A Geometric Perspective on Machine Learning

Partha Niyogi

The University of Chicago

Thanks: M. Belkin, A. Caponnetto, X. He, I. Matveeva, H. Narayanan, V. Sindhwani, S. Smale, S. Weinberger

High Dimensional Data

When can we avoid the curse of dimensionality?

- Smoothness

  rate $\approx (1/n)^{\frac{s}{d}}$

  splines,kernel methods, L_2 regularization...

- Sparsity

  wavelets, L_1 regularization, LASSO, compressed sensing..

- **Geometry**

  graphs, simplicial complexes, laplacians, diffusions

Geometry and Data: The Central Dogma

- Distribution of natural data is non-uniform and concentrates around low-dimensional structures.
- The shape (geometry) of the distribution can be exploited for efficient learning.

Manifold Learning

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

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

  connected components, min cut

- Classification: $\mathcal{M} \rightarrow \{-1, +1\}$

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

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

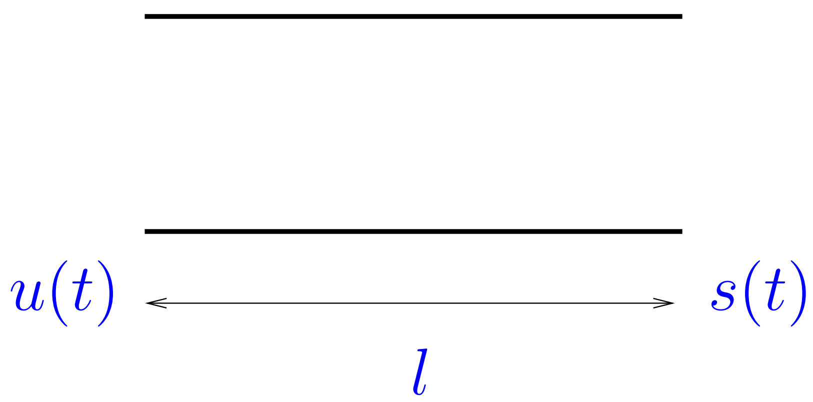
- $\mathcal{M}$ unknown: what can you learn about $\mathcal{M}$ from data?

  e.g. dimensionality, connected components

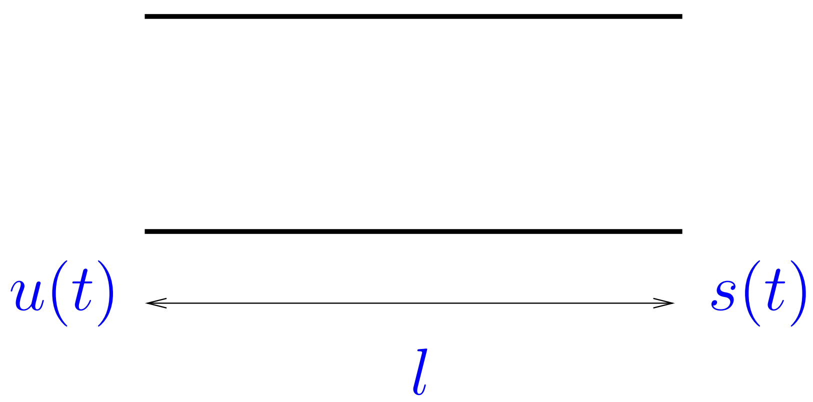
  holes, handles, homology

  curvature, geodesics

An Acoustic Example



An Acoustic Example

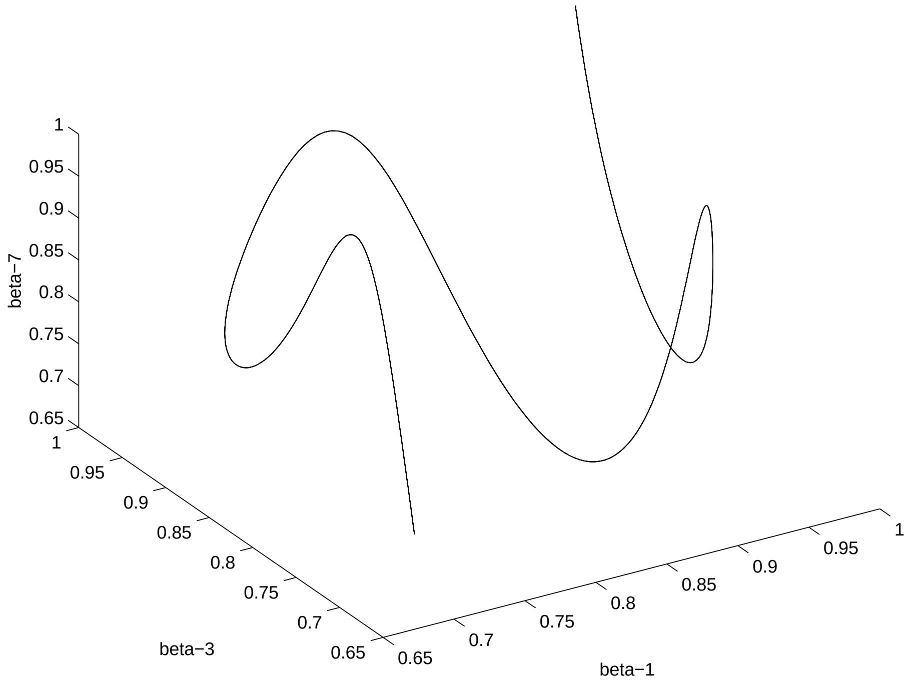


One Dimensional Air Flow

(i) $\frac{\partial V}{\partial x} = -\frac{A}{\rho c^2}\frac{\partial P}{\partial t}$ (ii) $\frac{\partial P}{\partial x} = -\frac{\rho}{A}\frac{\partial V}{\partial t}$

$V(x,t) =$ volume velocity
$P(x,t) =$ pressure

Solutions



$$u(t) = \sum_{n=1}^{\infty} \alpha_n \sin(n\omega_0 t) \in l_2$$

$$s(t) = \sum_{n=1}^{\infty} \beta_n \sin(n\omega_0 t) \in l_2$$

Formal Justification

- Speech

  speech $\in l_2$ generated by vocal tract

  Jansen and Niyogi (2005)

- Vision

  group actions on object leading to different images

  Donoho and Grimes (2004)

- Robotics

  configuration spaces in joint movements

- Graphics

Manifold + Noise may be generic model in high dimensions.

