Geometric and Graph methods for High-Dimensional Data

Mauro Maggioni

Departments of Mathematics and Applied Mathematics, Johns Hopkins University
1. Diffusion Geometry
2. Multiscale Methods for Data
3. Multiscale Analysis on & of Graphs
4. Model Reduction for Stochastic Systems
1. Diffusion Geometry

Graph and manifold models for high-dimensional data
Data Sets in High-Dimensions

Large data sets arise in wide variety of applications:
- signal processing (images, biomedical data, audio, sonar, . . . )
- text documents, web pages
- network traffic and events, transactions, recommendation systems
- high-dimensional dynamical systems (e.g. fluid flows, biomolecules, . . . )
- biological systems (reaction networks, gene array data . . .)

Often we may model data as samples $\{x_i\}_{i=1}^n$ from a probability distribution $\mu$ in $\mathbb{R}^D$. 
Curse of dimensionality

Data as samples $\{x_i\}_{i=1}^n$ from a probability distribution $\mu$ in $\mathbb{R}^D$

In 1 dimension, this would correspond to have a histogram (possibly “continuous”) where the height of a column at a location is the probability of seeing a point at that location.

To estimate this histogram with accuracy $\epsilon$, under reasonable conditions we need bins of width $\epsilon$ and at least constant number of points in each bin, for a total of $O(\epsilon^{-1})$ points.

Unfortunately in $D$ dimensions, there are $O(\epsilon^{-D})$ boxes of size $\epsilon$. So we need $O(\epsilon^{-D})$ points. This is way too many: for $\epsilon = 10^{-1}$ and $D = 100$, we would need $10^{100}$ points.

Can we reduce the dimensionality?
Intrinsically low-dimensional Geometric “Models” for data
Principal Component Analysis

Curse of dimensionality: sample size for important statistical tasks scales exponentially in the ambient dimension. Hope: data concentrates near intrinsically low-dimensional sets.

Represent data as a matrix $X \in \mathbb{R}^{D \times n}$, one sample per column. The Singular Value Decomposition $X = U \Sigma V^T$ reveals if the data is concentrated near a $k$-dimensional plane. It is the case iff $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \gg \sigma_{k+1} \geq \ldots$. The first $k$ columns of $U$ span the desired $k$-plane.

$U$: orthogonal $D \times D$

$\Sigma$: diagonal $D \times n$

Diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$

are called singular values.

$V$: orthogonal $n \times n$

system of coordinates for features

system of coordinates for points

1901, K. Pearson
Perform dimension reduction by projecting onto top k singular vectors, and then perform learning tasks in the reduced space.
PCA and intrinsic dimension

Blobs in random position

PCA: Fourier
Singular values do not decay rapidly; many singular vectors are needed to achieve good accuracy.

However:
the intrinsic dimension of the data is 2.
Beyond PCA

Data on low-dimensional plane (Pearson, 1901)

Data on multiple low-dimensional planes

Data on low-dimensional curved structure
Principal Component Analysis

We need to reduce the dimensionality!

Hope: \( \mu \) supported on (or near) low-dimensional sets, such as hyperplanes, or manifolds, or unions thereof.

Represent data as a matrix \( X \in \mathbb{R}^{D \times n} \), one sample per column. If data lies on a \( k \)-dimensional hyperplane, compute the singular value decomposition \( X = U\Sigma V^T \): then only \( k \) singular values are nonzero, and the first \( k \) columns of \( U \) span the desired hyperplane.

How to parametrize?
Will use graphs and local connectivities (nonlinear construction!)
Eigen-analysis of graphs then very useful (back to linear!)
Suppose $\mu$ supported on (or near) low-dimensional sets, such as hyperplanes, or manifolds, or unions thereof. We observe noisy samples $x_i + \eta_i$, where $x_i \sim_{i.i.d.} \mu$ and the $\eta_i$'s model noise.

How to estimate the intrinsic low-dimension?

Lots of work on this problem (physics, statistics, machine learning), albeit without noise. Mostly volume arguments: number of samples in $B_z(r)$ should scale like $r^k$ if the data is $k$-dimensional.

The problem is multiscale: same data may have different dimensions at different scales (and locations).

