Mathematical Advances in Manifold Learning

Nakul Verma
University of California, San Diego
naverma@cs.ucsd.edu

June 03, 2008

Abstract

Manifold learning has recently gained a lot of interest by machine learning practitioners. Here we provide a mathematically rigorous treatment of some of the techniques in unsupervised learning in context of manifolds. We will study the problems of dimension reduction and density estimation and present some recent results in terms of fast convergence rates when the data lie on a manifold.

1 Introduction

With increase in the volume of data, both in terms of number of observations as well as number of measurements, traditional learning algorithms are now faced with new challenges. One may expect that more data should lead to more accurate models, however a large collection of irrelevant and correlated features just add on to the computational complexity of the algorithm, without helping much to solve the task at hand. This makes the learning task especially difficult. In an attempt to alleviate such problems, a new model in terms of *manifolds* for finding relevant features and representing the data by a few parameters is gaining interest by machine learning and signal processing communities.

Most common examples of superficially high dimensional data are found in the fields of data mining and computer vision. Consider the problem of estimating the face and body pose in humans. Knowing where a person is looking gives a wealth of information to an automated agent regarding where the object of interest is – whether the person wants to interact with the agent or whether she is conversing with another person. The task of deciding where someone is looking seems quite challenging given the fact that the agent is only receiving a large array of pixels. However, knowing that a person's orientation only has one degree of freedom, the relevant information in this data can be expressed by just a single number – the angle of the turn, i.e. the orientation of the body.

In a typical learning scenario the task is slightly more complicated as the agent only gets to see a few samples from which it somehow needs to interpolate and generalize various possible scenarios. In our example this translates to the agent only having access to few of the body poses, from which it needs to predict where the person is looking. Thus the agent is faced with the difficulty to find an appropriate (possibly non-linear) basis to represent this data compactly. *Manifold learning* can be broadly described as the study of algorithms that use and inferring the properties of data that is sampled from an underlying manifold.

The goal of this survey is to study different mathematical techniques by which we can estimate some global properties of a manifold from a few samples. We will start by studying random projections as a nonadaptive linear dimensionality reduction procedure, which provides a probabilistic guarantee on preserving the interpoint distances between all points on a manifold. We will then focus on analyzing the spectrum of Laplace-Beltrami operator on functions on a manifold for finding non-linear embeddings and simplifying its structure. Lastly we will look at kernel density estimation to estimate high density regions on a manifold.

It is worth mentioning that our survey is by no means comprehensive and we simply highlight some of the recent theoretical advances in manifold learning. Most notably we do not cover the topics of regularization, regression and clustering of data belonging to manifolds. In the topic of dimensionality reduction, we are skipping the analysis of classic techniques such as LLE (Locally Linear Embedding), Isomap and their variants.

1.1 Preliminaries

We begin by introducing our notation which we will use throughout the paper.

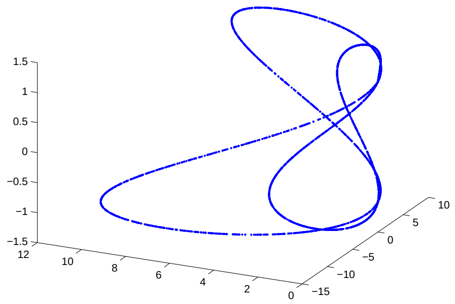

Figure 1: A 1-manifold in $\mathbb{R}^3$

Definition 1. *We say a function $f : U \mapsto V$ is a diffeomorphism, if it is smooth*[1] *and invertible with a smooth inverse.*

Definition 2. *A subset $M \subset \mathbb{R}^D$ is said to be a smooth n-manifold if M is locally diffeomorphic to $\mathbb{R}^n$, that is, at each $p \in M$ we can find an open neighborhood $U \subset \mathbb{R}^D$ such that there exist a diffeomorphic map between $U \cap M$ and $\mathbb{R}^n$.*

It is always helpful to have a picture in mind. See figure 1 for an example of 1-manifold in $\mathbb{R}^3$. Notice that locally any small segment of the manifold "looks like" an interval in $\mathbb{R}^1$.

Definition 3. *A tangent space at a point $p \in M$, denoted by T_pM, is the affine subspace formed by collection of all tangent vectors to M at p.*

For the purposes of this survey we will restrict ourselves to the discussion of manifolds whose tangent space at each point is equipped with an inner product. Such manifolds are called Riemannian manifolds and allow us to define various notions of length, angles, curvature, etc. on the manifold.

Since we will largely be dealing with samples from a manifold, we need to define

Definition 4. *A sequence $x_1, \ldots, x_n \subset M \subset \mathbb{R}^D$ is called independent and identically distributed (i.i.d.) when each x_i is picked independently from a fixed distribution $\mathcal{D}$ over M.*

With this mathematical machinery in hand, we can now demonstrate that manifolds incorporate a wide array of important examples – we present two such examples that serve as a motivation to study these objects.

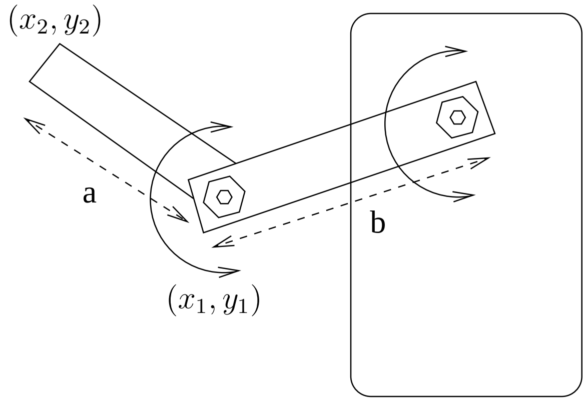

Figure 2: Movement of a robot's arm traces out a 2-manifold in $\mathbb{R}^4$

1.2 Some examples of manifolds

Movement of a robotic arm: Consider the problem of modelling the movement of a robotic arm with two joints (see figure 2). For simplicity let's restrict the movement to the $2D$-plane. Since there are two degrees of freedom, intuitively one should suspect that the movement should trace out a 2-manifold. We now confirm this in detail.

Let's denote the fixed shoulder joint as the origin, the position of the elbow joint as (x_1, y_1) and the position of wrist as (x_2, y_2). To see that the movement of the robotic arm traces out a 2-manifold, consider the map $f : \mathbb{R}^4 \to \mathbb{R}^2$ defined as $(x_1, y_1, x_2, y_2) \mapsto (x_1^2 + y_1^2, (x_2 - x_1)^2 + (y_2 - y_1)^2)$. Clearly $M \subset \mathbb{R}^4$, s.t. $M = f^{-1}(b^2, a^2)$ is the desired manifold. We can verify that locally M is diffeomorphic to $\mathbb{R}^2$ by looking at its derivative map $Df = 2\begin{pmatrix} x_1 & y_1 & 0 & 0 \\ x_1 - x_2 & y_1 - y_2 & x_2 - x_1 & y_2 - y_1 \end{pmatrix}$ and observing that it has maximal rank for non-degenerate values of a and b.

