LETTER Communicated by Joshua B. Tenenbaum

Laplacian Eigenmaps for Dimensionality Reduction and Data Representation

Mikhail Belkin
misha@math.uchicago.edu
Department of Mathematics, University of Chicago, Chicago, IL 60637, U.S.A.

Partha Niyogi
niyogi@cs.uchicago.edu
Department of Computer Science and Statistics, University of Chicago, Chicago, IL 60637 U.S.A.

One of the central problems in machine learning and pattern recognition is to develop appropriate representations for complex data. We consider the problem of constructing a representation for data lying on a low-dimensional manifold embedded in a high-dimensional space. Drawing on the correspondence between the graph Laplacian, the Laplace Beltrami operator on the manifold, and the connections to the heat equation, we propose a geometrically motivated algorithm for representing the high-dimensional data. The algorithm provides a computationally efficient approach to nonlinear dimensionality reduction that has locality-preserving properties and a natural connection to clustering. Some potential applications and illustrative examples are discussed.

1 Introduction

In many areas of artificial intelligence, information retrieval, and data mining, one is often confronted with intrinsically low-dimensional data lying in a very high-dimensional space. Consider, for example, gray-scale images of an object taken under fixed lighting conditions with a moving camera. Each such image would typically be represented by a brightness value at each pixel. If there were n^2 pixels in all (corresponding to an $n \times n$ image), then each image yields a data point in $\mathbb{R}^{n^2}$. However, the intrinsic dimensionality of the space of all images of the same object is the number of degrees of freedom of the camera. In this case, the space under consideration has the natural structure of a low-dimensional manifold embedded in $\mathbb{R}^{n^2}$.

Recently, there has been some renewed interest (Tenenbaum, de Silva, & Langford, 2000; Roweis & Saul, 2000) in the problem of developing low-dimensional representations when data arise from sampling a probability distribution on a manifold. In this letter, we present a geometrically

motivated algorithm and an accompanying framework of analysis for this problem.

The general problem of dimensionality reduction has a long history. Classical approaches include principal components analysis (PCA) and multidimensional scaling. Various methods that generate nonlinear maps have also been considered. Most of them, such as self-organizing maps and other neural network–based approaches (e.g., Haykin, 1999), set up a nonlinear optimization problem whose solution is typically obtained by gradient descent that is guaranteed only to produce a local optimum; global optima are difficult to attain by efficient means. Note, however, that the recent approach of generalizing the PCA through kernel-based techniques (Schölkopf, Smola, & Müller, 1998) does not have this shortcoming. Most of these methods do not explicitly consider the structure of the manifold on which the data may possibly reside.

In this letter, we explore an approach that builds a graph incorporating neighborhood information of the data set. Using the notion of the Laplacian of the graph, we then compute a low-dimensional representation of the data set that optimally preserves local neighborhood information in a certain sense. The representation map generated by the algorithm may be viewed as a discrete approximation to a continuous map that naturally arises from the geometry of the manifold.

It is worthwhile to highlight several aspects of the algorithm and the framework of analysis presented here:

- The core algorithm is very simple. It has a few local computations and one sparse eigenvalue problem. The solution reflects the intrinsic geometric structure of the manifold. It does, however, require a search for neighboring points in a high-dimensional space. We note that there are several efficient approximate techniques for finding nearest neighbors (e.g., Indyk, 2000).

- The justification for the algorithm comes from the role of the Laplace Beltrami operator in providing an optimal embedding for the manifold. The manifold is approximated by the adjacency graph computed from the data points. The Laplace Beltrami operator is approximated by the weighted Laplacian of the adjacency graph with weights chosen appropriately. The key role of the Laplace Beltrami operator in the heat equation enables us to use the heat kernel to choose the weight decay function in a principled manner. Thus, the embedding maps for the data approximate the eigenmaps of the Laplace Beltrami operator, which are maps intrinsically defined on the entire manifold.

- The framework of analysis presented here makes explicit use of these connections to interpret dimensionality-reduction algorithms in a geometric fashion. In addition to the algorithms presented in this letter, we are also able to reinterpret the recently proposed locally linear em-

bedding (LLE) algorithm of Roweis and Saul (2000) within this framework.

The graph Laplacian has been widely used for different clustering and partition problems (Shi & Malik, 1997; Simon, 1991; Ng, Jordan, & Weiss, 2002). Although the connections between the Laplace Beltrami operator and the graph Laplacian are well known to geometers and specialists in spectral graph theory (Chung, 1997; Chung, Grigor'yan, & Yau, 2000), so far we are not aware of any application to dimensionality reduction or data representation. We note, however, recent work on using diffusion kernels on graphs and other discrete structures (Kondor & Lafferty, 2002).

- The locality-preserving character of the Laplacian eigenmap algorithm makes it relatively insensitive to outliers and noise. It is also not prone to short circuiting, as only the local distances are used. We show that by trying to preserve local information in the embedding, the algorithm implicitly emphasizes the natural clusters in the data. Close connections to spectral clustering algorithms developed in learning and computer vision (in particular, the approach of Shi & Malik, 1997) then become very clear. In this sense, dimensionality reduction and clustering are two sides of the same coin, and we explore this connection in some detail. In contrast, global methods like that in Tenenbaum et al. (2000), do not show any tendency to cluster, as an attempt is made to preserve all pairwise geodesic distances between points.

  However, not all data sets necessarily have meaningful clusters. Other methods such as PCA or Isomap might be more appropriate in that case. We will demonstate, however, that at least in one example of such a data set (the "swiss roll"), our method produces reasonable results.

- Since much of the discussion of Seung and Lee (2000), Roweis and Saul (2000), and Tenenbaum et al. (2000) is motivated by the role that nonlinear dimensionality reduction may play in human perception and learning, it is worthwhile to consider the implication of the previous remark in this context. The biological perceptual apparatus is confronted with high-dimensional stimuli from which it must recover low-dimensional structure. If the approach to recovering such low-dimensional structure is inherently local (as in the algorithm proposed here), then a natural clustering will emerge and may serve as the basis for the emergence of categories in biological perception.

- Since our approach is based on the intrinsic geometric structure of the manifold, it exhibits stability with respect to the embedding. As long as the embedding is isometric, the representation will not change. In the example with the moving camera, different resolutions of the camera (i.e., different choices of n in the $n \times n$ image grid) should lead to embeddings of the same underlying manifold into spaces of very dif-

ferent dimension. Our algorithm will produce similar representations independent of the resolution.

The generic problem of dimensionality reduction is the following. Given a set $\mathbf{x}_1, \ldots, \mathbf{x}_k$ of k points in $\mathbb{R}^l$, find a set of points $\mathbf{y}_1, \ldots, \mathbf{y}_k$ in $\mathbb{R}^m$ ($m \ll l$) such that $\mathbf{y}_i$ "represents" $\mathbf{x}_i$. In this letter, we consider the special case where $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathcal{M}$ and $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^l$.

We now consider an algorithm to construct representative $\mathbf{y}_i$'s for this special case. The sense in which such a representation is optimal will become clear later in this letter.

2 The Algorithm

Given k points $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in $\mathbb{R}^l$, we construct a weighted graph with k nodes, one for each point, and a set of edges connecting neighboring points. The embedding map is now provided by computing the eigenvectors of the graph Laplacian. The algorithmic procedure is formally stated below.