Noise renders the problem meaningless at small scales, and so does curvature at large scales; finally, sampling makes the problem ill-posed at small scales.

We use the behavior of the singular values in the SVD of the samples in $B_z(r)$ as a function of $r$. 

Intrinsic dimension
Model: data \( \{x_i\}_{i=1}^n \) is sampled from a manifold \( \mathcal{M} \) of dimension \( k \), embedded in \( \mathbb{R}^D \), with \( k \ll D \). We receive \( \tilde{X}_n := \{x_i + \eta_i\}_{i=1}^n \), where \( \eta_i \sim_{\text{i.i.d}} N \) is \( D \)-dimensional noise (e.g. Gaussian). **Objective:** estimate \( k \).

Consider \( \sigma_i^{\tilde{z},r} = \lambda_i(\text{cov}(\tilde{X}_n \cap B_{\tilde{z}}(r))) \) for all \( r > 0 \), and all \( i = 1, \ldots, D \).

I.e.: for every \( z \) and every \( r > 0 \) consider the points in \( B_z(r) \) and look at the singular values of their covariance matrix:

Multiscale Singular Values

Their behavior as a function of \( r \) contains a lot of information about the geometry of \( \mathcal{M} \).
Multiscale SVD

Model: data $\{x_i\}_{i=1}^n$ is sampled from a manifold $\mathcal{M}$ of dimension $k$, embedded in $\mathbb{R}^D$, with $k \ll D$. We receive $\tilde{X}_n := \{x_i + \eta_i\}_{i=1}^n$, where $\eta_i \sim_{i.i.d} N$ is $D$-dimensional noise (e.g. Gaussian). Objective: estimate $k$.

Consider $\sigma_{i,j}^{\tilde{z},r} = \lambda_i(\text{cov}(\tilde{X}_n \cap B_{\tilde{z}}(r)))$
for all $r > 0$, and all $i = 1, \ldots, D$.

Green: where data is
Red: where noisy data is
Blue: volume in ball

In a ball of radius $r$, we show we only need a number of samples linear in the intrinsic dimension. Use concentration of measure phenomena and random matrix theory.
Multiscale SVD: sphere

Example: consider \( S^9(100,1000,0) \): 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with no noise.
Multiscale SVD: sphere+noise

Example: consider $\mathbb{S}^9(100,1000,0.1)$: 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise $\mathcal{N}(0, 0.1I_{100})$. Observe that $\mathbb{E}[||\eta||^2] \sim 0.1^2 \cdot 100 = 1$. 

Empirical quantities concentrate rapidly around these expected ones.
Multiscale SVD: sphere+noise

Example: consider \( S^9(100, 1000, 0.1) \): 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise \( \mathcal{N}(0, 0.1I_{100}) \). Observe that \( \mathbb{E}[||\eta||^2] \sim 0.1^2 \cdot 100 = 1 \).
Blobs in random position

Data: blobs in random position

Multiscale SVD around a point. The intrinsic dimension (2) is clearly revealed.
Blobs in random position

Data: blobs in random position

Multiscale SVD around a point. The intrinsic dimension (2) is clearly revealed.

Exercise: reproduce this example, and then add noise. Explore different regimes for $n$, $D$, and level of noise.
A sketch of the guarantees

We seek finite sample results, with high probability: given $n, D$ and other parameters of the problem, promise that with probability at least $1 - e^{-ct}$ we return the correct answer.

We prove that: if

- $M$ has small “curvature” $\kappa$
- the noise $\eta$ has small std $\sigma$,

so that there is a range of good scales $r$ for which the manifold does look “long” in $k$ directions and “thin” in the others, then

With high probability that range of scales survives sampling and noise, and we are able to determine $k$ given only $n_r \gtrsim k \log k$ points in a ball of radius $r$, for $r$ a good scale.

Moreover, if $\kappa \sim k^{-1}$ and $\sigma \sim D^{-\frac{1}{2}}$, then our results are dimension-free, i.e. the above is true w.h.p. for $n \gtrsim k \log k$ with the implicit constants independent of $k, D$. 
Graphs from point clouds
Principal Component Analysis

We need to reduce the dimensionality!