Take Home Message

- Geometrically motivated approach to learning

  nonlinear, nonparametric, high dimensions

- Emphasize the role of the Laplacian and Heat Kernel
  - Semi-supervised regression and classification
  - Clustering and Homology
  - Randomized Algorithms and Numerical Analysis

Pattern Recognition

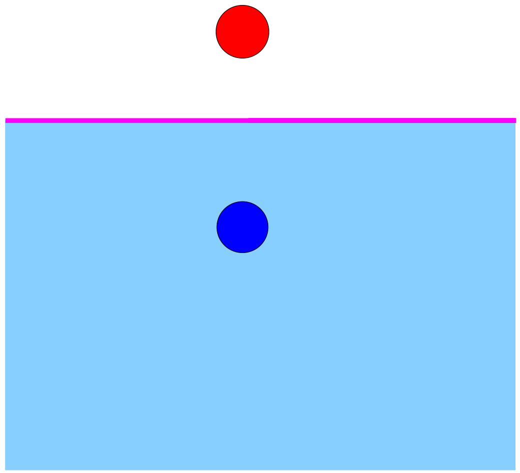
P on $X \times Y$ $\quad X = \mathbb{R}^N; Y = \{0, 1\}, \mathbb{R}$

(x_i, y_i) labeled examples

find $f : X \to Y$ $\quad$ Ill Posed

Simplicity

Simplicity



Regularization Principle

$$f = \arg\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

Splines

Ridge Regression

SVM

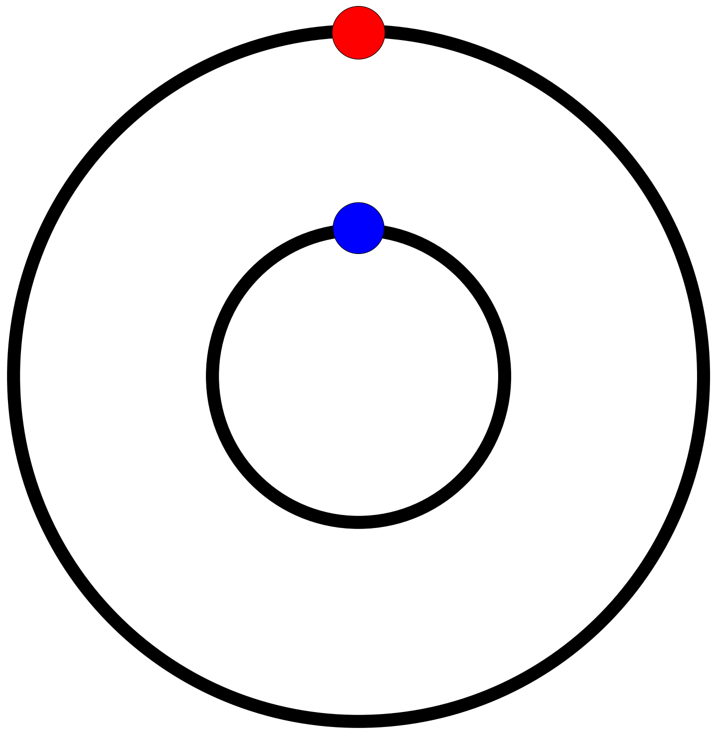
- $K : X \times X \to \mathbb{R}$ is a p.d. kernel

  e.g. $e^{-\frac{\|x-y\|^2}{\sigma^2}}$, $(1 + x \cdot y)^d$, etc.

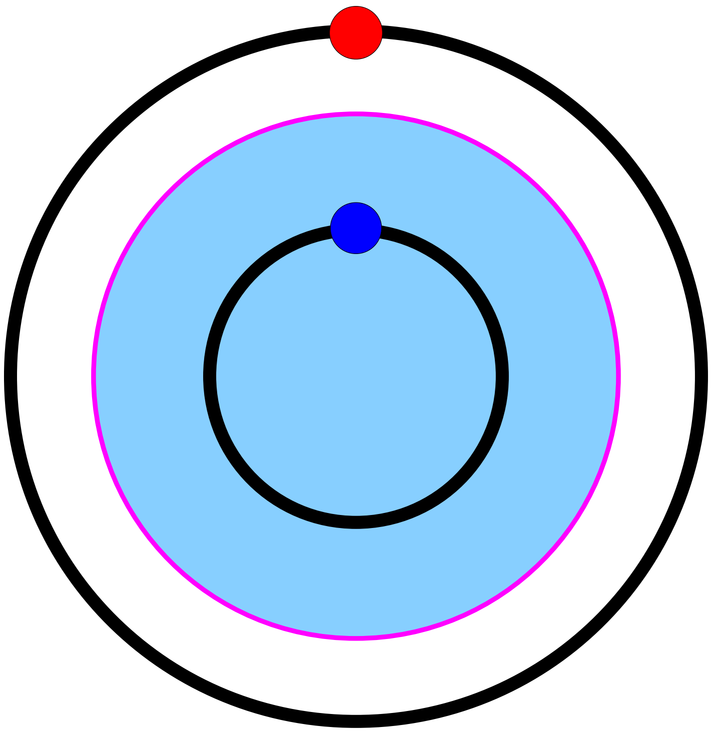
- H_K is a corresponding RKHS

  e.g., certain *Sobolev* spaces, polynomial families, etc.

Simplicity is Relative



Simplicity is Relative



Intuitions

- supp P_X has manifold structure
- *geodesic* distance v/s *ambient* distance
- geometric structure of data should be incorporated
- f versus $f_{\mathcal{M}}$

Manifold Regularization

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2$$

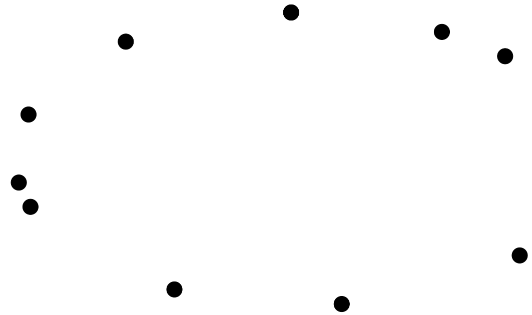
$$\|f\|_I^2 = \begin{cases} \text{Laplacian} & \int \langle \mathrm{grad}_{\mathcal{M}} f, \mathrm{grad}_{\mathcal{M}} f \rangle = \int f \Delta_{\mathcal{M}} f \\ \text{Iterated Laplacian} & \int f {\Delta_{\mathcal{M}}}^i f \\ \text{Heat kernel} & \mathrm{e}^{-\Delta_{\mathcal{M}} t} \\ \text{Differential Operator} & \int f(Df) \end{cases}$$

Representer Theorem: $f = \sum_{i=1}^{n} \alpha_i K(x, x_i) + \int_{\mathcal{M}} \alpha(y) K(x, y)$

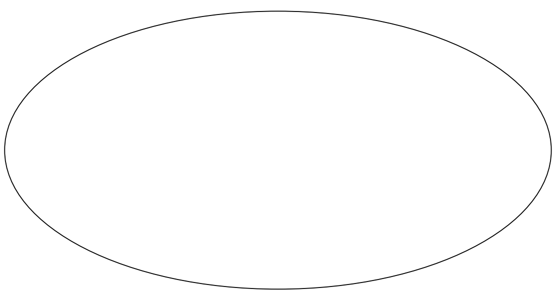
Belkin, Niyogi, Sindhwani (2004)

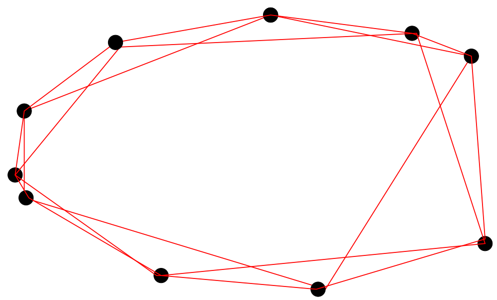
Approximating $\|f\|_I^2$

$\mathcal{M}$ is unknown but $x_1 \ldots x_M \in \mathcal{M}$



$$\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle \quad \approx \quad \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2$$