1. Step 1 (constructing the adjacency graph). We put an edge between nodes i and j if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close." There are two variations:
   (a) ϵ-neighborhoods (parameter $\epsilon \in \mathbb{R}$). Nodes i and j are connected by an edge if $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \epsilon$ where the norm is the usual Euclidean norm in $\mathbb{R}^l$. *Advantages*: Geometrically motivated, the relationship is naturally symmetric. *Disadvantages*: Often leads to graphs with several connected components, difficult to choose ϵ.
   (b) n nearest neighbors (parameter $n \in \mathbb{N}$). Nodes i and j are connected by an edge if i is among n nearest neighbors of j or j is among n nearest neighbors of i. Note that this relation is symmetric. *Advantages*: Easier to choose; does not tend to lead to disconnected graphs. *Disadvantages*: Less geometrically intuitive.
2. Step 2 (choosing the weights).[1] Here as well, we have two variations for weighting the edges:
   (a) Heat kernel (parameter $t \in \mathbb{R}$). If nodes i and j are connected, put
   $$W_{ij} = e^{-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{t}};$$
   otherwise, put $W_{ij} = 0$. The justification for this choice of weights will be provided later.

[1] In a computer implementation of the algorithm, steps 1 and 2 are executed simultaneously.

(b) Simple-minded (no parameters ($t = \infty$)). $W_{ij} = 1$ if vertices i and j are connected by an edge and $W_{ij} = 0$ if vertices i and j are not connected by an edge. This simplification avoids the need to choose t.

3. Step 3 (eigenmaps). Assume the graph G, constructed above, is connected. Otherwise, proceed with step 3 for each connected component. Compute eigenvalues and eigenvectors for the generalized eigenvector problem,

$$L\mathbf{f} = \lambda D\mathbf{f}, \tag{2.1}$$

where D is diagonal weight matrix, and its entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_j W_{ji}$. $L = D - W$ is the Laplacian matrix. Laplacian is a symmetric, positive semidefinite matrix that can be thought of as an operator on functions defined on vertices of G.

Let $\mathbf{f}_0, \ldots, \mathbf{f}_{k-1}$ be the solutions of equation 2.1, ordered according to their eigenvalues:

$$\begin{aligned} L\mathbf{f}_0 &= \lambda_0 D\mathbf{f}_0 \\ L\mathbf{f}_1 &= \lambda_1 D\mathbf{f}_1 \\ &\cdots \\ L\mathbf{f}_{k-1} &= \lambda_{k-1} D\mathbf{f}_{k-1} \\ 0 = \lambda_0 &\leq \lambda_1 \leq \cdots \leq \lambda_{k-1}. \end{aligned}$$

We leave out the eigenvector $\mathbf{f}_0$ corresponding to eigenvalue 0 and use the next m eigenvectors for embedding in m-dimensional Euclidean space:

$$\mathbf{x}_i \rightarrow (\mathbf{f}_1(i), \ldots, \mathbf{f}_m(i)).$$

3 Justification

3.1 Optimal Embeddings. Let us first show that the embedding provided by the Laplacian eigenmap algorithm preserves local information optimally in a certain sense.

The following section is based on standard spectral graph theory. (See Chung, 1997, for a comprehensive reference.)

Recall that given a data set, we construct a weighted graph $G = (V, E)$ with edges connecting nearby points to each other. For the purposes of this discussion, assume the graph is connected. Consider the problem of mapping the weighted graph G to a line so that connected points stay as close together as possible. Let $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ be such a map. A reasonable

criterion for choosing a "good" map is to minimize the following objective function,

$$\sum_{ij}(y_i - y_j)^2 W_{ij},$$

under appropriate constraints. The objective function with our choice of weights W_{ij} incurs a heavy penalty if neighboring points $\mathbf{x}_i$ and $\mathbf{x}_j$ are mapped far apart. Therefore, minimizing it is an attempt to ensure that if $\mathbf{x}_i$ and $\mathbf{x}_j$ are "close," then y_i and y_j are close as well.

It turns out that for any $\mathbf{y}$, we have

$$\frac{1}{2}\sum_{i,j}(y_i - y_j)^2 W_{ij} = \mathbf{y}^T L \mathbf{y}, \tag{3.1}$$

where as before, $L = D - W$. To see this, notice that W_{ij} is symmetric and $D_{ii} = \sum_j W_{ij}$. Thus,

$$\begin{aligned}\sum_{i,j}(y_i - y_j)^2 W_{ij} &= \sum_{i,j}(y_i^2 + y_j^2 - 2y_i y_j) W_{ij} \\ &= \sum_i y_i^2 D_{ii} + \sum_j y_j^2 D_{jj} - 2\sum_{i,j} y_i y_j W_{ij} = 2\mathbf{y}^T L \mathbf{y}.\end{aligned}$$

Note that this calculation also shows that L is positive semidefinite.

Therefore, the minimization problem reduces to finding

$$\underset{\substack{\mathbf{y} \\ \mathbf{y}^T D \mathbf{y} = 1}}{\operatorname{argmin}} \mathbf{y}^T L \mathbf{y}.$$

The constraint $\mathbf{y}^T D \mathbf{y} = 1$ removes an arbitrary scaling factor in the embedding. Matrix D provides a natural measure on the vertices of the graph. The bigger the value D_{ii} (corresponding to the ith vertex) is, the more "important" is that vertex. It follows from equation 3.1 that L is a positive semidefinite matrix, and the vector $\mathbf{y}$ that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$L\mathbf{y} = \lambda D \mathbf{y}.$$

Let $\mathbf{1}$ be the constant function taking 1 at each vertex. It is easy to see that $\mathbf{1}$ is an eigenvector with eigenvalue 0. If the graph is connected, $\mathbf{1}$ is the only eigenvector for $\lambda = 0$. To eliminate this trivial solution, which collapses all vertices of G onto the real number 1, we put an additional constraint of orthogonality and look for

$$\underset{\substack{\mathbf{y}^T D \mathbf{y} = 1 \\ \mathbf{y}^T D \mathbf{1} = 0}}{\operatorname{argmin}} \mathbf{y}^T L \mathbf{y}.$$

Thus, the solution is now given by the eigenvector with the smallest nonzero eigenvalue. The condition $\mathbf{y}^T D\mathbf{1} = 0$ can be interpreted as removing a translation invariance in $\mathbf{y}$.

Now consider the more general problem of embedding the graph into m-dimensional Euclidean space. The embedding is given by the $k \times m$ matrix $Y = [\mathbf{y}_1 \mathbf{y}_2, \ldots, \mathbf{y}_m]$, where the ith row provides the embedding coordinates of the ith vertex. Similarly, we need to minimize

$$\sum_{i,j} \|\mathbf{y}^{(i)} - \mathbf{y}^{(j)}\|^2 W_{ij} = \operatorname{tr}(Y^T L Y),$$

where $\mathbf{y}^{(i)} = [\mathbf{y}_1(i), \ldots, \mathbf{y}_m(i)]^T$ is the m-dimensional representation of the ith vertex. This reduces to finding

$$\operatorname*{argmin}_{Y^T D Y = I} \operatorname{tr}(Y^T L Y).$$

For the one-dimensional embedding problem, the constraint prevents collapse onto a point. For the m-dimensional embedding problem, the constraint presented above prevents collapse onto a subspace of dimension less than $m-1$ (m if, as in one-dimensional case, we require orthogonality to the constant vector). Standard methods show that the solution is provided by the matrix of eigenvectors corresponding to the lowest eigenvalues of the generalized eigenvalue problem $L\mathbf{y} = \lambda D\mathbf{y}$.

3.2 The Laplace Beltrami Operator. The Laplacian of a graph is analogous to the Laplace Beltrami operator on manifolds. In this section, we provide a justification for why the eigenfunctions of the Laplace Beltrami operator have properties desirable for embedding.

Let $\mathcal{M}$ be a smooth, compact, m-dimensional Riemannian manifold. If the manifold is embedded in $\mathbb{R}^l$, the Riemannian structure (metric tensor) on the manifold is induced by the standard Riemannian structure on $\mathbb{R}^l$.

