Advances in Manifold Learning

Presented by:

Nakul Verma

June 10, 2008

Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

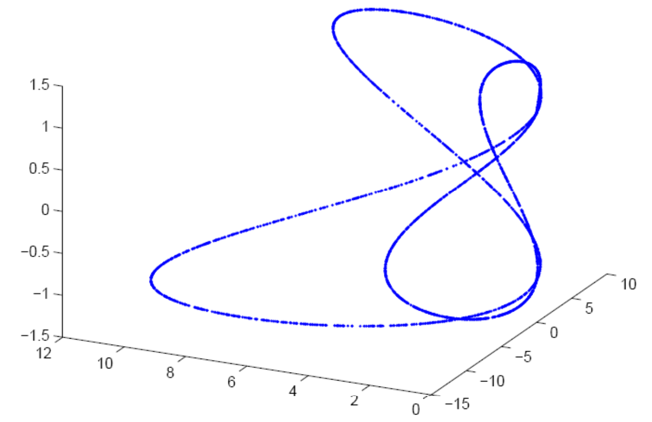
Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

What are manifolds?

Manifolds are geometric objects with that locally look like n-dimensional subspace. More formally:

$M \subseteq \Re^D$, is considered a n-dimensional manifold, if for all $p \in M$, we can find a smooth bijective map between $\Re^n$ and a neighborhood around p.



An example of a 1-dimensional manifold in $\Re^3$

- Manifolds are useful in modeling data:

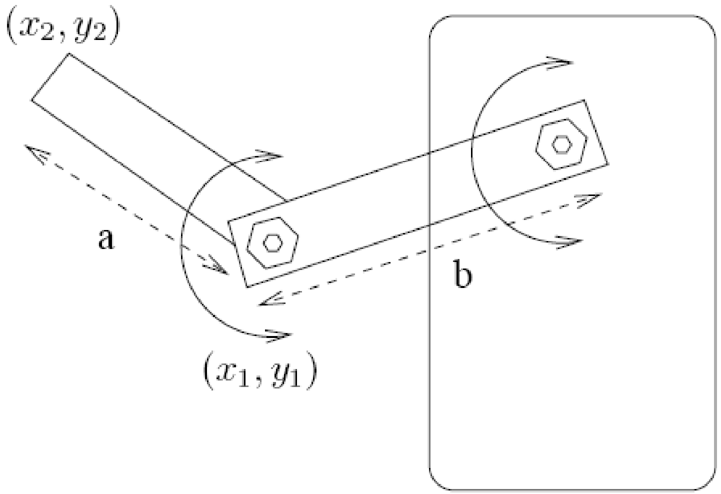
    Measurements we make for a particular observation are generally correlated and have few degrees of freedom.

    Say we make D measurements and there are n degrees of freedom, then such data can be modeled as a n-dimensional manifold in $\Re^D$

Some examples of manifolds

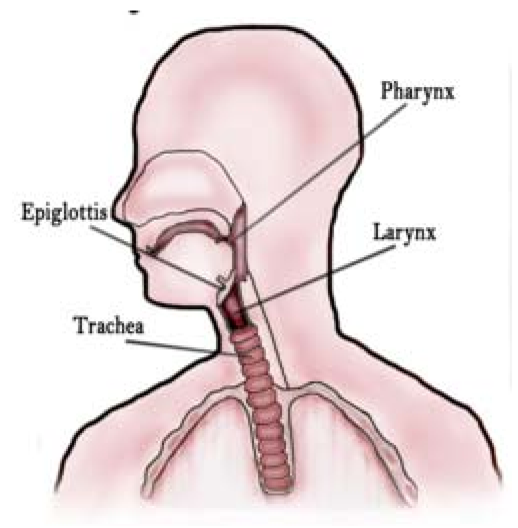
Modeling movement of a robotic arm

- Measurements taken on joints and elsewhere
- There are two degrees of freedom
- Set of all possible valid positions traces out a 2-dimensional manifold in the measurement space.