Set of orthogonal $n \times n$ matrices: We present this example to demonstrate that manifolds are not only good for representing physical processes with small degrees of freedom but also to better understand some of the abstract objects which we regularly encounter. Consider the problem of understanding the geometry the set of orthonormal matrices in the space of real $n \times n$ matrices. Note that the set of $n \times n$ orthonormal matrices is also called the orthogonal group, and is denoted by $O(n)$. We claim that this set forms a $k(k-1)/2$-manifold in $\mathbb{R}^{n^2}$.

To see this, consider the map $f : \mathbb{R}^{n^2} \to \mathbb{R}^{n(n+1)/2}$ defined by $(A)_{ij} \mapsto A^TA$. Now $M \subset \mathbb{R}^{n^2}$ such that

[1] recall that a function is smooth if all its partial derivatives $\partial^n f/\partial x_{i_1} \ldots \partial x_{i_n}$ exist and are continuous.

$M = f^{-1}(I_{n\times n})$ is exactly $O(n)$. To see that M is in fact a manifold, observe that the derivative map $Df_A.B = B^T A + A^T B$ is regular.

Observe that the examples above required us to know the mapping f a priori. However in the context of machine learning, the task is typically to estimate properties about M *without* having access to f.

1.3 Outline

The paper is organized as follows. We will discuss some linear and non-linear dimensionality reduction methods on manifolds with a special focus on random projections in section 2. We will then study Laplacian-eigenmaps as a process to simplify manifold structure in section 3, followed by nonparametric density estimation techniques on manifolds in section 4. We will finally conclude by discussing the significance of the results and some directions for future work in section 5.

2 Random projections for linear dimension reduction

Dimension reduction is an important preprocessing step in data analysis that has been studied extensively. Here we provide the motivation for why dimension reduction of data is desirable. We briefly discuss different techniques that have been employed for dimension reduction on data coming from an underlying manifold and examine a recently analyzed technique of random projections.

2.1 Dimensionality reduction

We know that learning algorithms scale poorly with the dimension of the data. This makes *dimension reduction* a popular preprocessing step – first map the data into a lower dimensional space while preserving the relevant information, and then run the regular learning algorithms in the smaller projected space.

One reasonable criterion to measure the quality of our low dimensional mapping is to test how well doest the mapping preserves pairwise distances. The basic intuition is that the distances between points in space relate to the dissimilarity between the corresponding observations. Thus, it is undesirable that two points that are far apart in the original space get mapped close to each other by performing a dimension reduction. Similarly, we would not want points that were close originally to get mapped far apart.

As one might expect, finding a mapping that preserves all distances of an arbitrary dataset can be a difficult task. Luckily in our case, the saving grace comes from observing that the data has a manifold structure. We are only required to preserve distances between points that lie on the manifold and not the whole ambient space.

2.1.1 Dimension reduction of manifold data

In the past decade, numerous methods for manifold dimension reduction have been proposed. The classic techniques such as Locally Linear Embeddings (LLE) and Isomaps, and newer ones such as Laplacian Eigenmaps and Hessian Eigenmaps, all share a common intuition – all these methods try to capture the local manifold geometry by constructing the adjacency graph on the sampled data. They all benefit from the observation that inference done on this neighborhood graph corresponds approximately to the inference on the underlying manifold. For a comprehensive survey we refer the readers to [8].

Note that these methods are examples of non-linear dimensionality reduction techniques on manifolds. However, we will present a *linear* dimension reduction technique that works surprisingly well on manifolds. The goal is to find a *linear* map $\Phi : \mathbb{R}^D \mapsto \mathbb{R}^d$, preferably $d \ll D$, which when applied to the data, preserves all interpoint distances. More formally we want to give guarantees of the form: for all $x, y \in M$, $\|x - y\| \approx \|\Phi x - \Phi y\|$.

2.1.2 Issues with principal component analysis

Arguably the most popular linear dimension reduction technique is the Principal Component Analysis (PCA). The main idea is to find an affine subspace of a specified dimension that captures maximum amount of variance in the data. It turns out that this optimization problem can be solved efficiently in closed form, and the desired optimal subspace is given by the span of the top d eigenvectors (corresponding to the top eigenvalues) of the covariance matrix of the data [17].

Unfortunately PCA, like all deterministic linear projection methods, is not suited for asserting global distance preservation guarantees on all pairwise points. One can easily construct examples where distances among far away points in the original space get collapsed in the projected space (see figure 3). Instead we will look at projecting the data onto a *random subspace*.

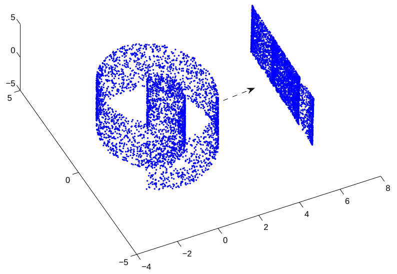

Figure 3: PCA projection can sometimes collapse distances between faraway points, making it an undesirable choice for distance preserving dimension reduction.

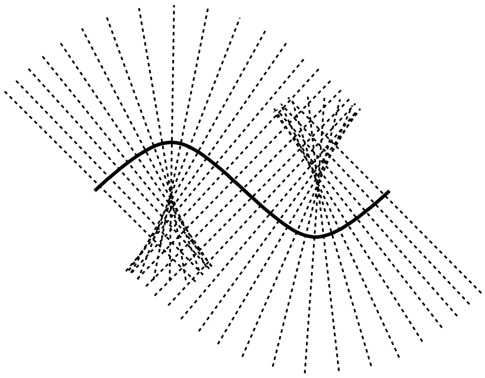

Figure 4: Tubular neighborhood of a manifold. Note that the normals (dotted lines) of a particular length incident at each point of the manifold (solid line) will intersect if the manifold is too curvy.

2.1.3 Random projections of manifolds

As the name suggests, random projections is concerned with projecting the data onto a *random* subspace of a fixed dimension d. We would be able to conclude that if the data lie on a manifold M, with high probability, projecting the data downto a sufficiently large random subspace would approximately preserve all interpoint distances. At a first glance this result appears very counter-intuitive – after all how can projecting the data onto a random subspace, that doesn't even take into account the samples, has the capability to preserve distances?

Starting point of such a counter-intuitive result is the much celebrated theorem of Johnson and Lindenstrauss which states that any point-set of size m in $\mathbb{R}^D$ can be embedded in $\mathbb{R}^{O(\log m)}$ with small distortion by using a linear map. Moreover, this linear map is essentially a random subspace of the desired embedding dimension.