As we did with the graph, we are looking here for a map from the manifold to the real line such that points close together on the manifold are mapped close together on the line. Let f be such a map. Assume that $f\colon \mathcal{M} \to \mathbb{R}$ is twice differentiable.

Consider two neighboring points $\mathbf{x}, \mathbf{z} \in \mathcal{M}$. They are mapped to $f(\mathbf{x})$ and $f(\mathbf{z})$, respectively. We first show that

$$|f(\mathbf{z}) - f(\mathbf{x})| \leq \operatorname{dist}_{\mathcal{M}}(\mathbf{x}, \mathbf{z}) \|\nabla f(\mathbf{x})\| + o(\operatorname{dist}_{\mathcal{M}}(\mathbf{x}, \mathbf{z})). \tag{3.2}$$

The gradient $\nabla f(x)$ is a vector in the tangent space $T\mathcal{M}_x$, such that given another vector $\mathbf{v} \in T\mathcal{M}_x$, $df(\mathbf{v}) = \langle \nabla f(x), \mathbf{v} \rangle_{\mathcal{M}}$.

Let $l = \operatorname{dist}_{\mathcal{M}}(\mathbf{x}, \mathbf{z})$. Let $c(t)$ be the geodesic curve parameterized by length connecting $\mathbf{x} = c(0)$ and $\mathbf{z} = c(l)$. Then

$$f(\mathbf{z}) = f(\mathbf{x}) + \int_0^l df(c'(t))\, dt = f(\mathbf{x}) + \int_0^l \langle \nabla f(c(t)), c'(t) \rangle\, dt.$$

Now by Schwartz inequality,

$$\langle \nabla f(c(t)), c'(t) \rangle \leq \|\nabla f(c(t))\| \, \|c'(t)\| = \|\nabla f(c(t))\|.$$

Since $c(t)$ is parameterized by length, we have $\|c'(t)\| = 1$. We also have $\|\nabla f(c(t))\| = \|\nabla f(\mathbf{x})\| + O(t)$ (by Taylor's approximation). Finally, by integrating, we have

$$|f(\mathbf{z}) - f(\mathbf{x})| \leq l \|\nabla f(\mathbf{x})\| + o(l),$$

where both O and o are used in the infinitesimal sense.

If $\mathcal{M}$ is isometrically embedded in $\mathbb{R}^l$, then $\text{dist}_{\mathcal{M}}(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|_{\mathbb{R}^l} + o(\|\mathbf{x} - \mathbf{z}\|_{\mathbb{R}^l})$ and

$$|f(\mathbf{z}) - f(\mathbf{x})| \leq \|\nabla f(\mathbf{x})\| \, \|\mathbf{z} - \mathbf{x}\| + o(\|\mathbf{z} - \mathbf{x}\|).$$

Thus, we see that $\|\nabla f\|$ provides us with an estimate of how far apart f maps nearby points.

We therefore look for a map that best preserves locality on average by trying to find

$$\underset{\|f\|_{L^2(\mathcal{M})}=1}{\text{argmin}} \int_{\mathcal{M}} \|\nabla f(x)\|^2, \tag{3.3}$$

where the integral is taken with respect to the standard measure on a Riemannian manifold. Note that minimizing $\int_{\mathcal{M}} \|\nabla f(x)\|^2$ corresponds to minimizing $L\mathbf{f} = \frac{1}{2} \sum_{i,j} (f_i - f_j)^2 W_{ij}$ on a graph. Here, $\mathbf{f}$ is a function on vertices, and f_i is the value of $\mathbf{f}$ on the ith node of the graph.

It turns out that minimizing the objective function of equation 3.3 reduces to finding eigenfunctions of the Laplace Beltrami operator $\mathcal{L}$. Recall that

$$\mathcal{L} f \stackrel{def}{=} -\operatorname{div} \nabla(f),$$

where div is the divergence of the vector field. It follows from the Stokes' theorem that $-\operatorname{div}$ and ∇ are formally adjoint operators, that is, if f is a function and $\mathbf{X}$ is a vector field, then $\int_{\mathcal{M}} \langle \mathbf{X}, \nabla f \rangle = -\int_{\mathcal{M}} \operatorname{div}(\mathbf{X}) f$. Thus,

$$\int_{\mathcal{M}} \|\nabla f\|^2 = \int_{\mathcal{M}} \mathcal{L}(f) f.$$

We see that $\mathcal{L}$ is positive semidefinite. f that minimizes $\int_{\mathcal{M}} \|\nabla f\|^2$ has to be an eigenfunction of $\mathcal{L}$. The spectrum of $\mathcal{L}$ on a compact manifold $\mathcal{M}$ is known to be discrete (Rosenberg, 1997). Let the eigenvalues (in increasing order) be $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$, and let f_i be the eigenfunction corresponding to eigenvalue λ_i. It is easily seen that f_0 is the constant function that maps the entire manifold to a single point. To avoid this eventuality, we

require (just as in the graph setting) that the embedding map f be orthogonal to f_0. It immediately follows that f_1 is the optimal embedding map. Following the arguments of the previous section, we see that

$$\mathbf{x} \to (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$$

provides the optimal m-dimensional embedding.

3.3 Heat Kernels and the Choice of Weight Matrix. The Laplace Beltrami operator on differentiable functions on a manifold $\mathcal{M}$ is intimately related to the heat flow. Let $f\colon \mathcal{M} \to \mathbb{R}$ be the initial heat distribution and $u(x, t)$ be the heat distribution at time t ($u(x, 0) = f(x)$). The heat equation is the partial differential equation $(\frac{\partial}{\partial t} + \mathcal{L})u = 0$. The solution is given by $u(x, t) = \int_{\mathcal{M}} H_t(x, y) f(y)$, where H_t is the heat kernel, the Green's function for this partial differential equation. Therefore,

$$\mathcal{L}f(x) = -\mathcal{L}u(x, 0) = -\left(\frac{\partial}{\partial t}\left[\int_{\mathcal{M}} H_t(x, y) f(y)\right]\right)_{t=0}. \tag{3.4}$$

It turns out that in an appropriate coordinate system (exponential, which to the first order coincides with the local coordinate system given by a tangent plane in $\mathbb{R}^l$), H_t is approximately the gaussian:

$$H_t(x, y) = (4\pi t)^{-\frac{m}{2}} e^{-\frac{\|x-y\|^2}{4t}} (\phi(x, y) + O(t)),$$

where $\phi(x, y)$ is a smooth function with $\phi(x, x) = 1$. Therefore, when x and y are close and t is small,

$$H_t(x, y) \approx (4\pi t)^{-\frac{m}{2}} e^{-\frac{\|x-y\|^2}{4t}}.$$

See Rosenberg (1997) for details.