Hope: \( \mu \) supported on (or near) low-dimensional sets, such as hyperplanes, or manifolds, or unions thereof.

Represent data as a matrix \( X \in \mathbb{R}^{D \times n} \), one sample per column. If data lies on a \( k \)-dimensional hyperplane, compute the singular value decomposition \( X = U \Sigma V^T \): then only \( k \) singular values are nonzero, and the first \( k \) columns of \( U \) span the desired hyperplane.

How to parametrize?
Will use graphs and local connectivities (nonlinear construction!)
Eigen-analysis of graphs then very useful (back to linear!)
Random walks on graphs and data

We may construct weighted graphs from data: given

. **Data** $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$.

. **Local similarities** via a kernel function $W(x_i, x_j) \geq 0$.

Simplest example: $W_\sigma(x_i, x_j) = e^{-\|x_i - x_j\|^2 / \sigma}$. 
Random walks on graphs and data

We may construct weighted graphs from data: given

. Data $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$.

. Local similarities via a kernel function $W(x_i, x_j) \geq 0$.

Simplest example: $W_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.

Given a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := W(x_i, x_j)$, when positive (or above a threshold). Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W, \quad P_{ij} = \text{prob. of jumping } x_i \rightarrow x_j$$

Let $1 = \lambda_0 \geq \lambda_1 \geq \ldots$ and $\varphi_i$ the eigenval.’s and eigenvect.’s of $P$, i.e. $P\varphi_i = \lambda_i \varphi_i$. We consider the map

$$G \ni x \mapsto (\varphi_1(x), \ldots, \varphi_m(x)) \in \mathbb{R}^m$$
Random walks on graphs and data

Given a **weighted graph** \((G, E, W)\): vertices represent data points, edges connect \(x_i, x_j\) with weight \(W_{ij} := W(x_i, x_j)\), when positive (or above a threshold). Let \(D_{ii} = \sum_j W_{ij}\) and

\[
P = D^{-1}W, \quad P_{ij} = \text{prob. of jumping } x_i \rightarrow x_j
\]

Let \(1 = \lambda_0 \geq \lambda_1 \geq \ldots\) and \(\varphi_i\) the eigenval.’s and eigenvec.’s of \(P\), i.e. \(P\varphi_i = \lambda_i \varphi_i\). We consider the map

\[
G \ni x \mapsto (\varphi_1(x), \ldots, \varphi_m(x)) \in \mathbb{R}^m
\]

With suitable modifications, one can show that as \(n \rightarrow +\infty\), and if the samples \(\{x_i\}\) are drawn according to the stationary distribution of the process, \(P\) converges to the propagator of the Fokker-Planck equation.
Diffusion Distances

In some cases the geodesic distances $d_M$ may not capture interesting geometric information. For example here $d_M(A, B) \sim d_M(B, C)$. Can we define a new distance that may capture this type of geometric characteristic?
Diffusion distance

We may use random walks, and, for $t > 0$, define

$$d^{(t)}(x, y) = \| T^t(x, \cdot) - T^t(y, \cdot) \|_{L^2(G)} = \sqrt{\sum_{z \in G} |T^t(x, z) - T^t(y, z)|^2}$$

$$T \varphi_i = \lambda_i \varphi_i \implies \sqrt{\sum_{i=1}^{+\infty} \lambda_i^{2t}(\varphi_i(x) - \varphi_i(y))^2}$$

$$\sim \left\| \left( \lambda_i^t \varphi_i(x) \right)_{i=1}^m - \left( \lambda_i^t \varphi_i(y) \right)_{i=1}^m \right\|_{\mathbb{R}^m}$$

Therefore $\Phi_m^{(t)}$ defined by $\Phi_m^{(t)}(x) = (\lambda_i^t \varphi_i(x))_{i=1}^m$ satisfies

$$\left\| \Phi_m^{(t)}(x) - \Phi_m^{(t)}(y) \right\|_{\mathbb{R}^m} \sim d^{(t)}(x, y),$$

at least for $t$ large and $m$ large.