We can leverage this result and get the basic proof outline for preserving distances on a manifold [2]:

1. We will show that not just a pointset, but an *entire* subspace can be preserved by a random projection.

2. We will show that distances between points within a small region of the manifold, can be approximated by a subspace, and thus are well preserved.

3. By taking a ϵ-net of suitable resolution over the manifold, distances between points that are far away are also well preserved.

We can now provide the results in detail[2]. We will start by defining one extra piece of notation which would help our discussion.

Definition 5 ([26]). *The condition number of a manifold M is $\frac{1}{\tau}$, if the normals of length $r < \tau$ at any two distinct points $p, q \in M$ don't intersect.*

Look at figure 4 to see the normals of a manifold. Notice that long non-intersecting normals are possible only if manifold is relatively flat. Hence the condition number of M gives us a handle on how curvy can M be.

Lemma 6 (Johnson-Lindenstrauss [19], [12]). *For any $0 < \epsilon < 1$ and any integer m, let d be a positive integer such that $d = \Omega\left(\frac{\ln m}{\epsilon^2}\right)$. Then for any set V of m points in $\mathbb{R}^D$, there is a linear map $\Phi : \mathbb{R}^D \mapsto \mathbb{R}^d$ such that for all $x, y \in V$,*

$$(1-\epsilon) \leq \frac{\|\Phi x - \Phi y\|^2}{\|x-y\|^2} \leq (1+\epsilon)$$

A projection onto a random subspace (of d dimensions) will satisfy this with high probability.

Proof. Let $\Phi(x) = \sqrt{\frac{D}{d}} R^T x$, where R is a $D \times d$ Gaussian random matrix with entries $\gamma_{ij} \sim N(0,1)$ i.i.d. Note that $R^T x$ (for a fixed x) is distributed as a Gaussian random vector, and from concentration properties of Gaussians, it follows that

1. $\mathbf{Pr}\left[\|R^T x\|^2 \geq (1+\epsilon)\frac{d}{D}\|x\|^2\right] \leq e^{-\Omega\{d\epsilon^2\}}$

[2] for simplicity of the exposition we only provide a proof sketch here and refer the readers to the original papers for detailed proof arguments.

2. $\mathbf{Pr}\left[\|R^T x\|^2 \leq (1-\epsilon)\frac{d}{D}\|x\|^2\right] \leq e^{-\Omega\{d\epsilon^2\}}$

This immediately implies that (with high probability)

$$\begin{aligned} \|\Phi x - \Phi y\|^2 &= \|\Phi(x-y)\|^2 \\ &= \frac{D}{d}\|R^T(x-y)\|^2 \\ &\leq \frac{D}{d}(1+\epsilon)\frac{d}{D}\|x-y\|^2 \\ &= (1+\epsilon)\|x-y\|^2 \end{aligned}$$

Similarly we can also assert that $\|\Phi x - \Phi y\|^2 \geq (1-\epsilon)\|x-y\|^2$.

Now, requiring this property to hold for all pairwise distances between n points, a simple application of union bound gives the desired result. ▮

Lemma 7 (subspace preservation [1]). *Let L be a n-dimensional affine subspace of $\mathbb{R}^D$. Pick $\epsilon, \delta > 0$ and $d \geq \Omega\left(\frac{n}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$. If Φ is a random subspace of d dimensions, then with probability $> 1-\delta$, we have that for all $x \in L$,*

$$(1-\epsilon)\sqrt{d/D}\|x\| \leq \|\Phi x\| \leq (1+\epsilon)\sqrt{d/D}\|x\|$$

Proof. By linearity of norms, it suffices to prove the result for vectors of length 1. Let V be a $\epsilon/4$-cover of a ball B of radius 1. Note that B can be covered by a $\epsilon/4$-net of size $\leq (12/\epsilon)^n$. Applying lemma 6 from above with distortion $\epsilon/2$, immediately yields for all $v \in V$ (with high probability)

$$(1-\epsilon/2)\|v\|^2 \leq \|\Phi v\|^2 \leq (1+\epsilon/2)\|v\|^2$$

Let A be the smallest number such that $\|\Phi x\| \leq (1+A)\|x\|$ for all $x \in L, \|x\| \leq 1$. Note that

$$\|\Phi x\| \leq \|\Phi v\| + \|\Phi(x-v)\| \leq 1 + \epsilon/2 + (1+A)\epsilon/4$$

Now since A is the smallest such number, we have that $A \leq \epsilon/2 + (1+A)\epsilon/4$ or equivalently $A \leq \frac{3\epsilon/4}{1-\epsilon/4} \leq \epsilon$. Similarly we can obtain a lower bound, yielding the desired result. ▮

Lemma 8 (effects on close-by points [2]). *Suppose $S = M \cap B$, where ball B has radius r. Pick $\delta, \epsilon > 0$ and $d = \Omega\left(\frac{n}{\epsilon^2}\log\frac{1}{\epsilon} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$. If $r \leq \frac{\epsilon\tau}{4}\sqrt{\frac{k}{D}}$ and Φ is a random projection to d dimensions then with probability $> 1-\delta$, for all $x, y \in S$*

$$(1-\epsilon)\sqrt{\frac{d}{D}} \leq \frac{\|\Phi x - \Phi y\|}{\|x-y\|} \leq (1+\epsilon)\sqrt{\frac{d}{D}}$$

Proof. Since we have chosen S small enough, pick any $p \in S$ and consider its tangent space T_p. For any $x \in S$, let $\bar{x} \in \mathbb{R}^d$ be its projection onto T_p and $x^\perp = x - \bar{x}$. Note that for any $x, y \in S$, we have that $\frac{\|x^\perp - y^\perp\|}{\|x-y\|} \leq r/\tau$.

Now by applying subspace preservation lemma to T_p, we have that (with high probability)

$$\begin{aligned} \|\Phi x - \Phi y\| &\leq \|\Phi\bar{x} - \Phi\bar{y}\| + \|\Phi x^\perp - \Phi y^\perp\| \\ &\leq \|\bar{x} - \bar{y}\|\sqrt{\frac{d}{D}}(1+\epsilon/2) + \|x^\perp - y^\perp\| \\ &\leq \|x-y\|\sqrt{\frac{d}{D}}(1+\epsilon/2) + \|x-y\|r/\tau \\ &\leq \|x-y\|\sqrt{\frac{d}{D}}(1+\epsilon) \end{aligned}$$

Similarly we can bound $\|\Phi x - \Phi y\| \geq \|x - y\|\sqrt{\frac{d}{D}}(1-\epsilon)$, giving us the desired result. ▮

Theorem 9 (manifold preservation [2]). *Suppose M is a compact n-dimensional Riemannian manifold in $\mathbb{R}^D$ with condition number $1/\tau$. Suppose that for all $\epsilon > 0$, M has an ϵ-cover of size $\leq N_0\left(\frac{1}{\epsilon}\right)^n$. Pick any $\epsilon, \delta > 0$ and $d = \Omega\left(\frac{n}{\epsilon^2}\log\frac{D}{\epsilon\tau} + \frac{1}{\epsilon^2}\log\frac{N_0}{\delta}\right)$. Let Φ be a random subspace of d dimensions. Then with probability $> 1-\delta$, for all $x, y \in M$,*

$$(1-\epsilon)\sqrt{\frac{d}{D}} \leq \frac{\|\Phi x - \Phi y\|}{\|x-y\|} \leq (1+\epsilon)\sqrt{\frac{d}{D}}$$

Proof. For $\epsilon_0 = \frac{\epsilon^2\tau}{128}\sqrt{d/D}$, let $\mu_1, \ldots, \mu_N$, be an ϵ_0-cover of M. Note that $N < N_0\left(\frac{1}{\epsilon_0}\right)^n$.