Notice that as t tends to 0, the heat kernel $H_t(x, y)$ becomes increasingly localized and tends to Dirac's δ-function, that is, $\lim_{t\to 0} \int_{\mathcal{M}} H_t(x, y) f(y) = f(x)$. Therefore, for small t from the definition of the derivative, we have

$$\mathcal{L}f(x) \approx \frac{1}{t}\left[f(x) - (4\pi t)^{-\frac{m}{2}} \int_{\mathcal{M}} e^{-\frac{\|x-y\|^2}{4t}} f(y)\, dy\right].$$

If $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are data points on $\mathcal{M}$, the last expression can be approximated by

$$\mathcal{L}f(\mathbf{x}_i) \approx \frac{1}{t}\left[f(\mathbf{x}_i) - \frac{1}{k}(4\pi t)^{-\frac{m}{2}} \sum_{\substack{\mathbf{x}_j \\ 0<\|\mathbf{x}_j-\mathbf{x}_i\|<\epsilon}} e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{4t}} f(\mathbf{x}_j)\right].$$

The coefficient $\frac{1}{t}$ is global and will not affect the eigenvectors of the discrete Laplacian. Since the inherent dimensionality of $\mathcal{M}$ may be unknown, we put

$\alpha = \frac{1}{k}(4\pi t)^{-\frac{m}{2}}$. It is interesting to note that since the Laplacian of the constant function is zero, it immediately follows that $\frac{1}{\alpha} = \sum_{\substack{\mathbf{x}_j \\ 0<\|\mathbf{x}_j-\mathbf{x}_i\|<\epsilon}} e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{4t}}$ and

$$\alpha = \left(\sum_{\substack{\mathbf{x}_j \\ 0<\|\mathbf{x}_j-\mathbf{x}_i\|<\epsilon}} e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{4t}} \right)^{-1}.$$

This observation leads to several possible approximation schemes for the manifold Laplacian. In order to ensure that the approximation matrix is positive semidefinite, we compute the graph Laplacian with the following weights:

$$W_{ij} = \begin{cases} e^{-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{4t}} & \text{if } \|\mathbf{x}_i - \mathbf{x}_j\| < \epsilon \\ 0 & \text{otherwise} \end{cases}.$$

4 Connections to Spectral Clustering

The approach to dimensionality reduction considered in this letter uses maps provided by the eigenvectors of the graph Laplacian and eigenfunctions of Laplace Beltrami operator on the manifold. Interestingly, this solution may also be interpreted in the framework of clustering and has very close ties to spectrally based clustering techniques such as those used for image segmentation (Shi & Malik, 1997), load balancing (Hendrickson & Leland, 1993), and circuit design (Hadley, Mark, & Vanelli, 1992). A closely related algorithm for clustering has been recently proposed by Ng et al. (2002). The approach considered there uses a graph that is globally connected with exponentially decaying weights. The decay parameter then becomes very important. In many high-dimensional problems, the minimum and the maximum distances between points are fairly close, in which case the weight matrix will be essentially nonsparse for any rate of decay.

Here we briefly outline the ideas of spectral clustering. It is often of interest to cluster a set of n objects into a finite number of clusters. Thus, given a set of n objects (visual, perceptual, linguistic, or otherwise), one may introduce a matrix of pairwise similarities between the n objects. It is then possible to formulate a general graph-theoretic framework for clustering as follows. Let $G = (V, E)$ be a weighted graph, and W is the matrix of weights, where the vertices are numbered arbitrarily. The weight W_{ij} associated with the edge e_{ij} is the similarity between v_i and v_j. We assume that the matrix of pairwise similarities is symmetric and the corresponding undirected graph is connected.[2]

[2] If the graph is not connected, there are many algorithms for finding its connected components.

Let us consider clustering the objects into two classes. We wish to divide V into two disjoint subsets A, B, $A \cup B = V$, so that the "flow" between A and B is minimized. The flow is a measure of similarity between the two clusters, and the simplest definition of the flow or cut between A and B is the total weight of the edges that have to be removed to make A and B disjoint:

$$\text{cut}(A, B) = \sum_{u \in A, v \in B} W(u, v).$$

Trying to minimize the $\text{cut}(A, B)$ will favor cutting off weakly connected outliers, which tends to lead to poor partitioning quality. To avoid that problem, a measure on the set of vertices is introduced. The weight of a vertex is its "importance" relative to other vertices,

$$\text{vol}(A) = \sum_{u \in A, v \in V} W(u, v)$$

where $W(u, v)$ is the weight on the edge between u and v.

Shi and Malik (1997), define the normalized cut:

$$\text{Ncut}(A, B) = \text{cut}(A, B) \left(\frac{1}{\text{vol}(A)} + \frac{1}{\text{vol}(B)} \right).$$

The problem, as formulated by Shi and Malik (1997), is to minimize Ncut over all partitions of the vertex set V.[3]

It turns out that the combinatorial optimization problem as stated is NP-hard.[4] However, if we allow relaxation of the indicator functions to real values, the problem reduces to minimizing the Laplacian of the graph, which can be easily computed in polynomial time with arbitrary precision.

Recall that

$$\mathbf{x}^T L \mathbf{x} = \sum_{i,j} (x_i - x_j)^2 w_{ij}.$$

Let, as above, A, B be disjoint subsets of V, $A \cup B = V$, and $a = \text{vol}(A)$, $b = \text{vol}(B)$. Put

$$x_i = \begin{cases} \dfrac{1}{\text{vol}(A)}, & \text{if } V_i \in A \\ -\dfrac{1}{\text{vol}(B)}, & \text{if } V_i \in B \end{cases}.$$

[3] A similar and, perhaps, more geometrically motivated quantity is the Cheeger constant,

$$h_G = \min_{A \subset V} \frac{\text{cut}(A, \bar{A})}{\min(\text{vol}(A), \text{vol}((\bar{A})))},$$

where $\bar{A}$ is the complement of A in V. See Chung (1997) for further reference.

[4] A proof due to Papadimitrou can be found in Shi and Malik (1997).

We have

$$\mathbf{x}^T L\mathbf{x} = \sum_{i,j}(x_i - x_j)^2 w_{ij} = \sum_{V_i \in A, V_j \in B} \left(\frac{1}{a} + \frac{1}{b}\right)^2 cut(A, B).$$

Also,

$$\begin{aligned}\mathbf{x}^T D\mathbf{x} &= \sum_i x_i^2 d_{ii} = \sum_{V_i \in A} \frac{1}{a^2} d_{ii} + \sum_{V_i \in B} \frac{1}{b^2} d_{ii} \\ &= \frac{1}{a^2} vol(A) + \frac{1}{b^2} vol(B) = \frac{1}{a} + \frac{1}{b}.\end{aligned}$$

Thus,

$$\frac{\mathbf{x}^T L\mathbf{x}}{\mathbf{x}^T D\mathbf{x}} = \text{cut}(A, B)\left(\frac{1}{a} + \frac{1}{b}\right) = \text{Ncut}(A, B).$$

Notice that $\mathbf{x}^T D\mathbf{1} = \mathbf{0}$, where $\mathbf{1}$ is a column vector of ones.

The relaxed problem is to minimize $\frac{\mathbf{x}^T L\mathbf{x}}{\mathbf{x}^T D\mathbf{x}}$ under the condition that $\mathbf{x}^T D\mathbf{1} = \mathbf{0}$. Put $\mathbf{y} = D^{1/2}\mathbf{x}$. D is invertible, assuming G has no isolated vertices. Then

$$\frac{\mathbf{x}^T L\mathbf{x}}{\mathbf{x}^T D\mathbf{x}} = \frac{\mathbf{y}^T D^{-1/2} L D^{-1/2}\mathbf{y}}{\mathbf{y}^T\mathbf{y}},$$

where $\mathbf{x} \perp D^{1/2}\mathbf{1}$.

The matrix $\tilde{L} = D^{-1/2} L D^{-1/2}$ is the so-called normalized graph Laplacian. $\tilde{L}$ is symmetric positive semidefinite. Notice that $D^{1/2}\mathbf{1}$ is an eigenvector for $\tilde{L}$ with eigenvalue 0, which is the smallest eigenvalue of $\tilde{L}$. Thus, $\min_{\mathbf{y} \perp D^{1/2}\mathbf{1}} \frac{\mathbf{y}^T \tilde{L}\mathbf{y}}{\mathbf{y}^T\mathbf{y}}$ is achieved when $\mathbf{y}$ is an eigenvector corresponding to the second smallest eigenvalue of $\tilde{L}$. Of course, zero can be a multiple eigenvalue, which happens if and only if G has more than one connected component.