Manifolds and Graphs

$$\mathcal{M} \approx G = (V, E)$$

$e_{ij} \in E$ if $\|x_i - x_j\| < \epsilon$

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$\Delta_{\mathcal{M}} \approx L = D - W$$

$$\int \langle \text{grad} f, \text{grad} f \rangle \approx \sum_{i,j} W_{ij}(f(x_i) - f(x_j))^2$$

$$\int f(\Delta f) \approx \mathbf{f}^T L \mathbf{f}$$

Manifold Regularization

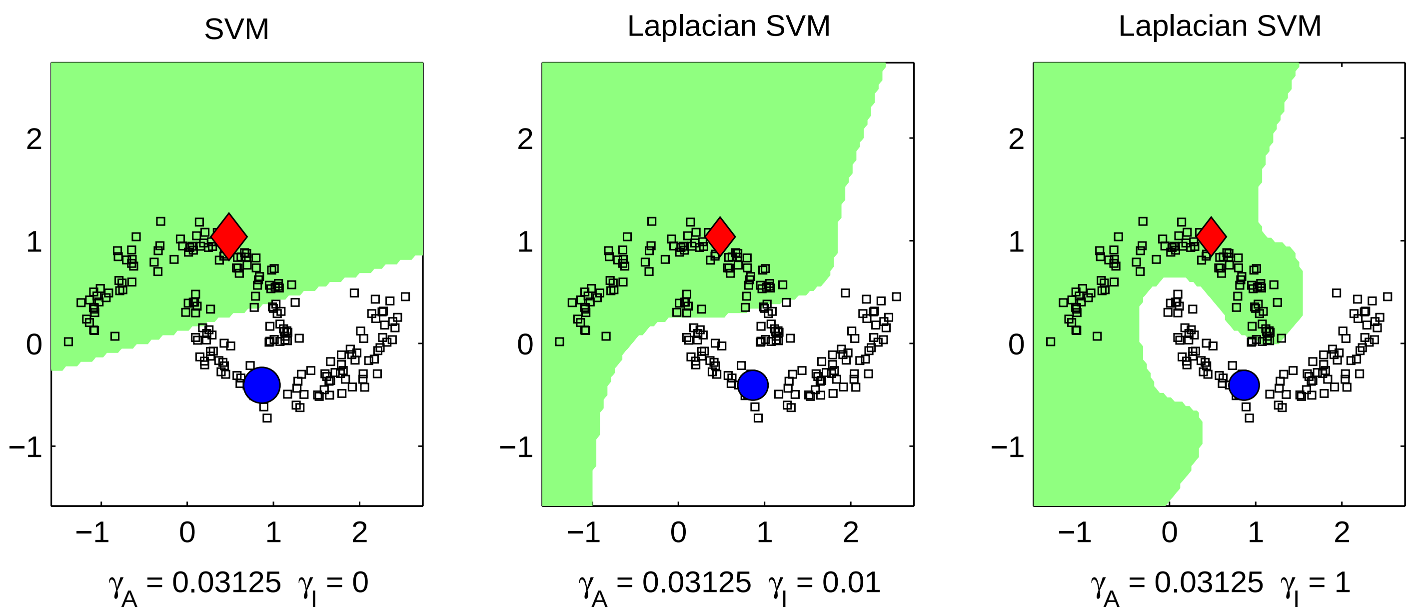
$$\frac{1}{n}\sum_{i=1}^{n} V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \gamma_I \sum_{i \sim j} W_{ij}(f(x_i) - f(x_j))^2$$

Representer Theorem: $f_{opt} = \sum_{i=1}^{n+m} \alpha_i K(x, x_i)$

$V(f(x), y) = (f(x) - y)^2$: Least squares

$V(f(x), y) = (1 - yf(x))_+$: Hinge loss (Support Vector Machines)