Let B_i be a ball of radius $\frac{\epsilon\tau}{4}\sqrt{\frac{d}{D}}$ centered at μ_i, we can apply lemma 8 to $B_1, \ldots, B_N$, to have distances within B_i be preserved upto $(1 \pm \epsilon)$.

Pick any $x, y \in M$, if $\|x-y\| \leq \epsilon\tau/8\sqrt{d/D}$, then $x, y \in B_i$ and thus the projected distances are preserved.

If $\|x-y\| > \epsilon\tau/8\sqrt{d/D}$, let μ_i and μ_j be their closest representatives. Then

$$\begin{aligned}\|\Phi x - \Phi y\| &\leq \|\Phi\mu_i - \Phi\mu_j\| + \\ &\quad \|\Phi x - \Phi\mu_i\| + \|\Phi y - \Phi\mu_j\| \\ &\leq \|\mu_i - \mu_j\| \sqrt{\frac{d}{D}}(1+\epsilon/2) + \\ &\quad \epsilon_0\sqrt{\frac{d}{D}}(1+\epsilon) + \epsilon_0\sqrt{\frac{d}{D}}(1+\epsilon) \\ &\leq (\|x-y\| + 2\epsilon_0)\sqrt{\frac{d}{D}}(1+\epsilon/2) + \\ &\quad 2\epsilon_0\sqrt{\frac{d}{D}}(1+\epsilon) \\ &\leq \|x-y\|\sqrt{\frac{d}{D}}(1+\epsilon)\end{aligned}$$

Similarly we can find a lower bound, yielding the final result. ▮

2.2 Discussion

Random projection of manifolds was first considered in [2] and the result was later improved in [10]. Note that the methodology of random projections provides a *nonadaptive* dimensionality reduction approach for manifold learning, where the projection map is *independent* of the actual data. The result presented is significant because data-independent projections are rarely seen in manifold learning literature. It is also worth mentioning that the main result on the minimum number of dimensions needed bears a strong resemblance to results seen in the area of Compressed Sensing for encoding sparse vectors (see [15] for more details) and some of the ideas presented in [2] are borrowed from Compressed Sensing literature.

Note that manifold learning practitioners are more interested in geodesic distances (distances along the manifold) rather than the standard Euclidian distances considered in the analysis above. The result of theorem 9 is easily extendible to geodesic distances by considering limits of sums of Euclidian distances [2].

Observe that the result presented here is a worst case analysis; it gives us an estimate of the minimum number of dimensions needed to preserve *all* interpoint distances within factor of $(1 \pm \epsilon)$. It would be interesting to see bound on the number of dimensions needed to preserve distances in an average sense.

3 Laplacian Eigenmaps for simplifying manifold structure

Laplacian Eigenmaps was recently proposed as a simple and intuitive algorithm for providing a low dimensional representation of data lying on a manifold. Like many manifold learning algorithms, it finds a low dimensional representation by performing computations on the adjacency graph of the sampled data. The basic intuition is that the graph constructed from the samples serves as a discrete approximation for the manifold; and inference based on the graph should correspond to desired inference on the underlying manifold. What sets Laplacian Eigenmaps apart is that the choice of weight used in constructing the graph and the subsequent spectral analysis is formally justified as a process which “simplifies” the manifold structure.

In contrast to random projections that explicitly attempts to preserve all pairwise distances, the optimization criterion of Laplacian Eigenmaps only incorporates the condition to preserve local distances. It turns out that the solution to this optimization criterion has a remarkable property of *smoothing* the manifold structure. More precisely, as we will observe in the following sections, this mapping has the property to reduce the curvature of high-curvature regions, transforming the manifold into a smoother, more manageable object.

3.1 Desirability of simple structure

As mentioned before, Laplacian Eigenmaps provide a non-linear mapping that, in essence, smooths out high curvature regions of the manifold. The power and success of such a mapping comes from noting that such regions can be thought as eccentricities in the collected data. Thus smoothing out these regions should provide a good generalization ability on manifolds.

Consider, for instance, a typical machine learning task of discriminating two classes on a manifold. Due to the inherent curvy manifold structure, it is difficult to find a simple classifier that can separate the classes. However, by first mapping the data via Laplacian Eigenmaps, one can find a simple classifier that can separate the classes well. See figure 5.

3.2 Geometric derivation of Laplacian Eigenmaps

Suppose we want to map M to a line such that nearby points get mapped close together. Let $f : M \to \mathbb{R}$ be a such a map. Then for any $x \in M$ and y in the

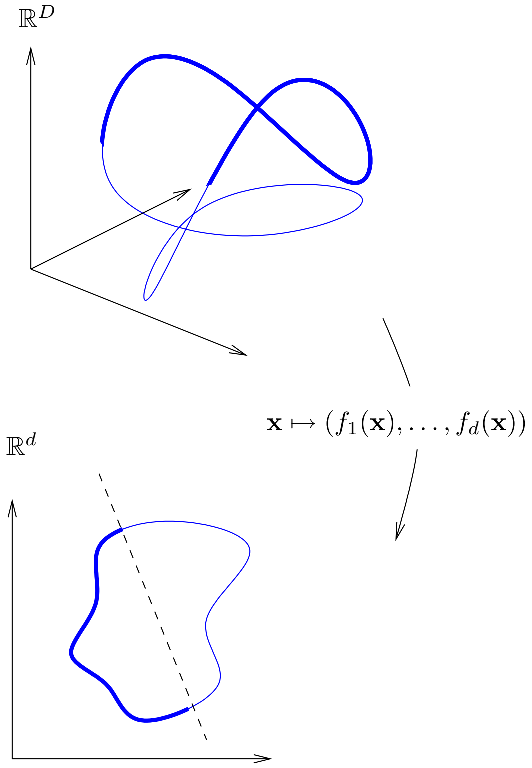
$\mathbb{R}^D$
$\mathbf{x} \mapsto (f_1(\mathbf{x}), \ldots, f_d(\mathbf{x}))$
$\mathbb{R}^d$

Figure 5: Laplacian eigenmaps maps a manifold of complex structure to a relatively simpler structure. This is beneficial for many learning tasks; the task of discriminating two classes on a manifold becomes easier, for instance.

neighborhood of x, we would like to have $|f(x)-f(y)|$ be bounded in terms of the original geodesic distance $d_M(x, y)$. Let $l = d_M(x, y)$, then using the Taylor expansion around x

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

Thus $\|\nabla f(x)\|$ provides us with an estimate of how far apart does f map nearby points. Hence in order to preserve distances, one should look for a map f that minimizes this quantity over all $x \in M$. One sensible minimizing criterion (in "sum-squared" sense) is $\arg\min_{\|f\|=1} \int_M \|\nabla f(x)\|^2$.