Remark. The central observation to be made here is that the process of dimensionality reduction that preserves locality yields the same solution as clustering. It is worthwhile to compare the global algorithm presented in Tenenbaum et al. (2000) with the local algorithms suggested here and in Roweis and Saul (2000). One approach to nonlinear dimensionality reduction as exemplified by Tenenbaum et al. attempts to approximate all geodesic distances on the manifold faithfully. This may be viewed as a global strategy. In contrast, the local approach presented here (as well as that presented in Roweis & Saul, 2000) attempts only to approximate or preserve neighborhood information. This, as we see from the preceding discussion, may also be interpreted as imposing a soft clustering of the data (which may be converted to a hard clustering by a variety of heuristic techniques). In this sense, the local approach to dimensionality reduction imposes a natural clustering of the data.

5 Analysis of Locally Linear Embedding Algorithm

We provide a brief analysis of the LLE algorithm recently proposed by Roweis and Saul (2000) and show its connection to the Laplacian.

Here is a brief description of their algorithm. As before, one is given a data set $\mathbf{x}_1, \ldots, \mathbf{x}_k$ in a high-dimensional space $\mathbb{R}^l$. The goal is to find a low-dimensional representation $\mathbf{y}_1, \ldots, \mathbf{y}_k \in \mathbb{R}^m$, $m \ll k$.

1. Step 1 (discovering the adjacency information). For each $\mathbf{x}_i$, find the n nearest neighbors in the data set, $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}$. Alternatively, $\mathbf{x}_{i_1}, \ldots, \mathbf{x}_{i_n}$ could be data points contained in an ϵ-ball around $\mathbf{x}_i$.
2. Step 2 (constructing the approximation matrix). Let W_{ij} be such that $\sum_j W_{ij}\mathbf{x}_{i_j}$ equals the orthogonal projection of $\mathbf{x}_i$ onto the affine linear span of $\mathbf{x}_{i_j}$'s. In other words, one chooses W_{ij} by minimizing

$$\sum_{i=1}^{l} \left\| \mathbf{x}_i - \sum_{j=1}^{n} W_{ij}\mathbf{x}_{i_j} \right\|^2$$

   under the condition that $\sum_j W_{ij} = 1$ for each i. Assume that W_{ij}'s are well determined. (Otherwise, as happens, for example, in the case when $n > k + 1$, the authors propose a heuristic that we will not analyze here.)
3. Step 3 (computing the embedding). Compute the embedding by taking eigenvectors corresponding to the k lowest eigenvalues of the matrix,

$$E = (I - W)^T(I - W).$$

   Notice that E is a symmetric positive semidefinite matrix.

E can be thought of as an operator acting on functions defined on the data points. We will now provide an argument that under certain conditions,

$$Ef \approx \frac{1}{2}\mathcal{L}^2 f.$$

Eigenvectors of $\frac{1}{2}\mathcal{L}^2$, of course, coincide with the eigenvectors of $\mathcal{L}$. We develop this argument over several steps:

Step 1: Let us fix a data point $\mathbf{x}_i$. We now show that

$$[(I - W)f]_i \approx -\frac{1}{2}\sum_j W_{ij}(\mathbf{x}_i - \mathbf{x}_{i_j})^T H(\mathbf{x}_i - \mathbf{x}_{i_j}),$$

where f is a function on the manifold (and therefore defined on the data points) and H is the Hessian of f at $\mathbf{x}_i$. To simplify the analysis, the neighbor-

ing points ($\mathbf{x}_{i_j}$'s) are assumed to lie on a locally linear patch on the manifold around $\mathbf{x}_i$.

Consider now a coordinate system in the tangent plane centered at $\mathbf{o} = \mathbf{x}_i$. Let $\mathbf{v}_j = \mathbf{x}_{i_j} - \mathbf{x}_i$. Since the difference of two points can be regarded as a vector with the origin at the second point, we see that $\mathbf{v}_j$'s are vectors in the tangent plane originating at $\mathbf{o}$. Let $\alpha_j = W_{ij}$. Since $\mathbf{x}_i$ belongs to the affine span of its neighbors and by construction of the matrix W, we have

$$\mathbf{o} = \mathbf{x}_i = \sum_j \alpha_j \mathbf{v}_j,$$

where

$$\sum \alpha_j = 1.$$

If f is a smooth function, its second-order Taylor approximation can be written as

$$f(\mathbf{v}) = f(\mathbf{o}) + \mathbf{v}^T \nabla f + \frac{1}{2}(\mathbf{v}^T H \mathbf{v}) + o(\|\mathbf{v}\|^2).$$

Here, $\nabla f = (\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n})^T$ is the gradient, and H is the Hessian, $H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}$ (both evaluated at $\mathbf{o}$). Therefore,

$$[(I - W)f]_i = f(\mathbf{o}) - \sum_j \alpha_j f(\mathbf{v}_j),$$

and using the Taylor approximation for $f(\mathbf{v}_j)$, we have

$$f(\mathbf{o}) - \sum_j \alpha_j f(\mathbf{v}_j) \approx f(\mathbf{o}) - \sum_j \alpha_j f(\mathbf{o}) - \sum_j \alpha_j \mathbf{v}_j^T \nabla f - \frac{1}{2} \sum_j \alpha_j \mathbf{v}_j^T H \mathbf{v}_j.$$

Since $\sum \alpha_j = 1$ and $\sum \alpha_j \mathbf{v}_j = \mathbf{o}$, we see that the first three terms disappear and

$$f(\mathbf{o}) - \sum_j \alpha_j f(\mathbf{v_j}) \approx -\frac{1}{2} \sum \alpha_j \mathbf{v}_j^T H \mathbf{v^j}.$$

Step 2: Now note that if $\sqrt{\alpha_i}\mathbf{v_i}$ form an orthonormal basis (which, of course, is not usually the case), then

$$\sum_j W_{ij} \mathbf{v}_j^T H \mathbf{v}_j = \mathrm{tr}(H) = \mathcal{L} f.$$

More generally, we observe that if $\mathbf{x}$ is a random vector, such that its distribution is uniform on every sphere centered at $\mathbf{x}_i$ (which is true, for example, for any locally uniform measure on the manifold), then the expectation $\mathrm{E}(\mathbf{v}^T H \mathbf{v})$ is proportional to $\mathrm{tr}\, H$.

Indeed, if $\mathbf{e}_1, \ldots, \mathbf{e}_n$ form an orthonormal basis for H corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$, then using the spectral theorem,

$$\mathrm{E}(\mathbf{v}^T H \mathbf{v}) = \mathrm{E}\left(\sum \lambda_i \langle \mathbf{v}, \mathbf{e}_i \rangle^2\right).$$

But since $\mathrm{E}\langle \mathbf{v}, \mathbf{e}_i \rangle^2$ is independent of i, put $\mathrm{E}\langle \mathbf{v}, \mathbf{e}_i \rangle^2 = r$, and the above expression reduces to

$$\mathrm{E}(\mathbf{v}^T H \mathbf{v}) = r\left(\sum_i \lambda_i\right) = r\,\mathrm{tr}(H) = r\mathcal{L}f.$$

Step 3: Putting steps 1 and 2 together, we see that

$$(I - W)^T (I - W) f \approx \frac{1}{2}\mathcal{L}^2 f.$$

LLE attempts to minimize $f^T (I - W)^T (I - W) f$, which reduces to finding the eigenfunctions of $(I-W)^T(I-W)$, which can now be interpreted as trying to find the eigenfunctions of the iterated Laplacian $\mathcal{L}^2$. Eigenfunctions of $\mathcal{L}^2$ coincide with those of $\mathcal{L}$.