Exercise: derive the above
In some cases the geodesic distance $d_G$ may not capture geometric information of interest. For example here $d_G(A, B) \sim d_G(B, C)$. The diffusion distance is sensitive to connectivity between clusters.
Spectral Clustering in one slide

\[ x \mapsto (\phi_2(x), \phi_3(x)) \]

Original

Every point is connected to its 5 nearest neighbors, obtaining a graph.

Flexibility + robustness

Diffusion space

\[ \phi_2 = 0 \text{ is a good cut!} \]

\[
\langle Lf, f \rangle = \sum_x \sum_{y \sim x} W(x, y) \left( \frac{f(x)}{\sqrt{d_x}} - \frac{f(y)}{\sqrt{d_y}} \right)^2
\]
Data, graphs and random walks

Simple examples in Matlab

Exercise: download the code, and the Diffusion Geometry toolbox, and go through these examples, and more, that are part of the toolbox
# Data Sets in High-Dimensions

A *deluge of data*: documents, customer databases, images, social network transactions, gene arrays, sensor networks, financial transactions...

*Data set*: often $X \subset \mathbb{R}^D$, $D$ very large ($10^2 - 10^8$).

<table>
<thead>
<tr>
<th>Data</th>
<th>Representation</th>
<th>Problems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approx. 1000 articles from ScienceNews. Representation: a document-term matrix, with about 1000 terms.</td>
<td><img src="image1" alt="Matrix representation" /></td>
<td>Automatically sort articles into categories, given only a small set labeled by experts. Navigate the library.</td>
</tr>
<tr>
<td>A database of ~60000 grayscale 28x28 images of handwritten digits 0-9.</td>
<td><img src="image2" alt="Digit images" /></td>
<td>Automatically recognize digits (e.g. for ZIP codes, checks, etc...)</td>
</tr>
</tbody>
</table>
Example: Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, $i$-th coordinate of document $d$ represents frequency in document $d$ of the $i$-th word in a dictionary. Point cloud of 1100 points in $\mathbb{R}^{1000}$. 
Example: Handwritten digits

Database of 60,000 pictures, with $28 \times 28$ pixels, of handwritten digits collected by USPS. Point cloud of 60,000 points in $\mathbb{R}^{728}$. 
Reduced models for Molecular Dynamics
The dynamics of a small peptide (12 atoms with $H$-atoms removed) in a bath of water molecules, is approximated by a Langevin system of stochastic equations

$$\dot{x} = -\nabla U(x) + \dot{w}$$

The set of configurations is a point cloud in $\mathbb{R}^{12 \times 3}$. 

Joint with C. Clementi, M. Rohrdanz, W. Zheng
Example: Molecular Dynamics Data

The dynamics of a small peptide (12 atoms with $H$-atoms removed) in a bath of water molecules, is approximated by a Langevin system of stochastic equations

$$\dot{x} = -\nabla U(x) + \sqrt{\frac{2}{\beta}} \dot{w}$$

$$\beta = \frac{1}{k_B T}$$

The set of configurations is a point cloud in $\mathbb{R}^{12 \times 3}$. 
Objectives

Complex, expensive, high-d simulator, at very fine (time)scale

Learning algorithm: allowed to run for a very long time, then somehow construct and output

Simple description of the system
Example: Alanine dipeptide


Free energy in terms of empirical coordinates

MSVD near transition state

MSVD near free energy minimum

Low intrinsic dimension: this settles certain questions/conjectures/intuitions by chemical physists; counting degrees of freedom/quality of approximation by low-dimensional representations

Different local intrinsic dimensions: in different regions of state space

Different local scales: quantifies intuitions about variability of entropy in state space

New very robust dimension reduction techniques: Locally Scaled Diffusion Maps that take this information into account; novel ways of approximating these systems by low-dimensional representations
Under suitable conditions, $H_{FP}$ has discrete spectrum $0 = \lambda_0 < \lambda_1 \leq \ldots \lambda_k \ll \lambda_{k+1} \leq \ldots$, and fundamental solution with eigen-expansion

$$p_t(x, y) = \phi_0(x) + \sum_{j=1}^{+\infty} \psi_j(y)\phi_j(x)e^{-\lambda_j t}.$$ 

