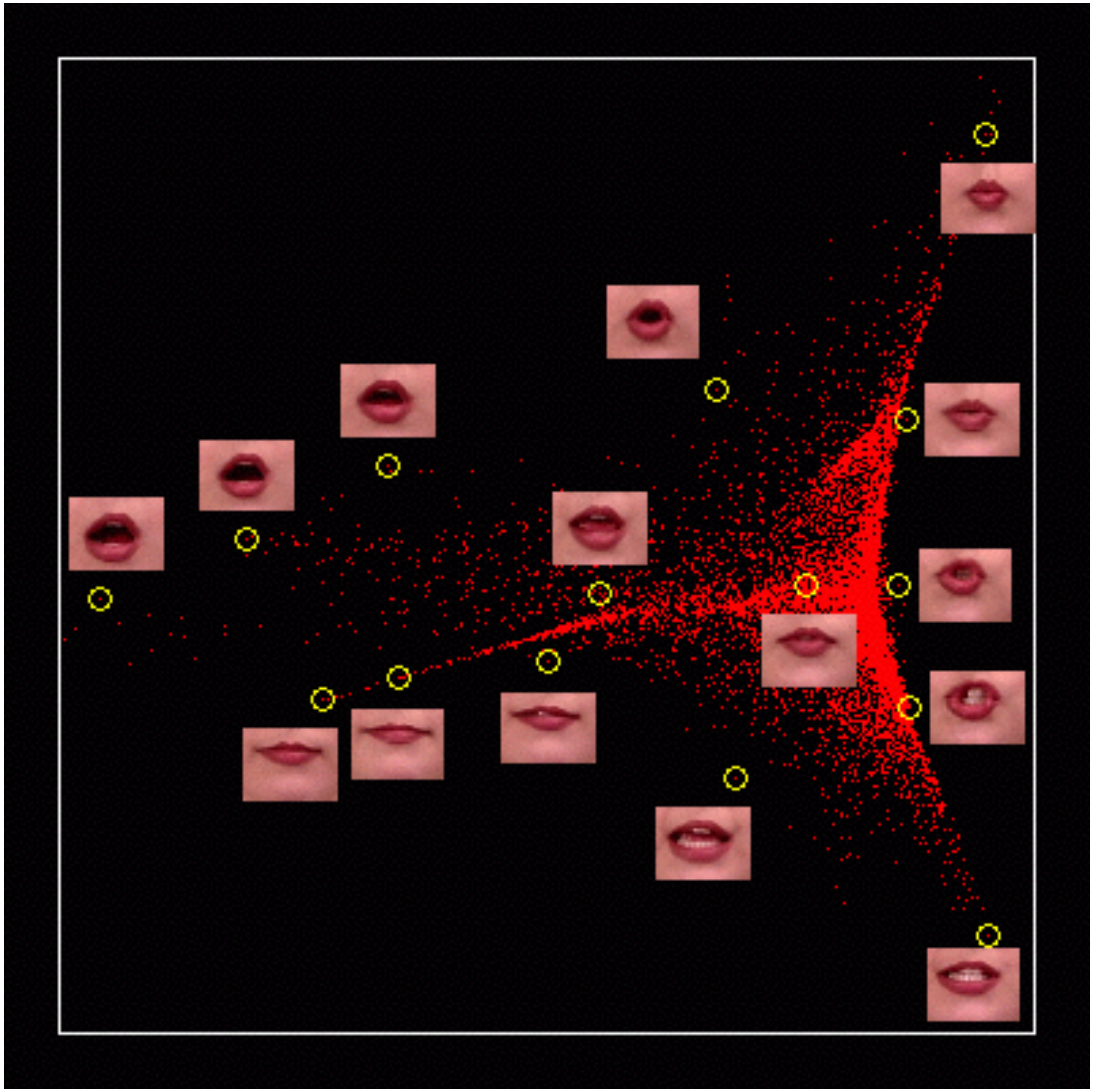
Remarks on LLE