6 Examples

We now briefly consider several possible applications of the algorithmic framework developed in this letter. We begin with a simple synthetic example of a "swiss roll" considered in Tenenbaum et al. (2000) and Roweis and Saul (2000). We then consider a toy example from vision with vertical and horizontal bars in a "visual field." We conclude with some low-dimensional representations constructed from naturally occurring data sets in the domains of speech and language.

We do not yet know of a principled way to choose the heat kernel parameter t. However, we conduct experiments on the "swiss roll" data set to demonstrate the effect of t and number of nearest neighbors N on the low-dimensional representation. It is clear that for very large values of N, it is critical to choose t correctly. It seems that choosing a smaller t tends to improve the quality of the representation for bigger but still relatively small N. For small values of N, the results do not seem to depend significantly on t.

In the rest of our experiments, we use the simplest version of the algorithm, $W_{ij} \in \{0, 1\}$ or $t = \infty$, which seems to work well in practice and does not involve a choice of a real-valued parameter.

6.1 A Synthetic Swiss Roll. The data set of 2000 points chosen at random from the swiss roll is shown in Figure 1. The swiss roll is a flat two-dimensional submanifold of $\mathbb{R}^3$. Two-dimensional representations of the swiss roll for different values of parameters N and t are shown in Figure 2.

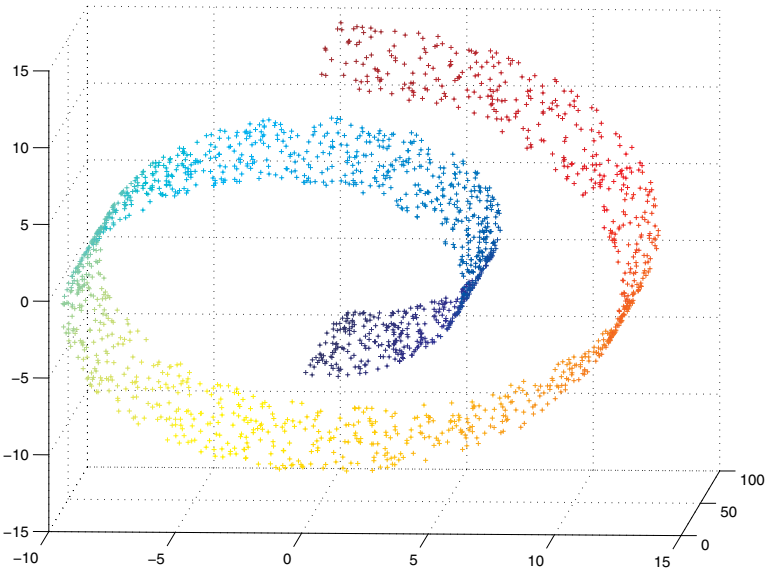

Figure 1: 2000 Random data points on the swiss roll.

Note that $t = \infty$ corresponds to the case when the weights are set to 1. Unlike Isomap, our algorithm does not attempt to isometrically embed the swiss roll into $\mathbb{R}^2$. However, it manages to unroll the swiss roll, thereby preserving the locality, although not the distances, on the manifold. We observe that for small values of N, we obtain virtually identical representations for different t's. However, when N becomes bigger, smaller values of t seemingly lead to better representations.

It is worthwhile to point out that an isometric embedding preserving global distances such as that attempted by Isomap is theoretically possible only when the surface is flat, that is, the curvature tensor is zero, which is the case with the swiss roll. However, a classical result due to gauss shows that even for a two-dimensional sphere (or any part of a sphere), no distance-preserving map into the plane can exist.

6.2 A Toy Vision Example. Consider binary images of vertical and horizontal bars located at arbitrary points in the visual field. Each image contains exactly one horizontal or vertical bar at a random location in the image plane. In principle, we may consider each image to be represented as a function

$$f\colon [0,1] \times [0,1] \to \{0,1\},$$

where $f(\mathbf{x}) = 0$ means the point $\mathbf{x} \in [0,1] \times [0,1]$ is white and $f(\mathbf{x}) = 1$ means the point is black. Let $v(x, y)$ be the image of a vertical bar. Then

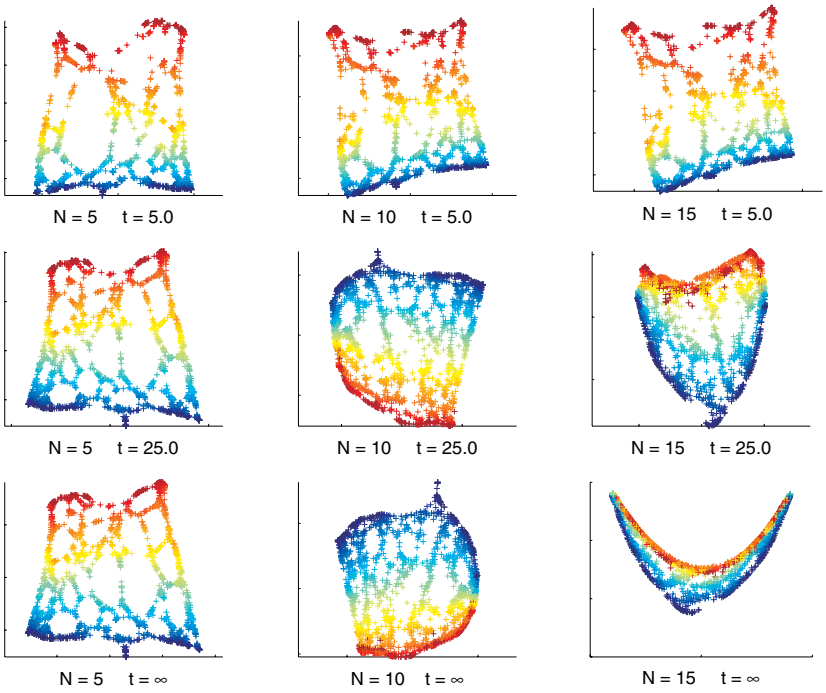

Figure 2: Two-dimensional representations of the swiss roll data, for different values of the number of nearest neighbors N and the heat kernel parameter t. $t = \infty$ corresponds to the discrete weights.

all images of vertical bars may be obtained from $v(x, y)$ by the following transformation:

$$v_t(x, y) = v(x - t_1, y - t_2).$$

The space of all images of vertical bars is a two-dimensional manifold, as is the space of all horizontal bars. Each of these manifolds is embedded in the space of functions ($L^2([0, 1] \times [0, 1])$). Notice that although these manifolds do not intersect, they come quite close to each other. In practice, it is usually impossible to tell whether the intersection of two classes is empty.

To discretize the problem, we consider a 40×40 grid for each image. Thus, each image may be represented as a 1600-dimensional binary vector. We choose 1000 images (500 containing vertical bars and 500 containing horizontal bars) at random. The parameter N is chosen to be 14 and $t = \infty$.