The dual system of eigenfunctions, which we pick as reaction coordinates, is

$$\psi_j(x) = \frac{\phi_j(x)}{\phi_0(x)}.$$
Molecular Dynamics & F-P. equation

Fokker-Planck equation & eigenfunctions $p_t(x, y) = \text{(density of) probability of molecule at } y \text{ at time } t$, having started at $x$

$$\frac{\partial p}{\partial t} = \frac{1}{\beta} \Delta p + \nabla(p \nabla U) = \mathbf{H}_{FP} p$$

$\beta = 1/(k_BT)$, $k_B$ is Boltzmann’s constant

Under suitable conditions, $\mathbf{H}_{FP}$ has discrete spectrum $0 = \lambda_0 < \lambda_1 \leq \ldots \lambda_k \ll \lambda_{k+1} \leq \ldots$, and fundamental solution with eigen-expansion

$$p_t(x, y) = \phi_0(x) + \sum_{j=1}^{+\infty} \psi_j(y) \phi_j(x) e^{-\lambda_j t}.$$ 

The dual system of eigenfunctions, which we pick as reaction coordinates, is

$$\psi_j(x) = \phi_j(x)/\phi_0(x).$$

**Diffusion Distance at time $t$**

$$d^t(x, y) = \|p_t(x, \cdot) - p_t(y, \cdot)\|_{L^2} = \sqrt{\sum_j e^{-\lambda_j t} |\psi_j(x) - \psi_j(y)|^2}$$
Diffusions maps for MD

R.R. Coifman, I.G. Kevrekidis, S. Lafon, MM, B. Nadler

\[ \dot{x} = -\nabla U(x) + \sqrt{\frac{2}{\beta}} \dot{w}, \quad x \in \mathbb{R}^{3N} \]

\[ \frac{\partial p}{\partial t} = \mathcal{L}p = \frac{1}{\beta} \Delta p + \nabla \cdot (p \nabla U) = -\mathbf{H}_{\text{FP}} p \]

\[ \mathbf{H}_{\text{FP}} \varphi_j = \lambda_j \varphi_j \quad \mathbf{H}_{\text{FP}}^* \psi_j = \lambda_j \psi_j \]

**Theorem.** Define the optimal $k$-dimensional approximation $p_{t}^{(k)}(x, y)$ of $p_t(x, y)$ by

\[ p_{t}^{(k)}(x, y) = \underset{\tilde{p}}{\arg \min} \sum_{y} \left[ \| p_t(x, y) - \tilde{p}_t(x, y) \|_2^2 \right] \]

where the argmin is over $\tilde{p} := \Pi p$, with $\Pi$ a rank $k$ orthogonal projector, $e^{-U/Z}$ is the stationary distribution. Then

\[ p_{t}^{(k)}(x, y) = \sum_{j=0}^{k-1} e^{-\lambda_j t} \psi_j(y) \varphi_j(x). \]

Empirical approximations, given samples distributed according to $\varphi_0$, can be obtained by suitably modifying diffusion maps (with guarantees).
The dynamics of a small peptide (12 atoms with $H$-atoms removed) in a bath of water molecules, is approximated by a Langevin system of stochastic equations

$$\dot{x} = -\nabla U(x) + \omega$$

The set of configurations is a point cloud in $\mathbb{R}^{12\times3}$. 

Example: Molecular Dynamics Data

![Molecular Dynamics Data](image)
Molecular Dynamics data for alanine

Given only trajectory data, we construct an empirical approximation to the generator of the Fokker-Planck, and compute its eigenvalues/vectors to obtain a low-dimensional embedding and reaction coordinates.

$$\frac{\partial p}{\partial t} = -\sum_i \frac{\partial}{\partial x_i} \left( \frac{1}{\beta} \frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i} \right) p = -H_{FP}p$$


Coordinate | $C_5$, $P_i$ | $\alpha_R$ | $\omega_P$ | $\alpha_R$ | $\omega_P$ | $C_5$, $P_i$
---|---|---|---|---|---|---
From simulation$^b$ | 0.023 | 0.047 | | | | |
$1^{st}$ DC | 0.023 ± 0.001 | 0.048 ± 0.003 | | | | |
$\Psi$ | 0.020 ± 0.001 | 0.040 ± 0.003 | | | | |

rate = $\left( \int_{\text{barrier}} ^{\text{well}} \frac{e^{\beta F(x)}}{D(x)} \, dx \int_{\text{well}} ^{\text{barrier}} e^{-\beta F(x')} \, dx' \right)^{-1}$

G. Hummer, 2005
We may plot the diffusion coordinates as functions of the physical observables given by the angles and notice they are essentially in one-to-one correspondence, with the diffusion coordinates emphasizing energy barriers separating minima.
∃ catch...
Random walks on graphs and data

We may construct weighted graphs from data: given

. Data $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$.