Natural process with physical constrains – speech

- Few anatomical characteristics, such as size of the vocal chords, pressure applied, etc. govern the speech signal.
- Whereas the standard representation of speech for recognition purposes, such as MFCC embed the data in fairly high dimensions.



Learning on manifolds

Learning on manifolds can be broadly defined as establishing methodologies and properties on samples coming from an underlying manifold.

Kinds of methods machine learning researchers look at:

- Finding a lower dimensional representation of manifold data
- Density estimation and regression on manifolds
- Performing classification tasks on manifolds
- and much more…

Here we will study some of these methods.

Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

Dimension reduction on manifolds

Why dimension reduction?

- Learning algorithms scale poorly with increase in dimension
- Representing the data in fewer dimensions while still preserving relevant information helps alleviate the computational issues
- It provides a simpler (shorter) description of the observations.

Dimension reduction types:

Non linear methods for dimension reduction

- For curvy objects such as manifolds, its more intuitive to have non-linear maps to lower dimension.
- Some popular techniques are: LLE, Isomap, Laplacian and Hessian Eigenmaps, etc.

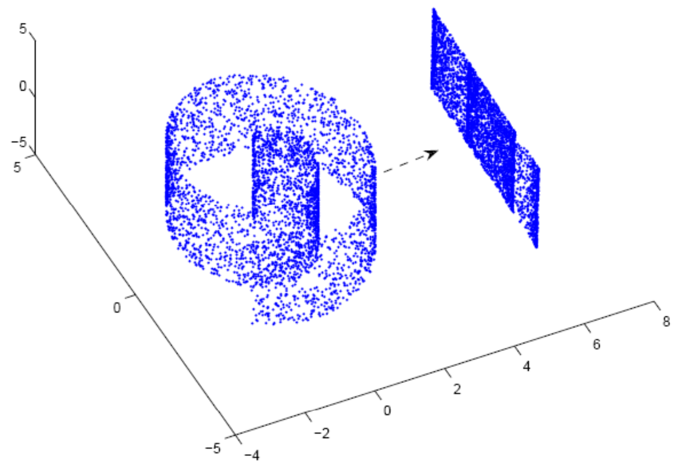
Linear Methods for dimension reduction

- Popular techniques are: PCA, random projections.

Issues with dimension reduction

Information Loss

- A low dimensional representation can result in information loss



Goal of dimension reduction

- Preserve as much relevant information as possible.
- In terms of machine learning, one good criterion is to preserve inter-point distances

Random projections of manifolds

What is Random Projections?

- Projecting the data orthogonally onto a random subspace of fixed dimension.
- Performing a random operation without even looking at the data seems questionable in preserving any kind of relevant information, we will see that this technique has strong theoretical guarantees in preserving inter-point distances!

Main Result (Baraniuk and Wakin [2])

Theorem: Let *M* be a *n*-dimensional manifold in $\Re^D$, Pick $\varepsilon > 0$ and let $d = \Omega(n/\varepsilon^2 \log D)$, then there is a linear map $f: \Re^D \rightarrow \Re^d$, such that for all $x, y \in M$,

$$(1-\varepsilon) \leq \|f(x) - f(y)\| / \|x - y\| \leq (1+\varepsilon)$$

(a projection onto a random *d* dim subspace will satisfy this with high probability)

