Laplacian Eigenmaps for Dimensionality Reduction and Data Representation

M. Belkin, P. Niyogi
Presented by: Alexandru Ionescu

07.05.2019

Introduction

- **Problem:** Constructing a representation for data lying on a low-dimensional manifold embedded in a high-dimensional space.
- **Solution:** An algorithm for representing the high-dimensional data based on the correspondence between the graph Laplacian, the Laplace Beltrami operator on the manifold, and the connections to the heat equation.
- **Advantages:** Locality-preserving properties and a natural connection to clustering.

Example

- A gray-scale image of an object taken under fixed lighting conditions with a moving camera would typically be represented by a brightness value at each pixel.
- For n^2 pixels in an $n \times n$ image, each image yields a data point in $\mathbb{R}^{n^2}$.
- The intrinsic dimensionality of the space of all images of the same object is the number of degrees of freedom of the camera.
- The space under consideration has the natural structure of a low-dimensional manifold embedded in $\mathbb{R}^{n^2}$.

The approach

- The algorithm builds a graph incorporating neighborhood information of the data set.
- Using the notion of the Laplacian of the graph, it computes *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*.
- 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 locality-preserving character of the Laplacian eigenmap algorithm makes it relatively insensitive to outliers and noise; also, it is not prone to short circuiting, as only the local distances are used.
- By trying to preserve local information in the embedding, the algorithm implicitly *emphasizes the natural clusters in the data*. In this sense, dimensionality reduction and clustering are connected.

General framework

- **Generic problem of dimensionality reduction:** Given a set $x_1, \ldots, x_k$ of k in $\mathbb{R}^l$, find a set of points $y_1, \ldots, y_k$ in $\mathbb{R}^m$, with $m \ll l$, such that y_i "represents" x_i.
- Special case for Laplacian eigenmaps algorithm: $x_1, \ldots, x_k \in \mathcal{M}$, where $\mathcal{M}$ is a manifold embedded in $\mathbb{R}^l$.
- $f : \mathcal{M} \to \mathcal{N}$ is called *embedding* of $\mathcal{M}$ in $\mathcal{N}$ if f is an injective immersion (its derivative is injective) which is a homeomorphism onto its image ($f : \mathcal{M} \to f(\mathcal{M})$ and its inverse are continuous functions); *embeddings preserve mathematical structures*.
- The algorithm constructs a weighted graph with k nodes, one for each point, and a set of edges connecting neighboring points. The embedding map is provided by computing the eigenvectors of the graph Laplacian.

The algorithm

1. Constructing the adjacency graph $G = (V, E)$ – $(i, j) \in E$ if x_i and x_j are "close":
   (a) ε–neighborhood: $||x_i - x_j||^2 < \varepsilon$
   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: i is among N nearest neighbors of j or viceversa.
   Advantages: Easier to choose; does not tend to lead to disconnected graphs.
   Disadvantages: Less geometrically intuitive.
2. Choosing the weights:
   (a) Heat kernel: $W_{ij} = e^{-\frac{||x_i - x_j||^2}{t}}$ if nodes i and j are connected and $W_{ij} = 0$ otherwise.
   (b) Simple-minded ($t = \infty$): $W_{ij} = 1$ if nodes i and j are connected by an edge and $W_{ij} = 0$ otherwise.
3. Eigenmaps: compute eigenvalues and eigenvectors for the generalized eigenvector problem

$$Lf = \lambda Df \tag{1}$$

- D – diagonal weight matrix, $D_{ii} = \sum_j W_{ji}$
- $L = D - W$ – Laplacian matrix (symmetric and positive semidefinite)
- $f_0, \ldots, f_{k-1}$ – solutions to equation (1), ordered according to their eigenvalues:

$$Lf_0 = \lambda_0 Df_0$$
$$Lf_1 = \lambda_1 Df_1$$
$$\ldots$$
$$Lf_{k-1} = \lambda_{k-1} Df_{k-1}$$

$$0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{k-1}.$$

- Leave out the eigenvector f_0 and use the next m eigenvectors for embedding in m–dimensional Euclidean space:

$$x_i \mapsto (f_1(i), \ldots, f_m(i)).$$

Optimal Embeddings

- Consider the one-dimensional problem of mapping a weighted graph G to a line such that connected points stay as close together as possible; let $y = (y_1, \ldots, y_n)^T$ be such a map.
- The problem translates into minimizing the objective function

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

which is an attempt to ensure that if x_i and x_j are "close", then y_i and y_j are close as well.