. Local similarities via a kernel function $W(x_i, x_j) \geq 0$.

Simplest example: $W_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$

Given a weighted graph $(G, E, W)$: vertices represent data points, edges connect $x_i, x_j$ with weight $W_{ij} := W(x_i, x_j)$, when positive (or above a threshold). Let $D_{ii} = \sum_j W_{ij}$ and

$$P = D^{-1}W, \quad P_{ij} = \text{prob. of jumping } x_i \to x_j$$

Let $1 = \lambda_0 \geq \lambda_1 \geq \ldots$ and $\varphi_i$ the eigenvalues’ and eigenvectors’ of $P$, i.e. $P\varphi_i = \lambda_i \varphi_i$. We consider the map

$$G \ni x \mapsto (\varphi_1(x), \ldots, \varphi_m(x)) \in \mathbb{R}^m$$
Example: Alanine dipeptide

Locally Scaled Diffusion Map

Construct the $N \times N$ matrix of transition probability kernels $K$, as

$$K_{ij} = e^{-\frac{d_{\text{RMSD}}(x_i,x_j)^2}{2\epsilon_i \epsilon_j}},$$

for $x_i$ and $x_j$ molecular configurations, $\epsilon_i$ and $\epsilon_j$ their local scales.

For each $x_i$, compute

$$P_i = \sum_{j=1}^{N} K_{ij},$$

which is proportional to a density estimation around $x_i$.

Normalize the kernel as

$$\tilde{K}_{ij} = P_i^{-\frac{1}{2}} K_{ij} P_j^{-\frac{1}{2}}.$$

Define the diagonal matrix $D$ as $D_i = \sum_{j=1}^{N} \tilde{K}_{ij}$, and construct a Markov matrix $M = D^{-1} \tilde{K}$,

$$M_{ij} = D_i^{-1} \tilde{K}_{ij}.$$

Compute largest eigenvalues and corresponding right eigenvectors of $M$. 

Example: SH-3

Free energy in terms of diffusion coordinates

Small polymer reversal rate

20 bead polymer confined in nanopore; the axis z of nanopore is the reaction coordinate. Previous work by Huang and Makarov used distance between bead ends, projected onto z, as empirical reaction coordinates, and noticed poor approximation of the transition rates.

Locally scaled diffusion maps yield a reaction coordinate along which a more accurate transition rate is estimated; this coordinate corresponds quite well with z, while the second diffusion coordinate is related to radius of gyration of the polymer.
Summary

“Models”: high-dimensional data may often be approximated by low-dimensional geometric objects

*Intrinsic dimension*: use Multiscale SVD to measure intrinsic dimension

*Coordinates*: global low-dimensional parametrizations via diffusion maps. This can be adapted to Langevin dynamics for low-dimensional descriptions of large-time dynamics.

The above requires long trajectories - typically expensive
Further connections: manifold learning semi supervised learning
Manifold Learning in 1 slide

Assume there is a domain \( \Omega \subset \mathbb{R}^k \) and a map \( \phi : \Omega \to \mathbb{R}^D \), and data is in the form \( x_i = \phi(p_i), \ p_i \in \Omega \). Given \( x_1, \ldots, x_n \), we are interested in finding the corresponding \( p_i \)'s, and learning the map \( \phi \) or \( \phi^{-1} \).

ISOMAP [Tenenbaum, DeSilva, Langford]: use Multi-Dimensional Scaling to map the \( x_i \)'s to Euclidean points having pairwise Euclidean distances close to the pairwise geodesic distances. Guaranteed to learn \( \phi \) when \( \Omega \) is convex, \( \phi \) is an isometry, and as \( n \to +\infty \).