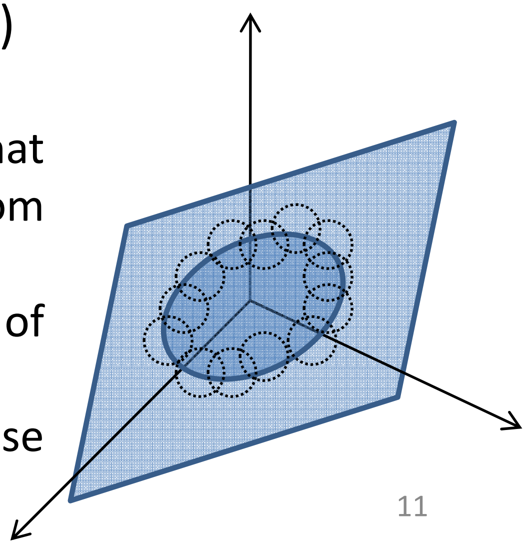
Proof Idea

1. A set of *m* points in $\Re^D$ can be embedded into d=Ω(log *m*) dimensions such that all interpoint distances are approximately preserved using a random projection (Johnson and Lindenstrauss [6], [5])
   - Consider a $D \times$ d Gaussian random matrix R, then for any $x \in \Re^D$, $\|R^T x\|^2$ is sharply concentrated around its expectation (= d/D $\|x\|^2$).
   - It follows that, if $f : x \mapsto \sqrt{D/d}\, R^T x$, then w.h.p.

$$\|f(x) - f(y)\|^2 = \frac{D}{d}\|R^T(x-y)\|^2 \le \frac{D}{d}(1+\varepsilon)\frac{d}{D}\|x-y\|^2$$

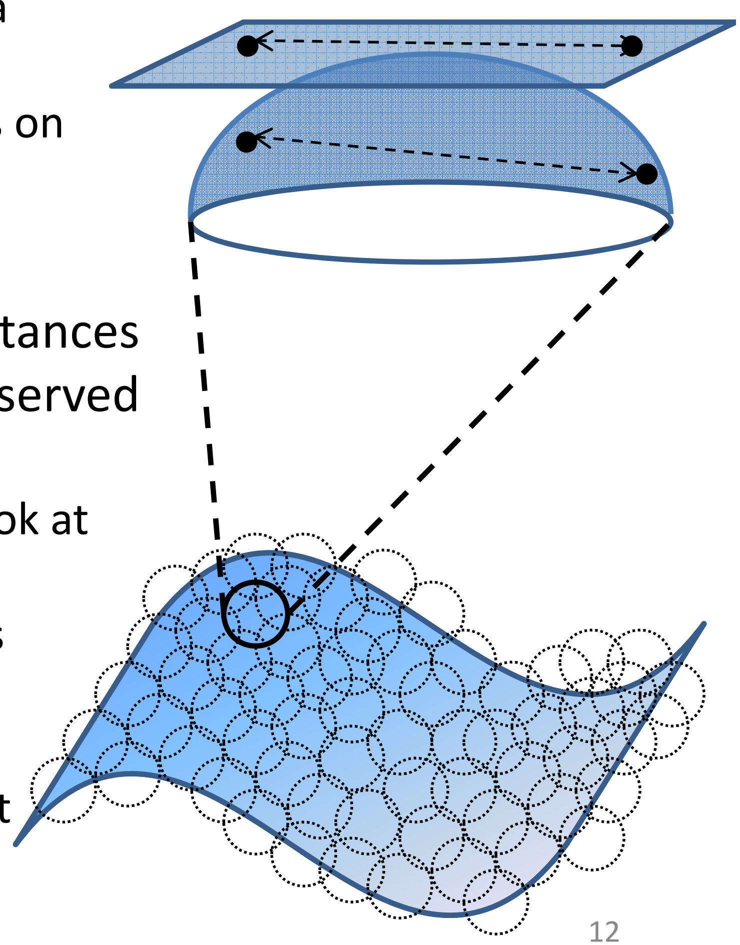
   - Similarly we can lower bound. Apply union bound on all O(m^2) pairs.

2. Not just a point-set, but an *entire* n-dimensional subspace of $\Re^D$ can be preserved by a random projection onto Ω(*n*) dimensions (Baraniuk, et.al. [1])
   - Due to linearity of norms, we only need to consider that length of a unit vector is preserved after a random projection.
   - Note that a unit ball in $\Re^n$, can be covered by $(1/\varepsilon)^n$ balls of radius ε. Apply step 1 to centers of these balls.
   - Any unit vector can be well approximated with one of these representatives (for a small enough ε)



Proof Idea (cont.)

3. Distances between points in a sufficiently small region of a manifold are well preserved (Baraniuk and Wakin [2]).
   - Assume manifold has bounded curvature, then a small enough region approximately looks like a subspace.
   - We can apply the step 2, to preserve distances on the subspace.