Note that $\int \|\nabla f(x)\|^2 = \langle \nabla f, \nabla f \rangle = \langle f, \Delta f \rangle$, where $\Delta = \nabla^2$ is defined to be the Laplace-Beltrami operator on $f(x)$. Thus minimizing this objective function is same as minimizing $\int_M f \Delta f$. Notice that this quantity has the same functional form as the Rayleigh quotient (with $\|f\|^2 = 1$). Hence the problem reduces to finding the eigenfunctions corresponding to the lowest eigenvalues of Δ [3].

This argument can be generalized for mappings to $\mathbb{R}^d$. For a compact M, the optimal d-dimensional embedding is given by the map $x \mapsto (f_1(x), \ldots, f_d(x))$, where f_i is the eigenfunction corresponding to the i^{th} lowest (non-zero) eigenvalue of Δ [3].

3.2.1 Laplace as a smoothness functional

Note that Δ also has the desirable property of being a smoothness functional [25]. Smoothness of a function f over, say, a unit circle S^1 can be defined as $S(f) := \int_{S^1} |f(x)'|^2 dx$. Then functions for which $S(f)$ is close to zero are considered smooth. Note that constant functions over S^1 are clearly smooth. In general, for any $f : M \to \mathbb{R}$,

$$S(f) := \int_M \|\nabla f(x)\|^2 dx = \int_M f \Delta f dx = \langle \Delta f, f \rangle_{L^2(M)}$$

Observe that the smoothness of a unit norm eigenfunction e_i of Δ is controlled by the corresponding eigenvalue λ_i, since $S(e_i) = \langle \Delta e_i, e_i \rangle = \lambda_i$. Therefore, approximating a function f in terms of its first d eigenfunctions of Δ is a way of controlling its smoothness.

So far we have established that spectrum of Δ of a manifold M, provides us with a desirable mapping of M. However, since we just have samples from M, we need a way to approximate Δ.

3.2.2 The graph Laplacian

Graph Laplacian is considered as a *discrete approximation* to the Laplace-Beltrami operator introduced in the previous section.

Let $x_1, \ldots, x_m$ be an independent sample from the uniform distribution over M and let t be a free parameter (optimized later). We can then construct a completely connected weighted undirected graph with samples as the vertices and edge weights w_{ij} as $e^{-\|x_i - x_j\|^2/4t}$. The corresponding graph Laplacian operator is given by the matrix [6]:

$$(L^t_m)_{ij} = \begin{cases} -w_{ij} & \text{if } i \neq j \\ \sum_k w_{ik} & \text{otherwise} \end{cases}$$

We may think of it as an operator on functions on points from the manifold. Let $p \in M$, and $f : M \mapsto \mathbb{R}$, then

$$L^t_m f(p) = f(p) \frac{1}{m} \sum_j e^{\frac{-\|p - x_j\|^2}{4t}} - \frac{1}{m} \sum_j f(x_j) e^{\frac{-\|p - x_j\|^2}{4t}}$$

We can now relate L^t_m to Δ_M[3] for any function f on M.

[3] For conciseness we will denote Δ operator on M as Δ_M.

3.2.3 Connecting together

Let $L^t f(p)$ be the continuous approximation of the graph Laplacian operator defined by

$$L^t f(p) := f(p) \int_M e^{\frac{-\|p-y\|^2}{4t}} d\nu y - \int_M f(y) e^{\frac{-\|p-y\|^2}{4t}} d\nu y$$

We can show that L^t_m is a functional approximation to Δ_M. The proof outline goes as follows ([5], [6]). For a fixed $p \in M$, and a smooth map f, (note that all statements are pointwise in p and f)

1. We will first deduce that L^t_m converges to L^t. This follows almost immediately from law of large numbers.
2. We will relate L^t to Δ_M by
   (a) Restricting the L^t integral to a small ball in M. This would help us express the L^t in a single local coordinate system.
   (b) By applying change of coordinates, we can express L^t as an integral in $\mathbb{R}^n$.
   (c) Finally we will relate the new integral in $\mathbb{R}^n$ to Δ_M.

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 (see [6] for details).

Lemma 10 (continuous approximation of L^t_m [6])**.** *Let L^t_m and L^t be defined as above, then for any $\epsilon > 0$*

$$\mathbf{Pr}\left[|L^t_m f(p) - L^t f(p)| > \epsilon\right] < 2e^{\Omega\{\epsilon^2 m\}}$$

Proof. Note that since L^t_m is the empirical average of m independent samples sampled uniformly from M and L^t is its expectation. Since M is compact, we can use Hoeffding's inequality to bound the deviation, giving the result. ▮

Lemma 11 (restricting L^t to local coordinates [6])**.** *Let $B \subset M$ be a sufficiently small open ball containing p such that B can be expressed in a single chart. For any $a > 0$, as $t \to 0$,*

$$\left| L^t f(p) - \int_B e^{-\frac{\|p-y\|^2}{4t}} (f(p) - f(y)) dy \right| = o(t^a)$$

Proof. For any point $x \in M - B$, let $d = \inf_{x \in M-B} \|p - x\|^2$. Note that $d > 0$ (since B is open). Hence the total contribution of such points to the integral is bounded by $Ce^{-d^2/4t}$ for some constant C. Note that as t tends to zero, this term decreases exponentially, giving the desired result. ▮

Since we have restricted the integral to a small enough ball, we can now use the local coordinate system around an open neighborhood of p. We can apply the canonical change of coordinates by using the exponential map $\exp_p : T_pM (\cong \mathbb{R}^n) \mapsto M$, that carries radial lines from $\mathbf{0}$ in T_pM into geodesics starting at p in M. Note that $\exp_p(0) = p$.