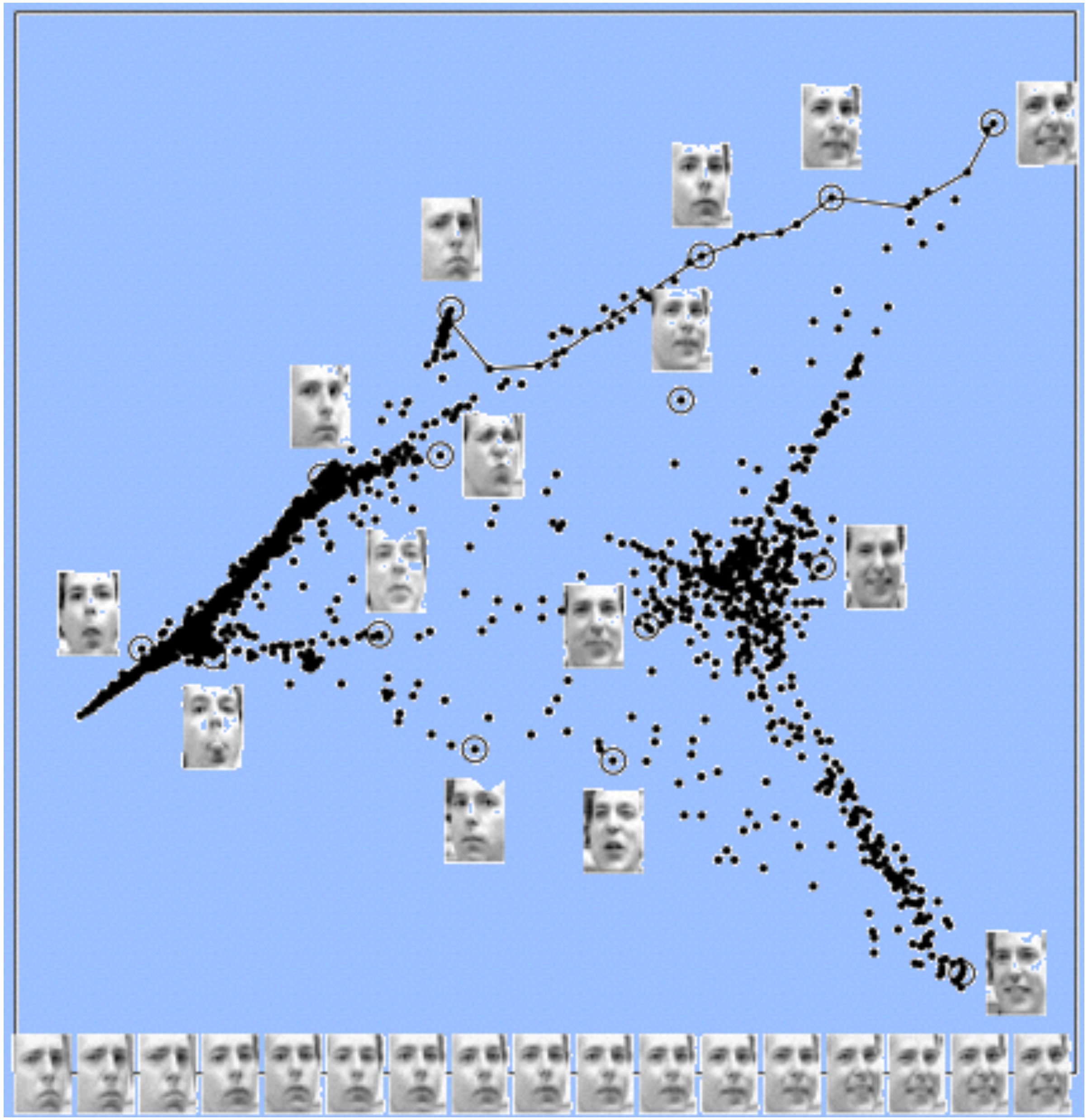
- Searching k-nearest neighbors is of O(kN)