4. Taking an ε-cover of the manifold, distances between far away points are also well preserved (Baraniuk and Wakin [2]).
   - For any two far away points *x* and *y*, we can look at their closest ε-cover representative.
   - Step 3 ensures that distance between *x* and its representative, and *y* and its representative is preserved.
   - Since ε-cover is a point-set, step 1 ensures that distances among representatives would be preserved.

Random projections on manifolds

We have shown:

- An orthogonal linear projection onto a random subspace has a remarkable property to preserve all interpoint distances on a manifold.
- This can be used to preserve geodesic distances as well.

It would be nice to know:

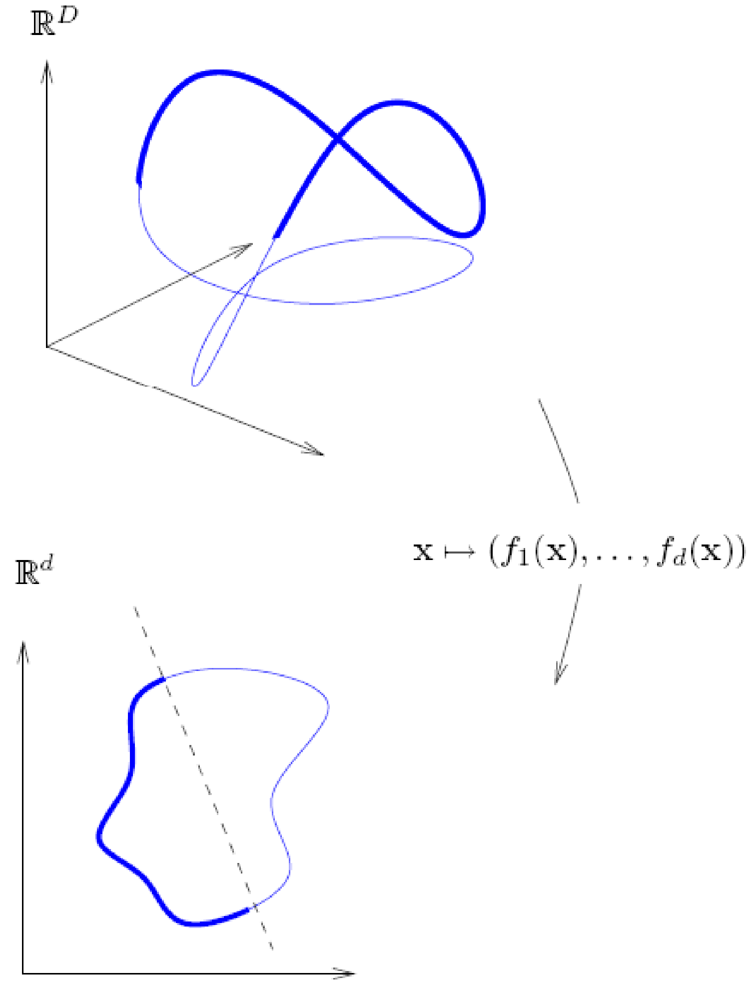
- What lower bounds (in terms of projection dimension) are achievable if we want to preserve ‘average’ distortion as opposed to worst case distortion.

Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

Laplacian Eigenmaps on manifolds

Laplacian Eigenmaps are a non-linear dimension reduction technique on manifold



Basic idea:

- To preserve the local geometry of the manifold.
- Has a remarkable effect of simplifying manifold structure.

Uses:

- Aids in classification tasks on data from a manifold.

Derivation of Laplacian Eigenmaps

Geometric derivation:

- Let $f : M \to \Re$ that maps nearby points on a manifold close together on a line.
- For any closeby $x,y \in M$, let $l=d_M(x,y)$ be the geodesic distance. Then,

$$|f(x) - f(y)| \leq l\|\nabla f(x)\| + o(l)$$

- Hence want to minimize $\|\nabla f(x)\|$ in 'sum squared sense'

$$\arg\min_{\|f\|=1} \int_M \|\nabla f(x)\|^2$$

- Now $\int \|\nabla f(x)\|^2 = \langle \nabla f, \nabla f \rangle = \langle f, \Delta f \rangle$, where Δ is the Laplace-Beltrami operator.