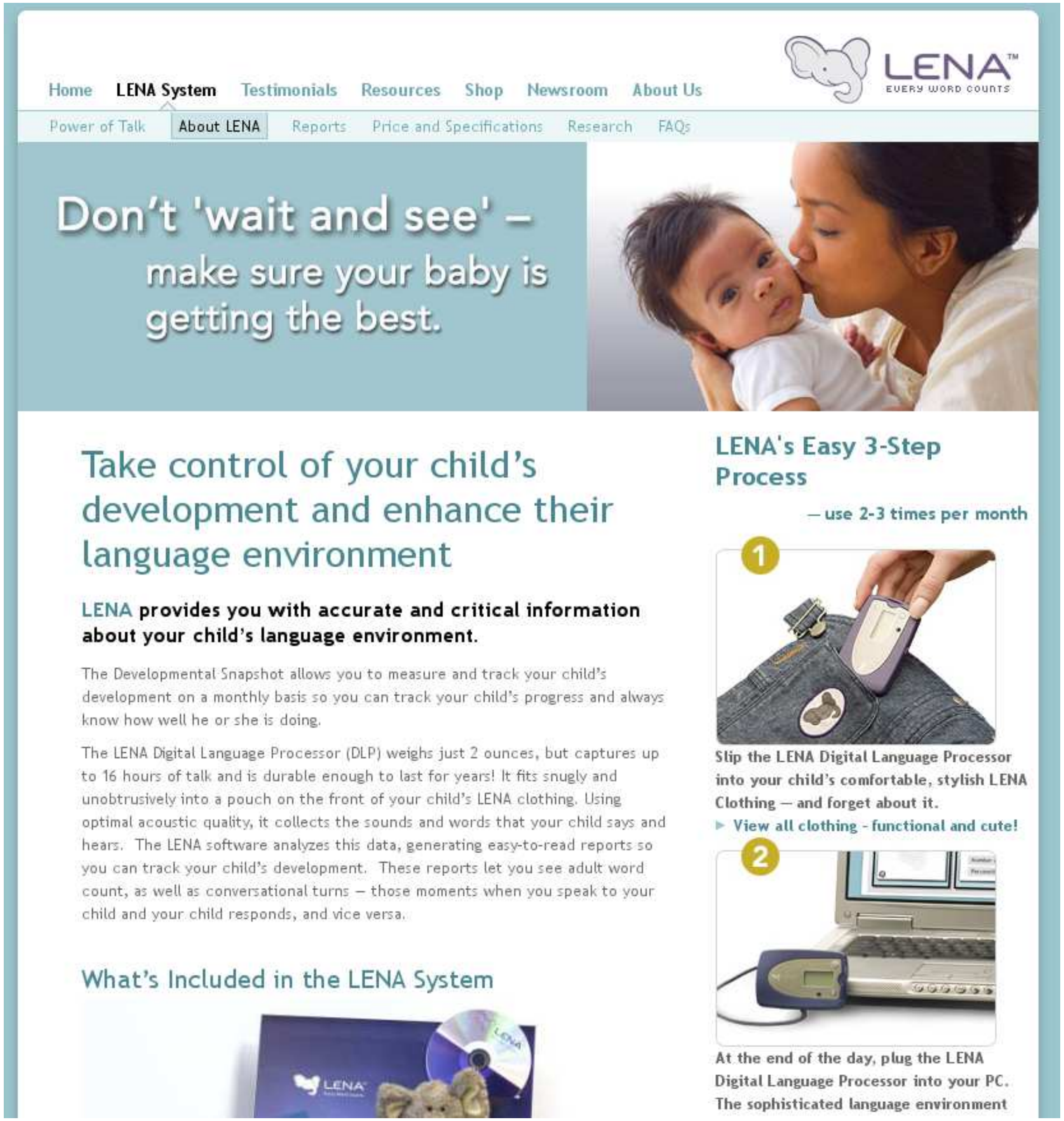
Ambient and Intrinsic Regularization



Experimental comparisons

Dataset → Algorithm ↓	g50c	Coil20	Uspst	mac-win	WebKB (link)	WebKB (page)	WebKB (page+link)
SVM (full labels)	3.82	0.0	3.35	2.32	6.3	6.5	1.0
RLS (full labels)	3.82	0.0	2.49	2.21	5.6	6.0	2.2
SVM (l labels)	8.32	24.64	23.18	18.87	25.6	22.2	15.6
RLS (l labels)	8.28	25.39	22.90	18.81	28.0	28.4	21.7
Graph-Reg	17.30	6.20	21.30	11.71	22.0	10.7	6.6
TSVM	6.87	26.26	26.46	7.44	14.5	8.6	7.8
Graph-density	8.32	6.43	16.92	10.48	-	-	-
∇TSVM	5.80	17.56	17.61	5.71	-	-	-
LDS	5.62	4.86	15.79	5.13	-	-	-
LapSVM	**5.44**	**3.66**	**12.67**	10.41	18.1	10.5	**6.4**
LapRLS	**5.18**	**3.36**	**12.69**	10.01	19.2	11.0	6.9
LapSVM$_{joint}$	-	-	-	-	**5.7**	**6.7**	**6.4**
LapRLS$_{joint}$	-	-	-	-	**5.6**	**8.0**	**5.8**

Real World


LENA™
EVERY WORD COUNTS
Home
LENA System
Testimonials
Resources
Shop
Newsroom
About Us
Power of Talk
About LENA
Reports
Price and Specifications
Research
FAQs
Don't 'wait and see' – make sure your baby is getting the best.
Take control of your child's development and enhance their language environment
LENA provides you with accurate and critical information about your child's language environment.
The Developmental Snapshot allows you to measure and track your child's development on a monthly basis so you can track your child's progress and always know how well he or she is doing.
The LENA Digital Language Processor (DLP) weighs just 2 ounces, but captures up to 16 hours of talk and is durable enough to last for years! It fits snugly and unobtrusively into a pouch on the front of your child's LENA clothing. Using optimal acoustic quality, it collects the sounds and words that your child says and hears. The LENA software analyzes this data, generating easy-to-read reports so you can track your child's development. These reports let you see adult word count, as well as conversational turns – those moments when you speak to your child and your child responds, and vice versa.
What's Included in the LENA System
LENA's Easy 3-Step Process
– use 2-3 times per month
1
Slip the LENA Digital Language Processor into your child's comfortable, stylish LENA Clothing – and forget about it.
View all clothing - functional and cute!
2
At the end of the day, plug the LENA Digital Language Processor into your PC. The sophisticated language environment

Graph and Manifold Laplacian

Fix $f : X \to \mathbb{R}$.
Fix $x \in \mathcal{M}$

$$[L_n f](x) = \frac{1}{n t_n (4\pi t_n)^{d/2}} \sum_j (f(x) - f(x_j)) e^{-\frac{\|x - x_j\|^2}{4 t_n}}$$

Put $t_n = n^{-1/(d+2+\alpha)}$, where $\alpha > 0$

with prob. 1, $\lim_{n \to \infty} (L_n f)|_x = \Delta_{\mathcal{M}} f|_x$

Belkin (2003), Belkin and Niyogi (2004,2005)

also Lafon (2004), Coifman et al,Hein, Gine and Koltchinski

Random Graphs and Matrices

Given $x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$

$$W_{ij} = \frac{1}{t(4\pi t)^{d/2}} e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$Eig[D - W] = Eig[L_n^{t_n}] \to Eig[\Delta_{\mathcal{M}}] \; O(\frac{1}{n^{1/(d+3)}})$$

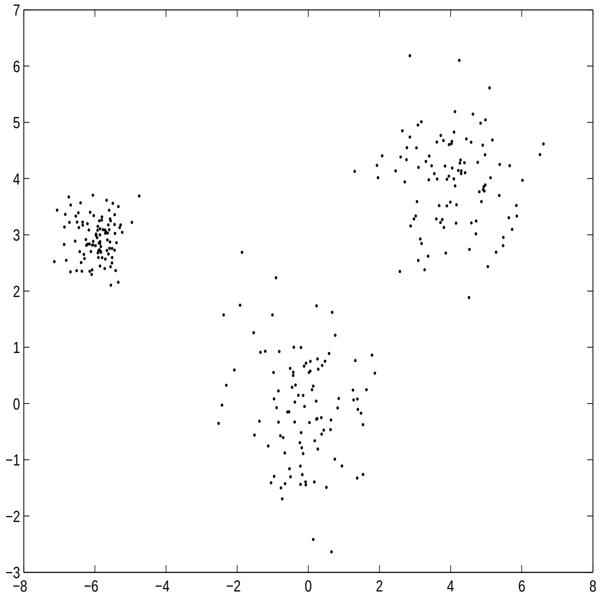
Belkin Niyogi 06,08

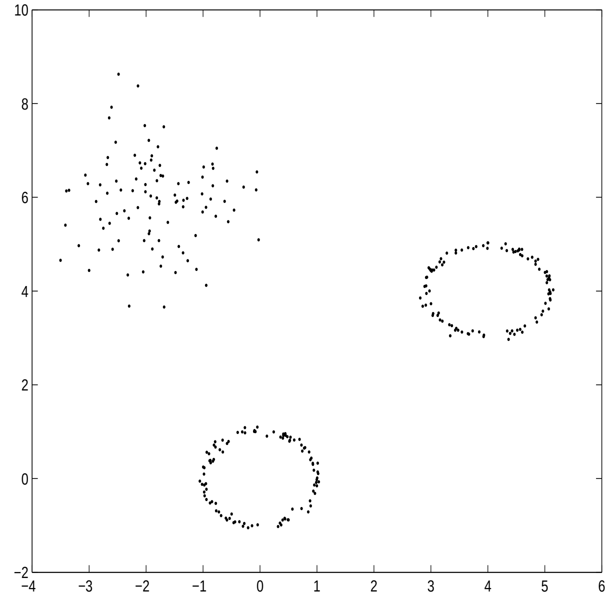
Allows us to reconstruct spaces of functions on the manifold.

(Patodi, Dodziuk: triangulated manifolds)

Manifold + Noise

Flexible, non-parametric, geometric probability model.