In Figure 3, the left panel shows a horizontal and vertical bar to provide a sense of the scale of the image. The middle panel is a two-dimensional representation of the set of all images using the Laplacian eigenmaps. Notice

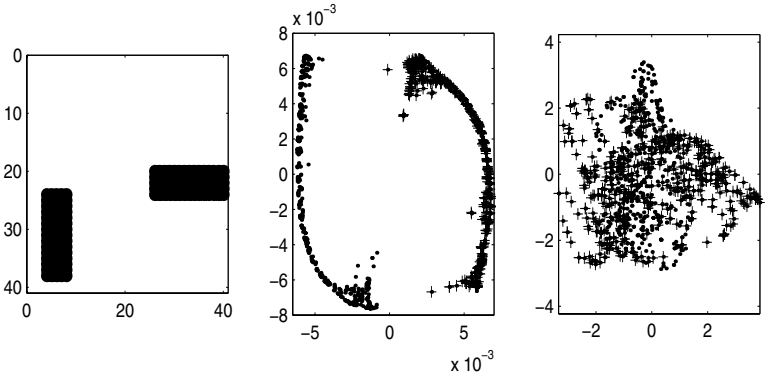

Figure 3: (Left) A horizontal and a vertical bar. (Middle) A two-dimensional representation of the set of all images using the Laplacian eigenmaps. (Right) The result of PCA using the first two principal directions to represent the data. Blue dots correspond to images of vertical bars, and plus signs correspond to images of horizontal bars.

that while the local graph is connected, the two-dimensional representation shows two well-defined components. The right panel shows the result of PCA using the first two principal directions to represent the data.

6.3 A Linguistic Example. An experiment was conducted with the 300 most frequent words in the Brown corpus—a collection of texts containing about 1 million words (not distinct) available in electronic format. Each word is represented as a vector in a 600-dimensional space using information about the frequency of its left and right neighbors (computed from the corpus). More precisely, let the 300 words be w_1 through w_{300}. Then the representation of w_i is a 600-dimensional vector $\mathbf{v}_i$ (say) where the first 300 dimensions of $\mathbf{v}_i$ characterize left neighbor relations and the next 300 characterize right neighbor relations. Thus, $\mathbf{v}_i(j)$ – the jth component ($j \leq 300$) of v_i is the number of times the sequence $w_j w_i$ occurs in the corpus (referred to as the bigram count). Similarly, $\mathbf{v}_i(j + 300)$ is the the count of the number of times the sequence $w_i w_j$ occurs in the corpus.

Thus, there are 300 vectors in $\mathbb{R}^{600}$. Of course, we do not claim that there is a natural low-dimensional manifold structure on these vectors. Nevertheless, it is useful for practical applications to construct low-dimensional representations of this data. For example, the well-known LSI (latent semantic indexing) approach uses PCA to represent the documents in a vector space model for purposes of search and information retrieval. Applying the Laplacian eigenmap with $N = 14; t = \infty$ to the data yields a low-dimensional representation shown in Figures 4 and 5. Note that words belonging to

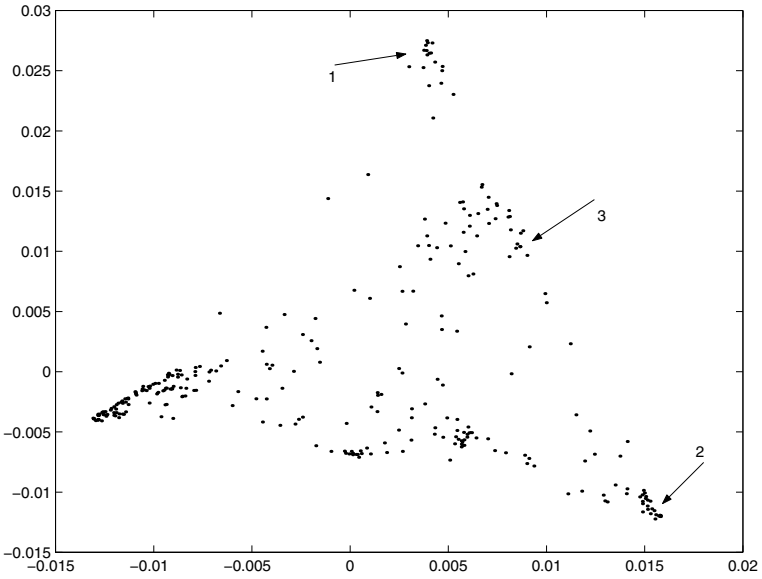

Figure 4: The 300 most frequent words of the Brown corpus represented in the spectral domain.

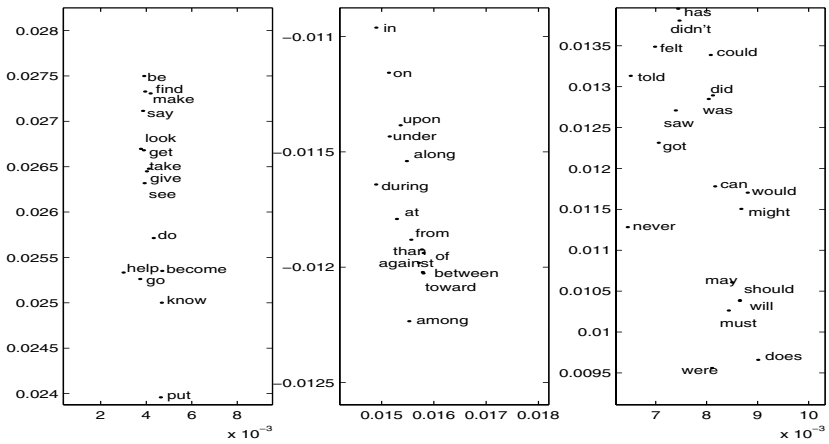

Figure 5: Fragments labeled by arrows: (left) infinitives of verbs, (middle) prepositions, and (right) mostly modal and auxiliary verbs. We see that syntactic structure is well preserved.

similar syntactic categories seem to cluster together, highlighting further the connections between clustering and dimensionality reduction as discussed in this letter.

6.4 Speech. We turn finally to an example from human speech. It has long been recognized that while the speech signal is high dimensional, the distinctive phonetic dimensions are few. An important open question in the field is to develop a low-dimensional representation of the speech signal that is correlated with phonetic content.

In this example, we consider the low-dimensional representations that arise by applying the Laplacian eigenmap algorithm to a sentence of speech sampled at 16 kHz. A short-time Fourier transform (with a 30 ms window) was computed from the speech signal at 5 ms intervals. This yielded a vector of Fourier coefficients for every 30 ms chunk of the speech signal. There were 685 such vectors in all. As a standard practice in speech recognition, the data were represented by the logarithm of these Fourier coefficients. Each vector contained 256 logs of Fourier coefficients. As before, we choose $N = 14; t = \infty$. Furthermore, each vector was labeled according to the identity of the phonetic segment it belonged to. These labels are not utilized by the Laplacian eigenmap algorithm, which finds a low-dimensional representation for the data. Shown in Figure 6 are the speech data points

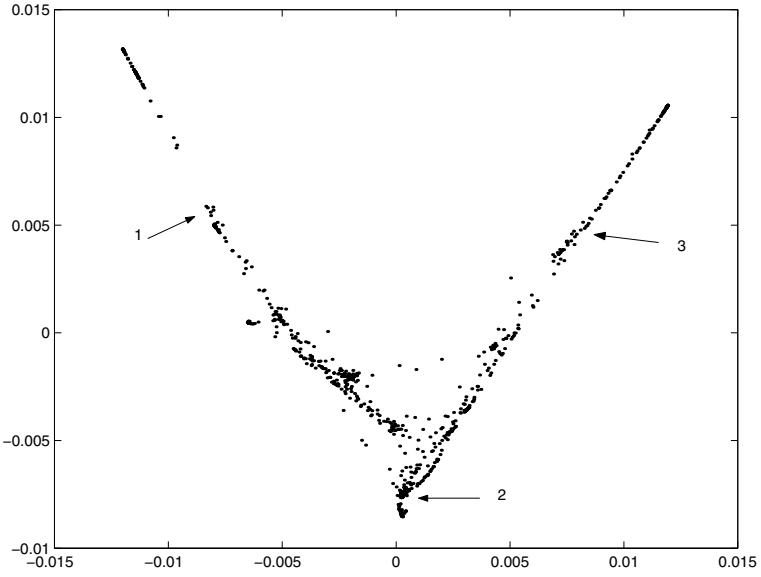

Figure 6: The 685 speech data points plotted in the two-dimensional Laplacian spectral representation.