- Thus, minimum of $\langle f, \Delta f \rangle$ is given by eigenfunction corresponding to the lowest eigenvalue of Δ.

- Generalizing to $\Re^d$, we can map $x \mapsto (f_1(x), \ldots, f_d(x))$ (f_i eigenfunction).

Derivation of Laplacian Eigenmaps

Laplace as smoothness functional:

- From theory of splins, we can measure the smoothness of a function as:

$$S(f) = \int_{S^1} \left| f(x)' \right|^2 dx$$

- This can be naturally extended for functions over a manifold

$$S(f) = \int_M \left\| \nabla f(x) \right\|^2 dx = \langle f, \Delta f \rangle$$

- Observe that smoothness of (unit norm) eigenfunction e_i is controlled by the corresponding eigenvalue. Since $S(e_i) = \langle e_i, \Delta e_i \rangle = \lambda_i$

- Thus, since $f = \sum c_i e_i$, we immediately get $S(f) = \langle \sum c_i e_i, \sum c_i \Delta e_i \rangle = \sum \lambda_i c_i^2$ so, first *d* eigenfunctions, gives a way to control smoothness.

Approximating Laplacian from samples

Graph Laplacian – a discrete approximation to Δ.

- Let $x_1,\ldots,x_m$ be sampled uniformly at random from a manifold. Let $\omega_{ij} = e^{-\|x_i - x_j\|^2/4t}$ then the matrix is called the graph Laplaican

$$\left(L_m^t\right)_{ij} = \begin{cases} -\omega_{ij} & \text{if } i \neq j \\ \sum_k \omega_{ik} & \text{otherwise} \end{cases}$$

- Note that, for any $p \in M$ and f on M :

$$L_m^t f\left(p\right) = f(p)\frac{1}{m}\sum_j e^{-\|p-x_j\|^2/4t} - \frac{1}{m}\sum_j f(x_j)e^{-\|p-x_j\|^2/4t}$$

Main Result (Belkin and Niyogi [4])

Theorem: For any $p \in M$, and a smooth map f, if $t \to 0$ sufficiently fast, then as $m \to \infty$:

$$L_m^t f\left(p\right) = \frac{1}{\mathrm{Vol}\left(M\right)}\Delta f\left(p\right)$$

Proof Idea

For a fixed $p \in M$, and a smooth map f,

1. Using concentration inequalities, we can deduce that L_m^t converges to its continuous version L^t.

$$L^t f(p) = f(p)\int e^{-\|p-x_j\|^2/4t} \mu dx - \int f(x_j) e^{-\|p-x_j\|^2/4t} \mu dx$$

   - This follows almost immediately from law of large numbers.

2. We can relate L^t with Δ by

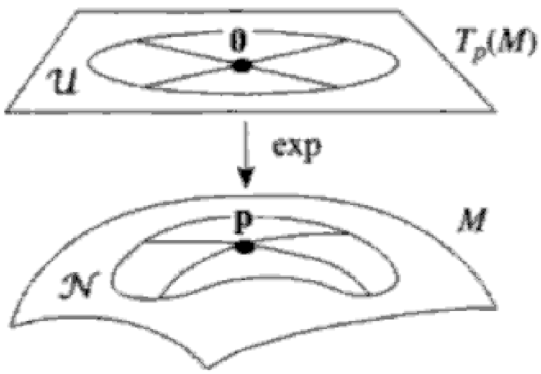
   (a) Reducing the entire integral to a small ball in M. This would help us express the L^t in a single local coordinate system.

   - Choosing t small enough guarantees that most of the contribution to the integral comes from points from a single local chart.

Proof Idea (cont.)

(b) Applying change of coordinates so that L^t can be expressed as a new integral in a n-dimensional Euclidian space.

- Canonical exponential map on manifolds sends vectors emanating from ***0*** in tangent space to geodesics from *p* in *M*.
- We can use the reverse exponential map to represent L^t in tangent space.