Remarks on Noise

1. Arbitrary probability distribution on the manifold: convergence to weighted Laplacian.
2. Noise off the manifold:

$$\mu = \mu_{\mathcal{M}} + \mu_{\mathbb{R}^N}$$

3. Noise off the manifold:

$$z = x + \eta \ (\sim N(0, \sigma^2 I))$$

We have

$$\lim_{t \to 0} \lim_{\sigma \to 0} L^{t,\sigma} f(x) = \Delta f(x)$$

Local and Global Analysis

$X =$ documents, signals, financial time series, sequences

$d(x, x')$ makes sense locally

- What is good global distance? What is global geometry/topology of X?
- What is good space of functions on X that is *adapted to geometry* of X?

Similarity Metrics

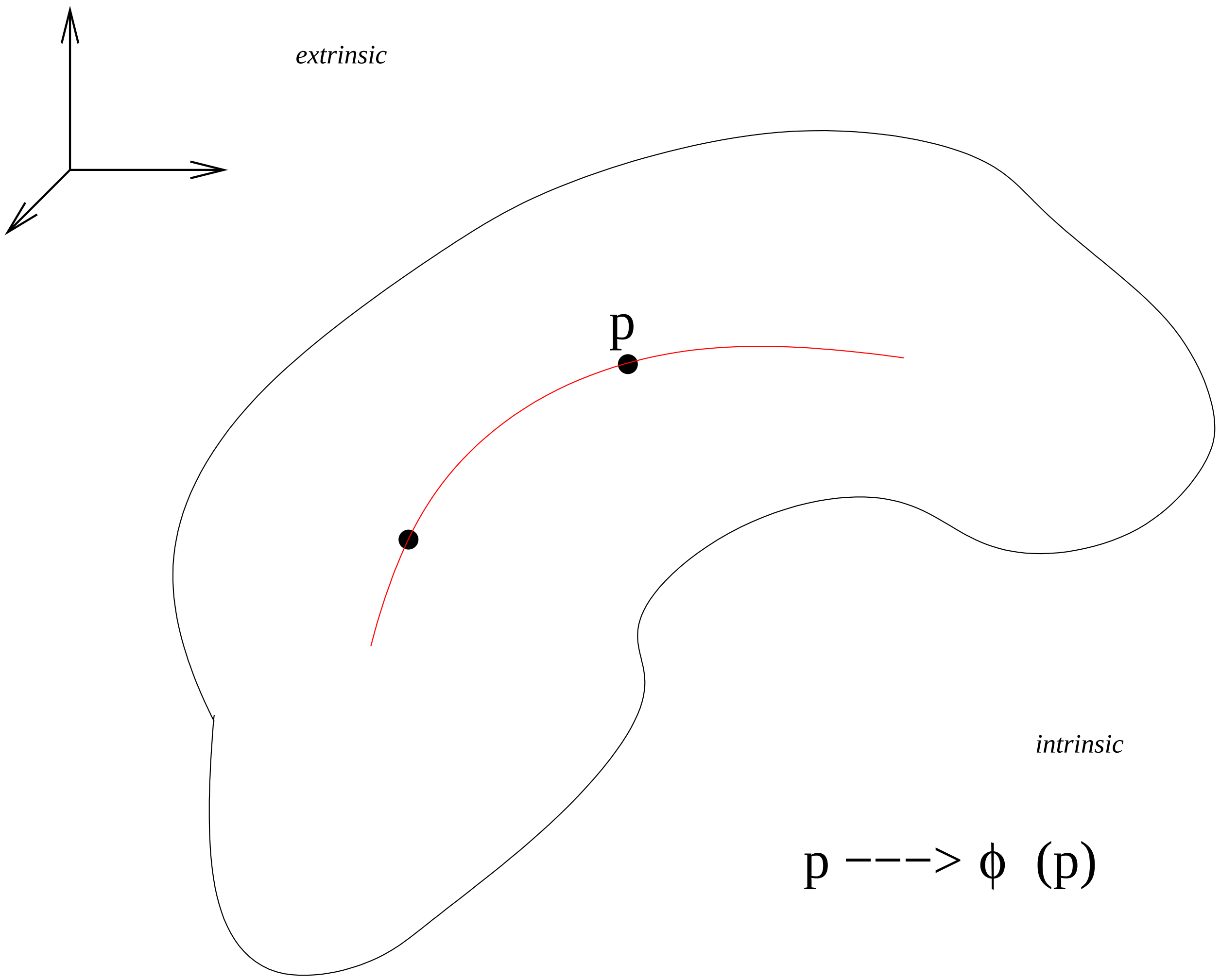
$$Sim_t(x,x') = K_t(x,x') = \alpha_t e^{\frac{-d^2(x,x')}{t}}$$

$$L_t f(x) = \int_X K_t(x,y)(f(x)-f(y))d\rho(y) \approx \frac{1}{n}\sum_{y\in X} K_t(x,y)(f(x)-f(y$$

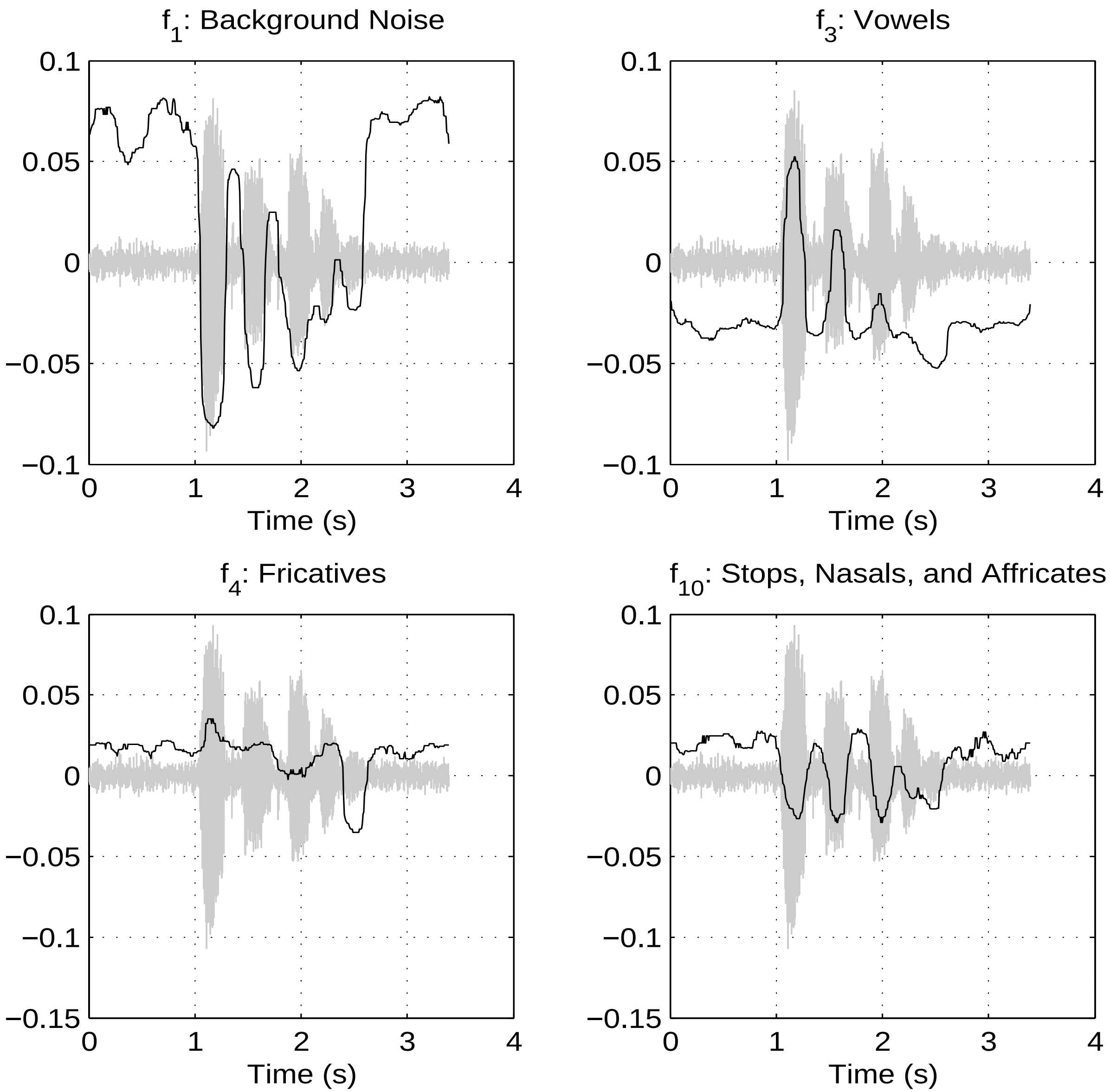
Choose t small $\rightarrow \lambda_i, \phi_i$
Choose T large $\rightarrow H_T(x,x') = \sum e^{-\lambda_i T}\phi_i(x)\phi_i(x')$

