ISOMAP and LLE
Matlab Dimensionality Reduction Toolbox

- [http://homepage.tudelft.nl/19j49/ Matlab_Toolbox_for_Dimensionality_Reduction.html](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.)
    
    - $B_{ij}(a) = -0.5 \, H(a) \, D \, H'(a)$, Hölder matrix $H(a) = I-1a'$
    
    - $a = 1_k$: $B_{ij} = -0.5 \, (D_{ij} - D_{ik} - D_{jk})$
    
    - $a = 1/n$: $B_{ij} = \frac{-1}{2} \left( D_{ij} - \frac{1}{N} \sum_{s=1}^{N} D_{sj} - \frac{1}{N} \sum_{t=1}^{N} D_{it} + \frac{1}{N^2} \sum_{s,t=1}^{N} D_{st} \right)$

- Eigendecomposition of $B = YY^T$

If we preserve the pairwise Euclidean distances do we preserve the structure??
Nonlinear Manifolds..

PCA and MDS see the Euclidean distance

What is important is the geodesic distance

Unfold the manifold
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:** \( \mathcal{M} \rightarrow \{1, \ldots, k\} \)  
  connected components, min cut

- **Classification/Regression:** \( \mathcal{M} \rightarrow \{-1, +1\} \) or \( \mathcal{M} \rightarrow \mathbb{R} \)  
  \( P \) on \( \mathcal{M} \times \{-1, +1\} \) or \( P \) on \( \mathcal{M} \times \mathbb{R} \)

- **Dimensionality Reduction:** \( f : \mathcal{M} \rightarrow \mathbb{R}^n \)  
  \( n \ll N \)

- **\( \mathcal{M} \) 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

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

$T_p \mathcal{M}^k \subset \mathbb{R}^N$
Tangent Vectors and Curves

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

\[ \frac{d\phi(t)}{dt} \bigg|_0 = V \]

Tangent vectors \(\rightarrow\) curves.
Riemannian Geometry

Norms and angles in tangent space.

\[ \langle v, w \rangle, \|v\|, \|w\| \]
**Geodesics**

Can measure length using **norm** in tangent space.

**Geodesic** — shortest curve between two points.

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

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

Tangent vectors $\mathbf{v}$

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

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

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

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

Tangent vectors $\longleftrightarrow$ 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 \( \phi(t) \)

\[ \phi(0) = p, \quad \phi(\|v\|) = q \quad \frac{d\phi(t)}{dt} \bigg|_0 = v \]
Laplacian-Beltrami Operator

\[ \Delta_M f(p) \equiv \sum_i \frac{\partial^2 f(\exp_p(x))}{\partial x_i^2} \]

Orthonormal coordinate system.

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

\[ \exp_p : T_p \mathcal{M}^k \rightarrow \mathcal{M}^k \]
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
Residual Variance vs. Intrinsic Dimension

- **Face Images**
- **SwisRoll**
- **Hand Images**

Graph A shows the residual variance for Face Images, Graph B for SwisRoll, Graph C for Hand Images, and Graph D with a label '2'.
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 $\gamma$ 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 $d_G \approx d_M$.  

14年10月9日星期四
Main Theorem

[ Bernstein, de Silva, Langford, and ]

Theorem 1: Let \( M \) be a compact submanifold of \( \mathbb{R}^n \) 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 \( \{x_i\} \) satisfies a \( \delta\)-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 \( \gamma \) 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 $\delta$-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 $\alpha$ sufficiently large:

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

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

• One need to compute pairwise shortest path between all sample pairs \((i,j)\)
  – Global
  – Non-sparse
  – Cubic complexity \(O(N^3)\)
Locally Linear Embedding

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

\[ \min_W \| X_i - \sum_{j=1}^{K} W_{ij} X_j \|^2 \] (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 (I)

(1) Construct a neighborhood graph

(2) Local fitting:
   - Pick up a point \( x_i \) and its neighbors \( N_i \)
   - Compute the local fitting weights
     \[
     \min_{\sum_{j \in N_i} w_{ij} = 1} \| x_i - \sum_{j \in N_i} w_{ij} (x_j - x_i) \|^2.
     \]
   This can be done by Lagrange multiplier method, \textit{i.e.} solving
     \[
     \min_{w_{ij}} \frac{1}{2} \| x_i - \sum_{j \in N_i} w_{ij} (x_j - x_i) \|^2 + \lambda (1 - \sum_{j \in N_i} w_{ij}).
     \]

   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 \( C^{(i)}_{jk} = \langle x_j - x_i, x_k - x_i \rangle \), whence the weights are
     \[
     w_i = C_i^\dagger (\bar{X}_i^T x_i + \lambda 1),
     \]
   where the Lagrange multiplier equals to
     \[
     \lambda = \frac{1}{1^T C_i^\dagger 1} \left( 1 - 1^T C_i^\dagger \bar{X}_i^T x_i \right),
     \]
   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 \( \mu > 0 \).
(3) Global alignment

Define a \( n \)-by-\( n \) weight matrix \( W \):

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

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?
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 $n \ll N$ landmark points from $\{x_i\}$ and compute a $n \times N$ 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
## Summary

<table>
<thead>
<tr>
<th>ISOMAP</th>
<th>LLE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do MDS on the geodesic distance matrix.</td>
<td>Model local neighborhoods as linear a patches and then embed in a lower dimensional manifold.</td>
</tr>
<tr>
<td>Global approach (O(N^3, \text{but L-ISOMAP}))</td>
<td>Local approach (O(N^2))</td>
</tr>
<tr>
<td>Might not work for nonconvex manifolds with holes</td>
<td>Nonconvex manifolds with holes</td>
</tr>
<tr>
<td>Extensions: Landmark, Conformal &amp; Isometric ISOMAP</td>
<td>Extensions: Hessian LLE, Laplacian Eigenmaps etc.</td>
</tr>
</tbody>
</table>

Both needs manifold finely sampled.
Reference

Acknowledgement

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