(c) Relating the new integral in $\Re^n$ to Δ.

- Using Taylor approximation and choosing *t* appropriately,

$$L^t f(p) \approx \frac{-1}{\mathrm{Vol}(M)} \int_B \left(x \nabla f + \frac{1}{2} x^T H x \right) e^{-\|x\|^2/4t} dx$$

$$= \frac{-tr(H)}{\mathrm{Vol}(M)} = \frac{1}{\mathrm{Vol}(M)} \Delta$$

Noting that since *M* is compact and any *f* can be approximated arbitrarily well by a sequence of functions f_i, we can get a uniform convergence for the entire *M* for any *f*.

Laplacian Eigenmaps on manifolds

We have shown:

- Preserving local distances yield a natural non-linear dimension reduction method that has a remarkable property of finding a smoother representation of the manifold.
- If the points are sampled uniformly at random from the underlying manifold, then the graph Laplacian approximates the true Laplacian.

It would be nice to know:

- What if the points are sampled independently from a non-uniform measure?
- We have seen that the spectrum of Laplacian basis gives a smooth approximation for functions on a manifold. What effects do Fourier basis or Lagrange basis have?

Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

Density estimation

Let f be an underlying density on $\Re^D$ and $\hat{f}_m$ be our estimate from *m* independent samples.

We can define quality of our estimate as $\mathrm{E}\int\left(\hat{f}_m(x) - f(x)\right)^2 dx$

This is also called the expected risk.

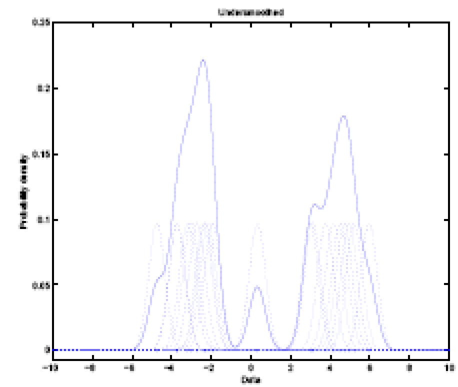
We are interested in how fast does expected risk decrease with increase in samples.

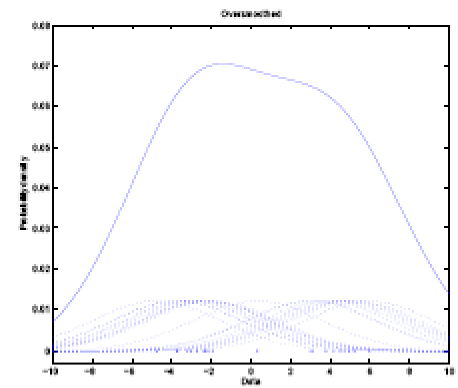
How to estimate $\hat{f}_m$ from samples?

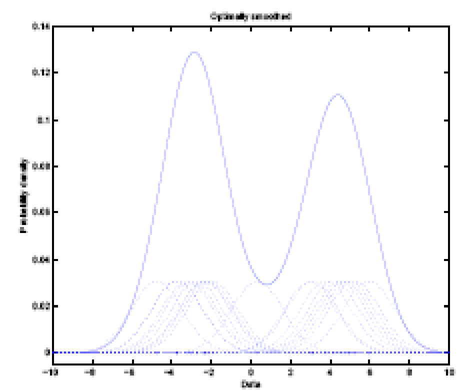
- Histograms
  - issues with smoothness
  - issues with grid placement
- Kernel density estimators

Kernel density estimation

- Density estimator that alleviates the problems of histograms
- Places a ’kernel function’ on each observed sample
  i.e. a function that is non-negative, has zero mean, finite variance, and integrates to one.
- Estimator is given by $f_{m,K}(x) = \frac{1}{mh^D} \sum_{i=1}^{m} K\left(\frac{\|x - x_i\|}{h} \right)$
  (*h* is a bandwidth parameter)







Properties:

- Bandwidth parameter is more important than the form of the kernel function for $\hat{f}_m$
- For optimal value of *h*, risk decreases as $O(m^{-4/4+D})$

Kernel Density estimation on manifolds

- We will use the following modified estimator:

$$f_{m,K}(p) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h^n \theta_{x_i}(p)} K\left(\frac{d_M(p, x_i)}{h}\right)$$

where $\theta_p(q)$ is the volume density function R $\exp^{-1}$(q) at p.