$$f = \sum_i \alpha_i \phi_i; \sum_i \alpha_i^2 g(\lambda_i)$$

Intrinsic Spectrograms

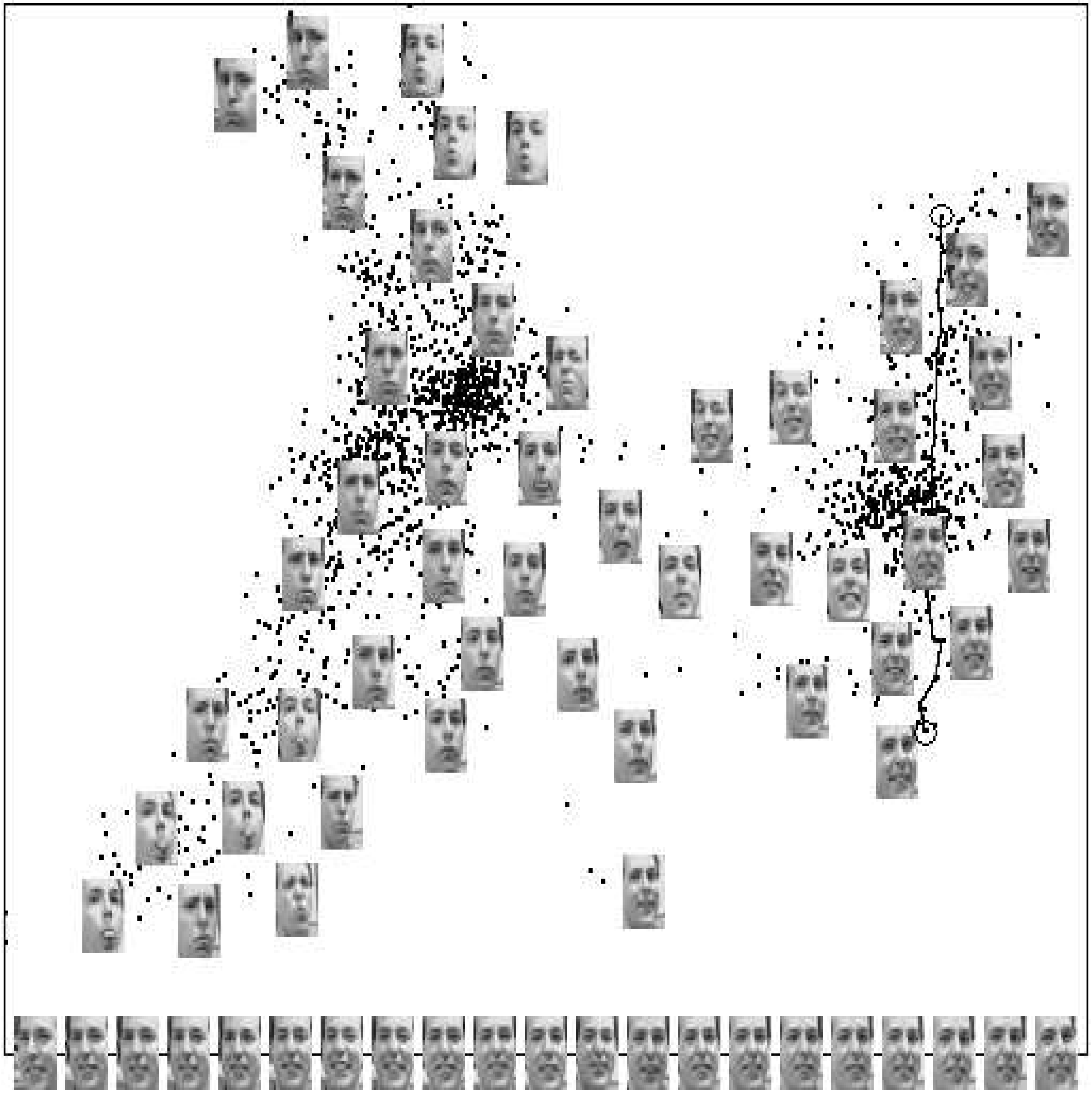


Speech and Intrinsic Eigenfunctions

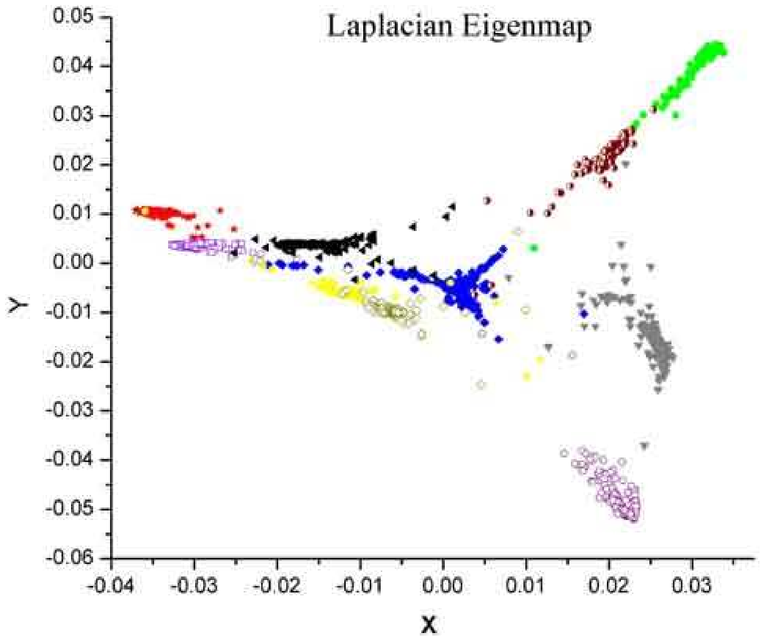


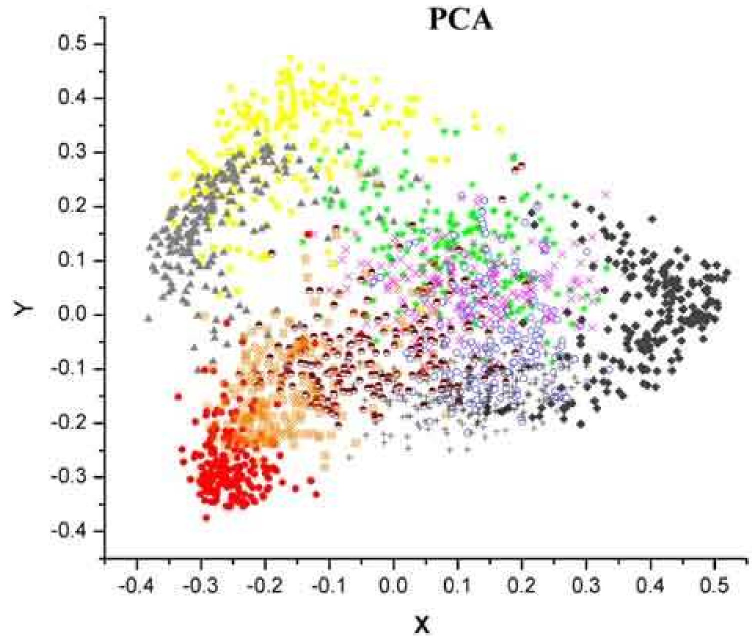
LaplacianFaces: Appearance Manifolds

X. He et al.



Visualizing Digits