Hessian eigenmap [Donoho & Grimes]: removes the requirement that \( \Omega \) is convex, but \( \phi \) still needs to be an isometry.

Diffusion Map/Laplacian Eigenmap [Coifman, Lafon, MM; Belkin, Niyogi]: use eigenfunctions of the Laplacian, connections with diffusion distances (rather than geodesic).
Heat and Eigenfunction maps

P.W. Jones, MM, R. Schul

One may use heat kernels or Laplacian eigenfunctions to obtain bi-Lipschitz maps from a large portion of a manifold $\mathcal{M}$ of dimension $d$, around any given point $z$, to Euclidean space $\mathbb{R}^d$. These maps are in the form

- $x \mapsto (K_t(x, x_i))_{i=1}^d$ where $K_t(\cdot, x_i)$ is the heat kernel on $\mathcal{M}$ centered at $x_i \in \mathcal{M}$ at time $t$. $t$ and the sources $x_i$ are easily depending only on $z$ and on the size of the portion to be mapped.

- $x \mapsto (\varphi_{j_i}(x))_{i=1}^d$, where the $\varphi_{j_i}$’s are eigenfunctions of the Laplacian $\Delta$ on $\mathcal{M}$, and the indices $j_i$ are well-chosen depending on the large region of $\mathcal{M}$ being mapped.

Note that these algorithms require $d$ as input.
Semi-Supervised Learning


Lots of work in the past 10 years in the machine learning, statistics, applied mathematics communities: manifold learning, dimensionality reduction, topological data analysis.
Function approximation on data

The geometry of the data may help the construction of useful priors and function spaces, for learning problems.

Approximation in high dimension requires large number of samples unless strong smoothness assumptions are made.

Another approach: use intrinsically low dimensional geometric structures.
Basic idea: map an image to a high-dimensional set, where each point is a $p \times p$ patch.

We can construct a GMRA dictionary for the image on the set of $p \times p$ patches, and then analyze the image accordingly. This has lead to state-of-art results in a variety of image processing tasks, such as denoising, in-painting, super-resolution, etc...
Map from $N \times N$ image $I$ to point cloud $X$ of $N^2 \ 7 \times 7$ patches. View $I$ as a function $X \rightarrow \mathbb{R}_+$ and denoise it by considering $P_X^t I$, with $P_X$ the random walk on $X$. 
Summary

- Curse of dimensionality
- Principal Component Analysis
- Multiscale Principal Component Analysis
- Intrinsic Dimensionality
- Construct graphs from point clouds by local connections
- Random walks, diffusion distances, spectral clustering
- Examples
- Connections with continuous limits
- Model reduction for molecular dynamics
Meet your TA

Wenjing Liao

Ph.D. in Mathematics under the supervision of Prof. Albert Fannjiang at University of California, Davis in 2013. At Duke ever since, and now at J.H.U.

Expertise:
- imaging and signal processing,
- applied harmonic analysis,
- machine learning and high dimensional data analysis.

TA Sessions:
- A mix of problems, discussions, numerical simulations.
- The main goal is to discuss ideas, and mathematical technique at a high level (a.k.a. “back-of-the-envelope” calculations).
References

- D. Donoho, *High-Dimensional Data Analysis: The curses and blessings of dimensionality*

- **Multiscale SVD**: *Multiscale Geometric Methods for Data Sets I: Multiscale SVD, Noise and Curvature*, A. V. Little, M. Maggioni, L. Rosasco. We use a lot of tools and results from R. Vershynin's *Introduction to non-asymptotic analysis of random matrices*


- **Diffusion geometry**: [http://www.math.jhu.edu/~mauro/diffusiongeometries.html](http://www.math.jhu.edu/~mauro/diffusiongeometries.html)
  - Articles using/citing diffusion geometry

- **Manifold learning**: See the seminal *Isomap* and *Local linear embedding* papers, with many others that followed. Belkin and Niyogi introduced Laplacian eigenmaps in a series of papers, for both dimensionality reduction and regression on manifolds, see e.g. *Laplacian eigenmaps for dimensionality reduction and data representation* or *Semisupervised learning on Riemannian manifolds*