R is the ratio of canonical measure to the Lebesgue measure

Main Result (Pelletier [7])

Theorem: Let f be the underlying density over a n-dimensional manifold in $\Re^D$ and $f_{m,K}$ as above, then:

$$\mathbf{E}\left\| \hat{f}_{m,K} - f \right\|^2 \leq C\left(\frac{1}{mh^n} + h^4\right)$$

setting $h \approx m^{-1/n+4}$, we get the rate of convergence of $O(m^{-4/n+4})$

Proof Idea

1. Separately bounding the squared bias and variance of the estimator.
   - We can bound the pointwise bias by applying change of coordinates via the exponential map and using Taylor approximation (as before).
   - Integrating the squared pointwise bias gives the following

$$\int_M b^2(p)dp \le O\left(h^4 \mathrm{Vol}(M)\right)$$

   - We can bound the pointwise variance by using $\mathrm{Var}(X) \le \mathrm{E}X^2$
   - Integrating variance and using properties of $\theta_p(q)$ gives the following

$$\int_M \mathrm{Var}\hat{f}_{m,K}(p)dp \le O\left(1/mh^n\right)$$

2. Decomposing the risk to its bias and variance components.
   - Note that

$$\mathbf{E}\left\| \hat{f}_{m,K} - f \right\|^2 = \int \left(\mathbf{E}\hat{f}_{m,K}(p) - f(p)\right)^2 dp + \int \mathrm{Var}\left(\hat{f}_{m,K}(p)\right) dp$$

Kernel density estimation on manifolds

We have shown:

- Rates of convergence of a kernel density estimator on manifolds are independent of the ambient dimension *D*.
- They depend exponentially on the manifold's intrinsic dimension *n*.

It would be nice to know:

- How to estimate $\theta_p(q)$?
- What about rates of convergence in ℓ_1 or ℓ_∞?

Outline

- Motivation
  - Manifolds
  - Manifold Learning
- Random projection of manifolds for dimension reduction
  - Introduction to random projections
  - Main result and proof
- Laplacian Eigenmaps for smooth representation
  - Laplacian eigenmaps as a smoothness functional
  - Approximating the Laplace operator from samples
- Manifold density estimation using kernels
  - Introduction to density estimation
  - Sample rates for manifold kernel density estimation
- Questions / Discussion

Summary of results

- Random projections for manifolds
  - An orthogonal linear projection onto a random subspace can preserve all interpoint distances on a manifold.
  - Random projections can also preserve geodesic distances.
- Laplacian Eigenmaps for manifold smoothness
  - Preserving local distances yield a natural non-linear dimension reduction method for finding a smoother representation of the manifold.
  - If the points are sampled uniformly at random from the underlying manifold, then the graph Laplacian approximates the true Laplacian.
- Manifold density estimation using kernels
  - Rates of convergence of a kernel density estimator on manifolds are independent of the ambient dimension D.
  - They depend exponentially on the manifold’s intrinsic dimension n.

Questions/Discussion

- What is the best (isometric) embedding dimension can we hope for?

- Results depend heavily on intrinsic manifold dimension. How to estimate this quantity?

- How can we relax the ‘manifold assumption’?

References

[1] R. Baraniuk, et. al. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 2008.

[2] R. Baraniuk and M. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 2007.

[3] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian based manifold methods. *Journal of Computer and System Sciences*, 2007.

[5] S. Dasgupta and A. Gupta. An elementary proof of the Johnson-Lindenstrauss lemma. *UC Berkeley Tech. Report 99-006*, March 1999.

[6] W. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Conf. in Modern Analysis and Probability*, pages 189–206, 1984.

[7] B. Pelletier. Kernel density estimation on Riemannian manifolds. *Statistics and Probability Letters*, 73:297–304, 2005.