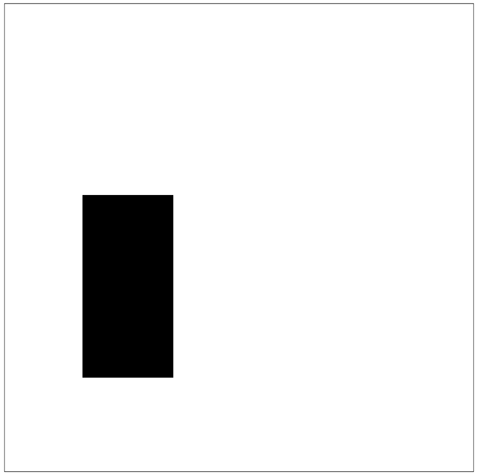




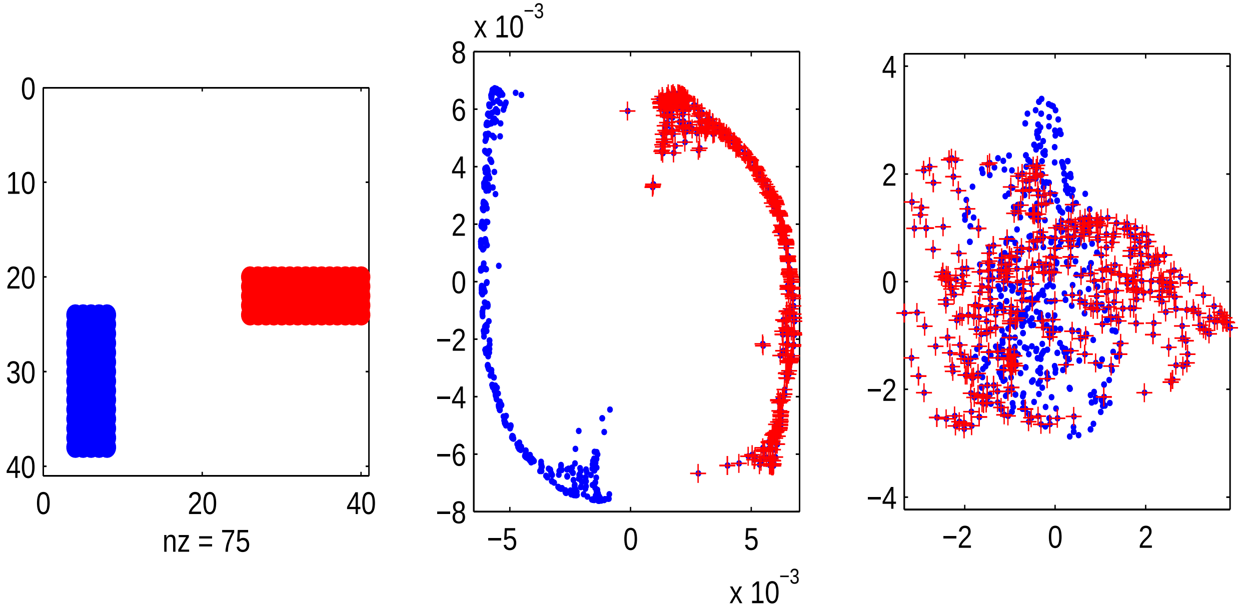
Vision Example

$$f : \mathbb{R}^2 \rightarrow [0, 1]$$

$$\mathcal{F} = \{f | f(x, y) = v(x - t, y - r)\}$$



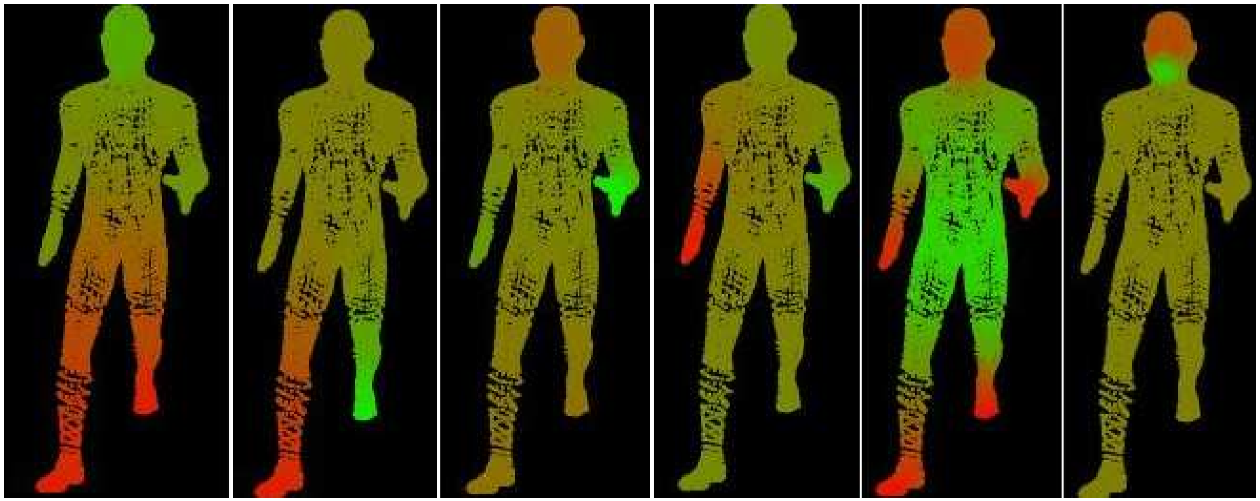
PCA versus Laplacian Eigenmaps



Computer Vision: Laplacian Eigenmaps

Machine vision: inferring joint angles.