To reduce the computations to $\mathbb{R}^n$, any $y \in M$ (in neighborhood of p) can be written as $\exp_p(x)$ for some $x \in T_pM$. Let $\bar{f}(x) := f(\exp_p(x))$. Then a key fact about Laplace-Beltrami operator is that $\Delta_M f(p) = \Delta_{\mathbb{R}^n} \bar{f}(0) = -\sum_i \frac{\partial^2 \bar{f}}{\partial x_i^2}(0)$. Hence we can analyze L^t in Euclidian space via the (inverse) exponential map [6]:

$$L^t = \frac{1}{\text{Vol}(M)} \int_{\bar{B}} e^{-\frac{\|x\|^2}{4t}} (\bar{f}(0) - \bar{f}(x))(1 + O(\|x\|^2)) dx$$

Using Taylor approximation about 0, we have that:

$$\bar{f}(x) - \bar{f}(0) = x \nabla \bar{f} + \frac{1}{2} x^T H x + O(\|x\|^3)$$

Hence for functions with bounded third order derivatives and letting $t \to 0$, we have that (see [6] for details)

$$\begin{aligned} L^t f(p) &= \frac{-1}{\text{Vol}(M)} \int_{\bar{B}} \left(x \nabla \bar{f} + \frac{1}{2} x^T H x \right) e^{\frac{-\|x\|^2}{4t}} dx \\ &= \frac{-tr(H)}{\text{Vol}(M)} = -\frac{1}{\text{Vol}(M)} \sum \frac{\partial^2 \bar{f}(0)}{\partial x_i^2} \end{aligned}$$

Combining above lemmas immediately yields the main result

Theorem 12 (relating L^t to Δ_M [6])**.** *Let L^t and Δ_M be as defined above. Then for any $p \in M$ and any smooth function f with bounded third order derivative, if $t \to 0$ sufficiently fast, then*

$$L^t f(p) = \frac{1}{\text{Vol}(M)} \Delta_M f(p)$$

3.3 A practical algorithm

As seen in previous sections, Laplacian Eigenmaps have a sound mathematical basis for simplifying data representation. [3] gives a practical algorithm for embedding the data in lower dimensions using this technique. Let $X = x_1, \ldots, x_m \in \mathbb{R}^D$ an independent sample drawn uniformly at random from M, d be the embedding dimension and t be the bandwidth parameter,

Algorithm 3.1: Laplacian Eignmaps (X, d, t)

```
1. Let $W_{ij} = \begin{cases} e^{-\|x_i - x_j\|^2/t} & \text{if } x_i \text{ and } x_j \text{ are close} \\ 0 & \text{otherwise} \end{cases}$
```

2. Let $L = A - W$, where A is a diagonal matrix $A_{ii} = \sum_j W_{ji}$.

3. Compute eigenvectors and eigenvalues for generalized eigenvector problem $L\mathbf{f} = \lambda B\mathbf{f}$. Let the solutions be column vectors of $\mathbf{F}$.

4. **return** $[\mathbf{F}]_{D\times d}$ eigenvectors corresponding to lowest non-zero eigenvalues.

This algorithm has been applied successfully to real-world datasets in [25], giving promising results.

3.4 Discussion

Laplacian Eigenmaps provide a sound low-dimensional representation of a manifold, which has the benefit of simplifying its structure. The corresponding algorithm presented here is simple and intuitive - it requires a few computations and one eigenvalue problem making it quite appealing.

One major limitation of the result presented is that points are sampled i.i.d. from the uniform measure over the manifold. In general, one would like to relax this condition and this problem is still open.

[25] exploits the fact that the embedding simplifies the structure of the manifold for semi-supervised learning on data generated from manifolds. They also show an improvement in classification accuracy for certain real-world datasets.

As discussed, Laplace-Beltrami operator Δ provides a good measure of smoothness, [7] and [25] have used this fact to develop a theory of regularization of functions on a manifold.

Just like the spectrum of Laplace-Beltrami operator yields a smoother representation of a manifold, it would be interesting to study what conditions are optimized if we explore a different basis for functions on a manifold. For instance, the benefits of approximating functions on a manifold using the Lagrange basis (for polynomials) or the Fourier basis (for square integrable functions) is largely unexplored.

4 Kernel methods for manifold density estimation

Many manifold learning methods rely heavily on having independent samples from uniform distribution on M. However, in general, we can't expect such restrictive conditions on the underlying density. Eventhough the analysis of many procedures in the non-uniform setting largely remains an open problem, we do, however, have a handle on estimating the underlying density from independent samples via the method of kernels.

Since we would like to make fewest possible assumptions on the underlying density, we will focus on nonparametric density estimation techniques in this section. We refer the readers to [14], [13], and [31] for an excellent treatment of the subject.

4.1 Density estimation

Density estimation is an important problem in statistics and machine learning. Here the goal is to estimate the underlying density from an i.i.d. sample. Let f be the true density and $\hat{f}_m$ be our estimate (from m samples). Note that we will make little assumptions about the structural form of f.

Given f and $\hat{f}_m$, we can evaluate the quality of our estimate by looking at the associated deviation (called the risk) of $\hat{f}_m$ from f. One popular way to analyze risk is by looking at the expected squared difference between the true density and our estimate. Thus, risk can be defined as

$$R = \mathbb{E}\int(\hat{f}_m(x) - f(x))^2 dx$$

Of course for any reasonable estimator, as the sample size gets larger, one would expect the risk to go down. Here we are interested in studying how fast does the risk go to zero for different estimators.

One intuitive estimator which works well in low dimensions is the histogram estimate. The idea is that we can grid the space and count the relative frequency of points falling into each bin. Though quite intuitive, histograms have their share of disadvantages which make them quite unappealing [31]. Primarily due to sharp boundary between adjacent bins, histogram estimates are not smooth. Moreover the estimator $\hat{f}_m$ is heavily dependent on the placement of the grid; by slightly moving the grid, one can get a wildly different estimator.

This motivates the study of kernel density estimators which largely alleviates these problems by giving smooth approximation to the underlying density, and doesn't suffer from choice of grid placement.

4.1.1 Kernel density estimation

As mentioned in previous section, kernel density estimation provides an attractive alternative to naive histogram estimates, that works well in practice.

The basic idea behind kernel estimate is as follows. To remove the dependence on the grid edges, kernel estimators center a "kernel function" at each sampled data point. By placing a smooth kernel function, the resulting estimator will have a smooth density estimate.

More formally, let K be a kernel function. That is, a smooth function with the properties:

1. Non-negative: $K(x) \geq 0$
2. Integrates to one: $\int K(x)dx = 1$
3. Zero mean: $\int xK(x)dx = 0$, and
4. Finite variance: $\int x^2 K(x)dx < \infty$
5. Maximum at zero: $\sup_x K(x) = K(0)$

Then a kernel density estimate on a sample $x_1, \ldots, x_m$ sampled independently from a fixed underlying distribution on $\mathbb{R}^D$ is given by

$$\hat{f}_{m,K}(x) = \frac{1}{mh^D} \sum_{i=1}^{m} K\left(\frac{\|x - x_i\|}{h}\right)$$

where h is the bandwidth parameter dependent on the number of samples. It is easy to check the $f_{m,K}$ is a well defined density function.