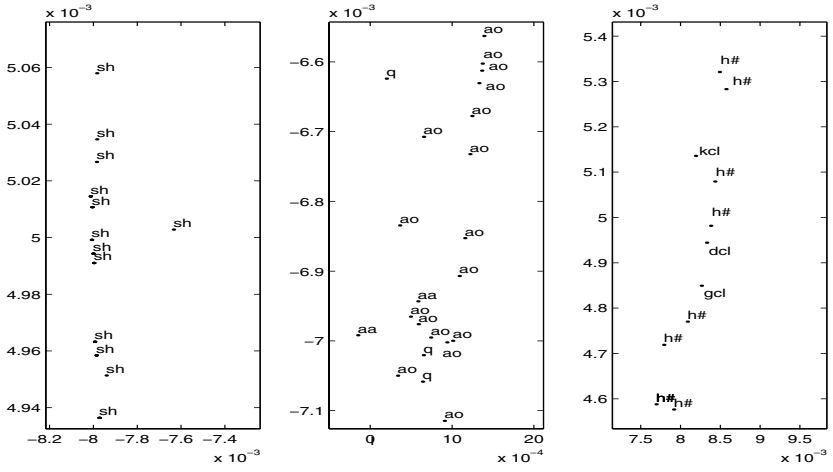

Figure 7: A blowup of the three selected regions corresponding to the arrows in Figure 6. Notice the phonetic homogeneity of the chosen regions. The data points corresponding to the same region have similar phonetic identity, though they may (and do) arise from occurrences of the same phoneme at different points in the utterance. The symbol *sh* stands for the fricative in the word *she*; *aa* and *ao* stand for vowels in the words *dark* and *all*, respectively; *kcl*, *dcl*, and *gcl* stand for closures preceding the stop consonants *k*, *d*, *g*, respectively. *h#* stands for silence.

plotted in the two-dimensional Laplacian representation. The two "spokes" correspond predominantly to fricatives and closures, respectively. The central portion corresponds mostly to periodic sounds like vowels, nasals, and semivowels. A natural clustering into the broad classes is obtained, and Figure 7 shows three different regions of the representation space. Note the phonetic homogeneity of the data points that lie in each of these regions. Points mapped to the same region in the representation space share similar phonetic features, though points with the same label may originate from different occurrences of the same phoneme.

7 Conclusions

In this letter, we introduced a coherent framework for dimensionality reduction for the case where data reside on a low-dimensional manifold embedded in a higher-dimensional space. A number of questions remain to be answered:

- Our approach uses the properties of Laplace Beltrami operator to construct invariant embedding maps for the manifold. Although such

maps have some demonstrable locality-preserving properties, they do not in general provide an isometric embedding. The celebrated Nash's embedding theorem (Nash, 1954) guarantees that an n-dimensional manifold admits an isometric C^1 embedding into a $2n+1$–dimensional Euclidean space.[5] However it remains unclear whether such an embedding is easily computable by a discrete algorithm. Furthermore, there are usually many possible isometric embeddings of a given manifold. For example, any knot in $\mathbb{R}^3$ is an isometric embedding of a circle. However, when the embedded manifold is isometric to a domain in $\mathbb{R}^k$, the canonical embedding is given by the exponential map. In that case, Isomap provides an embedding and guarantees convergence (Bernstein, de Silva, Langford, & Tenenbaum, 2000). In general, it is not clear how to discriminate between "good" and "bad" isometric embeddings. It would therefore be interesting to formulate more precisely what properties of an embedding make it desirable for pattern recognition and data representation problems.

- We have not given any consideration to other geometric invariants of the manifold that may potentially be estimated from data. For example, it is unclear how to estimate reliably even such a simple invariant as the intrinsic dimensionality of the manifold.
- There are further issues pertaining to our framework that need to be sorted out. First, we have implicitly assumed a uniform probability distribution on the manifold according to which the data points have been sampled. Second, it remains unclear how the algorithm behaves when the manifold in question has a boundary. Third, appropriate choices for N (or ϵ) and t and their effect on the behavior of the embeddings need to be better understood. Fourth, the convergence of the finite sample estimates of the embedding maps needs to be addressed.
- Finally, and most intriguing, while the notion of manifold structure in natural data is a very appealing one, we do not really know how often and in which particular empirical contexts the manifold properties are crucial to account for the phenomena at hand. Vastly more systematic studies of the specific problems in different application domains need to be conducted to shed light on this question.

Acknowledgments

We are very grateful to John Goldsmith for motivating us to consider the approach discussed here, to Peter Bickel for many insightful critical comments, and to Yali Amit, Lazslo Babai, Todd Dupont, Joshua Maher, and

[5] The C^1 condition is essential. If the embedding has to be infinitely differentiable, the required dimension is much higher (Nash, 1956).

Ridgway Scott for conversations. Belkin and Niyogi (2002) was an earlier version of this letter.

References

Belkin, M., & Niyogi, P. (2002). Laplacian eigenmaps and spectral techniques for embedding and clustering. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Bernstein, M., de Silva, V., Langford, J. C., & Tenenbaum, J. B. (2000). *Graph approximations to geodesics on embedded manifolds*. Available on-line: http://isomap.stanford.edu/BdSLT.pdf.

Chung, Fan R. K. (1997). *Spectral graph theory*. Providence, RI: American Mathematical Society.

Chung, Fan R. K., Grigor'yan, A., & Yau, S.-T. (2000). Higher eigenvalues and isoperimetric inequalities on Riemannian manifolds and graphs. *Communications on Analysis and Geometry, 8*, 969–1026.

Hadley, S. W., Mark, B. L., & Vanelli, A. (1992). An efficient eigenvector approach for finding netlist partitions. *IEEE Transactions on Computer-Aided Design, 11*(7), 885–892.

Haykin, S. (1999). *Neural networks: A comprehensive foundation*. Upper Saddle River, NJ: Prentice Hall.

Hendrickson, B., & Leland, R. (1993). Multidimensional spectral load balancing. In *Proceedings of the Sixth SIAM Conference on Parallel Processing for Scientific Computing* (pp. 953–961). Philadelphia: SIAM.

Indyk, P. (2000). *Dimensionality reduction techniques for proximity problems*. Paper presented at the Eleventh Symposium on Discrete Algorithms, San Francisco.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the ICML 2002*.

Nash, J. (1954). C^1 isometric imbeddings. *Annals of Mathematics, 56*, 383–396.

Nash, J. (1956). The imbedding problem for Riemannian Manifolds. *Annals of Mathematics, 63*, 20–63.

Ng, A. Y., Jordan, M., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Rosenberg, S. (1997). *The Laplacian on a Riemannian manifold*. Cambridge: Cambridge University Press.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science, 290*, 2323–2326.

Schölkopf, B., Smola, A., & Mülller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10*(5), 1299–1319.

Seung, H. S., & Lee, D. D. (2000). The manifold way of perception. *Science, 290*, 2268–2269.

Shi, J., & Malik, J. (1997). Normalized cuts and image segmentation. *IEEE Conf. Computer Vision and Pattern Recognition* (pp. 731–737).

Simon, H. D. (1991). Partitioning of unstructured problems for parallel processing. *Computing Systems in Engineering, 2*, 135–148.
Tenenbaum, J., de Silva, V., & Langford, J. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science, 290*, 2319–2323.

Received April 16, 2002; accepted November 1, 2002.