Corazza, Andriacchi, Stanford Biomotion Lab, 05, Partiview, Surendran



Isometrically invariant representation.

Connections and Implications

- Clustering and Topology

  sparse cuts, combinatorial Laplacians, complexes

  (Niyogi, Smale, Weinberger, 2006,2008; Narayanan, Belkin, Niyogi, 2006)

- Numerical Analysis

  heat flow based algorithms, sampling, PDEs

  (Belkin, Narayanan, Niyogi, 2006; Narayanan and Niyogi, 2008)

- Random Matrices and Graphs

  results on spectra

  Belkin and Niyogi, 2008

- Speech, Text, Vision

  Intrinsic versus Extrinsic

  He et al. 2005, Jansen and Niyogi, 2006

Learning Homology

$$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$$

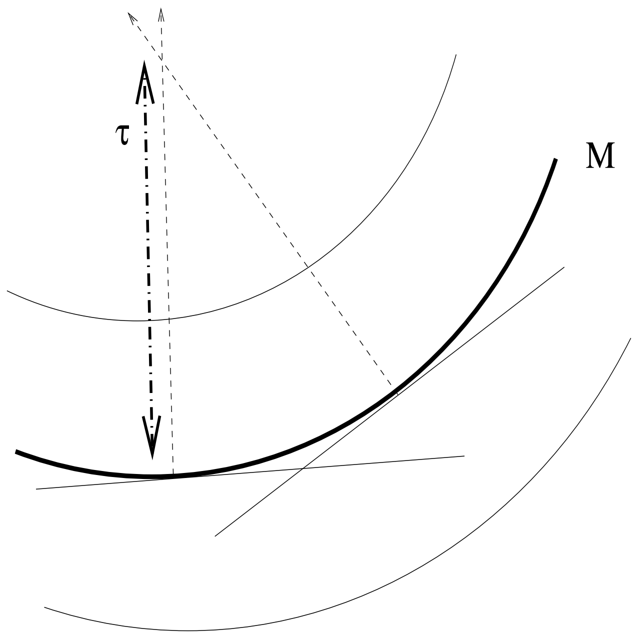
Can you learn qualitative features of $\mathcal{M}$?

- Can you tell a torus from a sphere?
- Can you tell how many connected components?
- Can you tell the dimension of $\mathcal{M}$?

(e.g. Carlsson, Zamorodian, Edelsbrunner, Guibas, Oudot, Lieutier, Chazal, Dey, Amenta,Choi,

Cohen-Steiner, de Silva etc.)

Well Conditioned Submanifolds



Tubular Neighborhood

Condition No. $\frac{1}{\tau}$

Min. distance to *medial axis*

Euclidean and Geodesic distance

$\mathcal{M} \subset \mathbb{R}^N$ condition $\sim \tau$

$p, q \in \mathcal{M}$ where $||p - q||_{\mathbb{R}^N} = d$.

For all $d \leq \frac{\tau}{2}$,

$$d_{\mathcal{M}}(p, q) \leq \tau - \tau\sqrt{1 - \frac{2d}{\tau}}$$

In fact, Second Fundamental Form Bounded by $\frac{1}{\tau}$

Homology

$$x_1, \ldots, x_n \in \mathcal{M} \subset \mathbb{R}^N$$

$$U = \cup_{i=1}^{n} B_\epsilon(x_i)$$

If ϵ well chosen, then U deformation retracts to $\mathcal{M}$.

Homology of U is constructed using the *nerve* of U and agrees with the homology of $\mathcal{M}$.

Theorem

$\mathcal{M} \subset \mathbb{R}^N$ with cond. no. τ

$\bar{x} = \{x_1, \ldots, x_n\} \sim$ uniformly sampled i.i.d.

$0 < \epsilon < \frac{\tau}{2}$ $\qquad$ $\beta = \frac{vol(\mathcal{M})}{(\sin^{-1}(\epsilon/2\tau))^d vol(B_{\epsilon/8})}$

Let $U = \cup_{x \in \bar{x}} B_\epsilon(x)$

$$n > \beta(\log(\beta) + \log(\frac{1}{\delta}))$$

with prob. $> 1 - \delta$,
homology of U equals the homology of $\mathcal{M}$

(Niyogi, Smale, Weinberger, 2004)

A Data-derived complex

$$x_1, \ldots, x_n \in \mathbb{R}^N$$

Pick $\epsilon > 0$ and balls $B_\epsilon(x_i)$

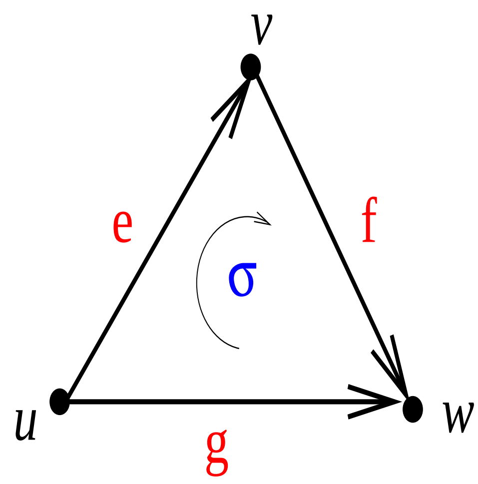
Put j-face for every $(i_0, \ldots, i_j)$ such that

$$\cap_{m=0}^{j} B_\epsilon(x_{i_m}) \neq \phi$$

Chains and the Combinatorial Laplacian

j chain is a formal sum $\sum_\sigma \alpha_\sigma \sigma$

C_j is the vector space of j-chains



$$\partial_j : C_j \to C_{j-1}$$

$$\partial_j^* : C_{j-1} \to C_j$$

$$\Delta_j = \partial_j^* \partial_j + \partial_{j+1} \partial_{j+1}^*$$

Noise

$$P \text{ on } \mathbb{R}^N$$

such that

$$P(x,y) = P(x)P(y|x) \text{ where } x \in \mathcal{M}, y \in N_x$$

$$a \leq P(x)$$

$$P(y|x) = \sigma^2 I_{N-d}$$

Small Noise

$$\sqrt{N-d}\sigma \leq c\tau$$

[Theorem]
There exists an algorithm that recovers homology that is polynomial in D.

Niyogi, Smale, Weinberger; 2008

Future Directions

- Machine Learning
  - Scaling Up
  - Multi-scale
  - Geometry of Natural Data
  - Geometry of Structured Data
- Algorithmic Nash embedding
- Random Hodge Theory
- Partial Differential Equations
- Graphics
- Algorithms