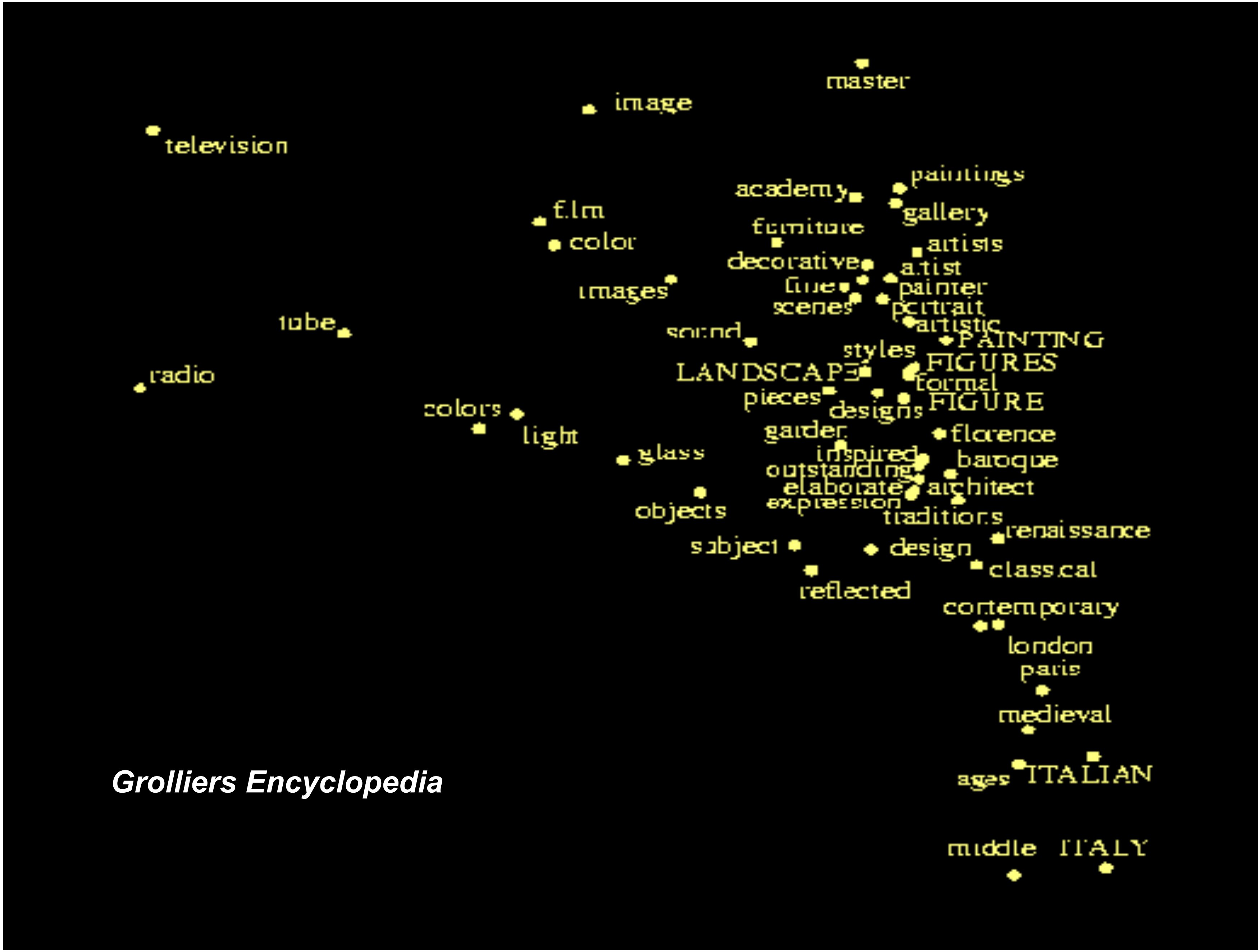
- W is sparse, kN/N^2=k/N nozeros
- W might be negative, additional nonnegative constraint can be imposed
- $B=(I-W)^T(I-W)$ is positive semi-definite (p.s.d.)

- Open Problem: exact reconstruction condition?









Grolliers Encyclopedia

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 Landmark-ISOMAP)	Local approach O(N^2)
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

Acknowledgement

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