It is known that the quality of the kernel estimate is particularly sensitive to the value of the bandwidth parameter h and less on the form of K [31]. Hence, the choice of bandwidth is important for a good approximation. See figure 6 to see how the changes in the bandwidth result in varying approximations to the underlying density. Small values of h lead to spiky estimates (without much smoothing) while larger h values lead to oversmoothing.

For the optimal choice of bandwidth, the risk decreases as $O(m^{-4/(4+D)})$ (see [31] for details). Note that due to the exponential dependence on D, the quality of the estimate decreases sharply with increase in the dimension; we require exponential number of points to get the same level of accuracy in high dimensions. This is generally referred as the curse of dimensionality.

In the context of manifolds, one would hope that since the manifold occupies a small fraction of the entire ambient space, better convergence rates should be possible. We will study this next.

4.2 Manifold density estimation using Kernels

For a curvy object such as a manifold, we need to define a modified version of the kernel density estimator [27]:

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

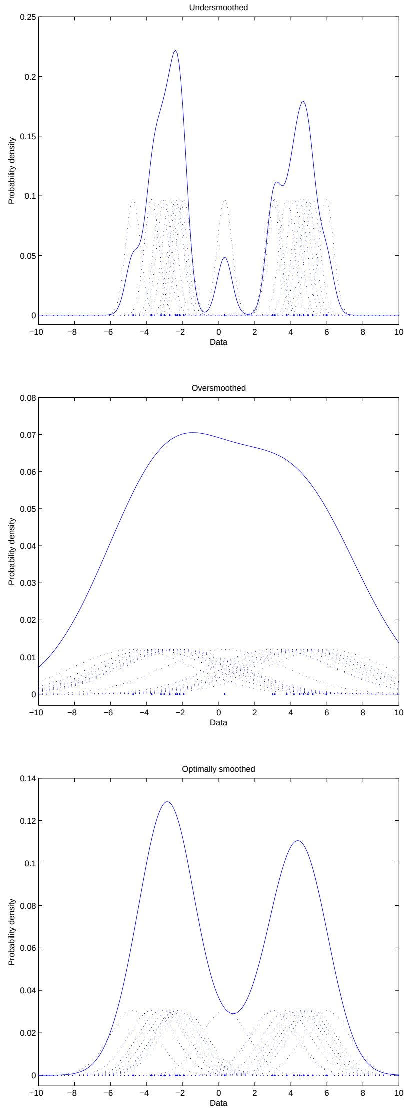

Figure 6: Kernel density estimate of one dimensional data generated from a mixture of two Gaussians. For a fixed independent sample of size 20, and using the Gaussian kernel function (dotted curves), we see that different choices of bandwidth yield significantly different kernel estimators (solid line). Top figure shows the effect of small bandwidths, middle figure shows the effect of large bandwidth, and the bottom figure shows the choice of optimal bandwidth. Note that the optimal bandwidth recovers that the underlying density is in fact a mixture of two Gaussians.

where $d_M(p,q)$ is the geodesic distance between $p,q \in M$ and $\theta_p(q)$ is the volume density function on M defined as $\mathcal{K}(\exp_p^{-1}(q))$ ($\mathcal{K}$ is the ratio of canonical measure of Riemannian metric on T_pM to Lebesgue measure of Euclidian metric on T_pM).

Note that this estimator is a well defined probability density. We will be able to relate it to the underlying true density by [27]:

1. Separately bounding the squared bias and the variance of the estimator. To do this, we will apply change of coordinates to express the integral in $\mathbb{R}^n$.
2. Decomposing the expected risk to its bias and variance components. We can then apply the calculated bounds, yielding the optimal convergence rates.

Lemma 13 (bounding the squared bias [27])**.** *Let f be a probability density on M and $\hat{f}_{m,K}$ be its estimator. If f is square integrable with bounded second covariant derivative, then there exists a constant C_1, such that*

$$\int_M \left(\mathbb{E}\hat{f}_{m,K}(p) - f(p)\right)^2 dp \le C_1 h^4$$

Proof. Consider the pointwise bias,

$$\begin{aligned} b(p) &= \mathbb{E}\hat{f}_{m,K}(p) - f(p) \\ &= \int_{q\in M} \frac{1}{\theta_q(p)h^n} K\left(\frac{d_M(p,q)}{h}\right) f(q)dq - f(p) \\ &= \int_{x\in T_pM} \frac{1}{h^n} K\left(\frac{\|x\|}{h}\right) O(\|x\|^2)dx \end{aligned}$$

Where the last step is by applying change of coordinates via the canonical exponential map $\exp_p : T_pM \mapsto M$, and applying Taylor approximation around 0, $f(\exp_p(x)) =: \bar{f}(x) = \bar{f}(0) + x\nabla\bar{f}(0) + O(\|x\|^2)$. Hence, by applying change of variables $y = x/h$:

$$\begin{aligned} \int_M b^2(p)dp &\le Ch^4 \left(\int_M \|y\|^2 K(\|y\|)dy\right)^2 \left(\int_M dp\right) \\ &\le C'h^4 \operatorname{Vol}(M) \end{aligned}$$

▮

Lemma 14 (bounding the variance [27])**.** *Let f and $\hat{f}_{m,K}$ be defined as above. Then there exists a constant C_2, such that*

$$\int_M \operatorname{Var} \hat{f}_{m,K}(p)dp \le C_2 \frac{1}{mh^n}$$

Proof. Since $\operatorname{Var}(X) \le \mathbb{E}X^2$, we have that for any $p \in M$,

$$\begin{aligned} \operatorname{Var} \hat{f}_{m,K}(p) &\le \frac{1}{mh^{2n}} \mathbb{E} \frac{1}{\theta_{x_1}^2(p)} K^2 \left(\frac{d_M(p,x_1)}{h}\right) \\ &= \frac{1}{mh^{2n}} \int_M \frac{f(q)}{\theta_q^2(p)} K^2 \left(\frac{d_M(p,q)}{h}\right) dq \end{aligned}$$

Integrating both sides over the entire M, we have that $\int_M \operatorname{Var} \hat{f}_{m,K}(p)dp$ is

$$\begin{aligned} &\le \int_{p\in M} \frac{1}{mh^{2n}} \int_{q\in M} \frac{f(q)}{\theta_q^2(p)} K^2 \left(\frac{d_M(p,q)}{h}\right) dqdp \\ &\le \frac{1}{mh^{2n}} K^2(0) \int_{q\in M} f(q) \int_{p\in M} \frac{1}{\theta_q^2(p)} dpdq \\ &\le \frac{Ch^n \operatorname{Vol}(S^n)}{mh^{2n}} K^2(0) \int_{q\in M} f(q)dq \end{aligned}$$

where last inequality is by letting $C = \sup_{p,q} \theta_q^{-1}(p)$ and noting $\int 1/\theta_q(p) = h^n \operatorname{Vol}(S^n)$. The desired result follows. ▮

Theorem 15 (kernel density estimation on manifolds [27])**.** *Let M be a compact n dimensional Riemannian manifold in $\mathbb{R}^D$, and let f, $\hat{f}_{m,k}$ be defined as above, then there exists a constant C such that*

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

Proof. By doing the standard bias-variance decomposition, we have that

$$\begin{aligned} \mathbb{E}\|\hat{f}_{m,K} - f\|^2 &= \int_M \left(\mathbb{E}\hat{f}_{m,K}(p) - f(p)\right)^2 dp \\ &\quad + \int_M \operatorname{Var}\left(\hat{f}_{m,K}(p)\right) dp \\ &\le C_1h^4 + C_2 \frac{1}{mh^n} \end{aligned}$$

where the last inequality is by applying previous two lemmas, immediately giving the desired result. ▮

Note that as a consequence, setting the bandwidth $h \approx m^{-1/(n+4)}$ results in optimal rate of convergence of $O(m^{-4/(n+4)})$, which is independent of the ambient space dimension D.

4.3 Discussion

Density estimation is a central topic in statistics and machine learning. In case of nonparametric density estimation, convergence rates to the true density are known to be exponential in the dimension. In case of manifolds, since the data is locally diffeomorphic to a smaller subspace, one may expect that a weaker dependence on the ambient space. The result presented here is noteworthy as the number of samples needed to gain desired accuracy is *independent* of the ambient dimension. Note that the exponential dependence on the intrinsic manifold dimension, although still unacceptable, is generally more manageable in a typical machine learning scenario.

[14] argues that one should look at ℓ_1 risk as it is invariant under monotone transformations. It would be interesting to see if these rates can be sharpened in ℓ_1, when the data is known to lie on a manifold. ℓ_2 and ℓ_∞ risks have been considered in [18] using Fourier analysis, though their estimator is not a proper probability density.

5 Conclusion and future work

In this survey we examined how some of the known mathematical techniques can be applied in a new context, when data is assumed to be sampled from a manifold. We observed that the manifold assumption leads to results that are significantly less dependent on the ambient dimension.

We looked at random projections as a *linear* dimensionality reduction procedure on manifolds, and concluded that a projection onto a space of dimension just $\Omega(n \log D)$ can preserve all pairwise distances on a manifold remarkably well. We then focused on analyzing the spectrum of the Laplace-Beltrami operator on functions on a manifold and concluded that the resulting eigenmap has the surprising property of simplifying the manifold structure, making it into a more manageable object. Lastly, we looked at kernel density estimation to estimate high density regions on a manifold and found that sample size needed to get desired accuracy can be made completely independent of the ambient dimension. As we can see, significant progress has been in the area of Manifold Learning in the last few years, though much still remains largely unexplored.

In terms of low dimensional mappings, [23], [24] proved that any Reimannian Manifold can be isometrically embedded in $2n+1$ dimensional Euclidian space. However finding such an embedding by a discrete algorithm still remains a hard open problem.

Note that all techniques mentioned in this survey and elsewhere in the literature crucially dependent on the knowledge of the intrinsic dimension of the manifold. However in a typical machine learning problem this quantity is unknown. Note that a poor estimate of n can render manifold learning methods useless; an underestimate will result in low accuracies and an overestimate will require impractically large sample sizes. Researchers have started looking into estimating the intrinsic dimension using likelihood and binpacking methods ([21], [20]), though further progress is needed for a more unified approach.

Researchers often find the "manifold assumption" (data lying exactly on a smooth manifold) too restrictive. In an attempt to relax this assumption, [11] recently proposed a new viewpoint of analyzing algorithms in terms of local covariance dimension. This framework effectively incorporates data that is not necessarily coming from an underlying manifold, but locally has low dimensional structure in an average sense. Both practical and theoretical analysis of machine learning problems in this promising framework is open.

Acknowledgments

The author would like to thank Sanjoy Dasgupta, Joe Pasquale and Lawrence Saul for insightful comments and suggestions.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. 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. Semi-supervised learning on Riemannian manifolds. *Machine Learning Journal*, 56:209–239, 2004.

[5] M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian based manifold methods. *Conference on Computational Learning Theory*, 2005.

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

[7] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. *International Conference on Artificial Intelligence and Statistics*, 2005.

[8] L. Cayton. Algorithms for manifold learning. *UCSD Technical Report CS2008-0923*, 2008.

[9] Y. Chikuse. *Statistics on special manifolds.* Springer, 2003.

[10] K. Clarkson. Tighter bounds for random projections of manifolds. *Computational Geometry*, 2007.

[11] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *ACM Symposium on Theory of Computing*, 2008.

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

[13] L. Devroye. *A course in density estimation.* Birkhauser Verlag AG, 1987.

[14] L. Devroye and L. Gyorfi. *Nonparametric density estimation: the $L1$ view.* Wiley and Sons, 1984.

[15] D. Donoho. Compressed sensing. 2004.

[16] D. Donoho and C. Grimes. Hessian eigenmaps: locally linear embedding techniques for high dimensional data. *Proc. of National Academy of Sciences*, 100(10):5591–5596, 2003.

[17] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley-Interscience, 2nd edition, 2000.

[18] H. Hendriks. Nonparametric estimation of probability density on a Riemannian manifold using fourier expansions. *Annals of Statistics*, 18(2):832–849, 1990.

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

[20] B. Kégl. Intrinsic dimension estimation using packing numbers. *Neural Information Processing Systems*, 14, 2002.

[21] E. Levina and P. Bickel. Maximum likelihood estimation of intrinsic dimension. *Neural Information Processing Systems*, 17, 2005.

[22] J. Milnor. *Topology from the differential viewpoint.* Univ. of Virginia Press, 1972.

[23] J. Nash. C^1 isometric imbeddings. *Annals of Mathematics*, 56, 1954.

[24] J. Nash. The imbedding problem for Riemannian manifolds. *Annals of Mathematics*, 63, 1956.

[25] P. Niyogi. Manifold regularization and semi-supervised learning: some theoretical analysis. *Technical Report TR-2008-01*, 2008.

[26] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Disc. Computational Geometry*, 2006.

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

[28] S. Rosenberg. *The Laplacian on a Riemannian manifold.* Cambridge University Press, 1997.

[29] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2000.

[30] J. Tenebaum, V. de Silva, and J. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000.

[31] L. Wasserman. *All of nonparametric statistics.* Springer, 2005.