- Since

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

where $L = D - W$, it follows that the minimization problem reduces to finding

$$\arg\min_{\substack{y \\ y^T D y = 1}} y^T L y.$$

- The constraint $y^T D y = 1$ removes an arbitrary scaling factor in the embedding.

- Equation (1) yields that the vector y that minimizes the objective function is given by the minimum eigenvalue solution to the generalized eigenvalue problem

$$Ly = \lambda Dy.$$

- If the graph is connected, then the constant function 1 is the only eigenvector for $\lambda = 0$, such that an additional constraint is neccesary to eliminate the trivial solution and the problem becomes

$$\underset{\substack{y^T Dy=1 \\ y^T D1=0}}{\arg\min}\, y^T Ly.$$

- The solution is now given by the eigenvector with the smallest nonzero eigenvalue.

- Consider now the general problem of embedding the graph into m–dimensional Euclidean space.
- The embedding is given by the $k \times m$ matrix $Y = [y_1, \ldots, y_m]$, where the ith row provides the embedding coordinates of the ith vertex.
- We need to minimize

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

where $y^{(i)} = [y_1(i), \ldots, y_m(i)]^T$ is the m–dimensional representation of the ith vertex.

- The problem reduces to finding

$$\underset{Y^T D Y = I}{\arg\min}\ Y^T L Y,$$

where the constraint prevents collapse onto a subspace of dimension less than $m - 1$ (as for the one-dimensional embedding problem, the constraint prevents collapse onto a point).

The Laplace Beltrami Operator

- Let $\mathcal{M}$ be a smooth, compact, m-dimensional Riemannian manifold.
- We are looking 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 $f : \mathcal{M} \to \mathbb{R}$ be such a map and let f be twice differentiable.
- The *gradient* $\nabla f(x)$ is a vector in the tangent space $T\mathcal{M}_x$ such that $df(v) =< \nabla f(x), v >_{\mathcal{M}}$ for any $v \in T\mathcal{M}_x$.
- It is known that $||\nabla f||$ provides an estimate of how far apart f maps nearby points, that is,

$$|f(z) - f(x)| \leq ||\nabla f(x)||\ ||z - x|| + o(||z - x||).$$

- The problem to find a map that best preserves locality on average means trying to find

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

which corresponds to minimizing $Lf = \frac{1}{2}\sum_{i,j}(f_i - f_j)^2 W_{ij}$ on a graph.

- If $\mathcal{L}$ is the Laplace Beltrami operator defined by $\mathcal{L}f = -\operatorname{div}\nabla(f)$, it is known that
$$\int_{\mathcal{M}} ||\nabla f||^2 = \int_{\mathcal{M}} \mathcal{L}(f)f.$$
- f that minimizes $\int_{\mathcal{M}} ||\nabla f(x)||^2$ has to be an eigenfunction of $\mathcal{L}$.
- Let the eigenvalues of $\mathcal{L}$ be $0 \leq \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots$ and let f_i be the eigenfunction corresponding to eigenvalue λ_i.
- f_0 is the constant function that maps the entire manifold to a single point, therefore we must require that the embedding map f is orthogonal to f_0.
- Hence, the optimal m-dimensional embedding is
$$x \mapsto (f_1(x), \ldots, f_m(x)).$$

Heat Kernel 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.
- $f : \mathcal{M} \to \mathbb{R}$ – the initial heat distribution; $u(x,t)$ – the heat distribution at time t ($u(x,0) = f(x)$).
- *Heat equation:* $\left(\frac{\partial}{\partial t} + \mathcal{L}\right) u = 0$; solution: $u(x,t) = \int_{\mathcal{M}} H_t(x,y) f(y)$, where H_t is the heat kernel; in an appropriate coordinate system:

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

- $\varphi(x,y)$ is a smooth function with $\varphi(x,x) = 1$, hence, for close x and y and small t,

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

- Since $\lim_{t\to 0} \int_{\mathcal{M}} H_t(x,y) f(y) = f(x)$, for small t we have

$$\begin{aligned} \mathcal{L}f(x) = \mathcal{L}u(x,0) &= -\left(\frac{\partial}{\partial t}\left[\int_{\mathcal{M}} H_t(x,y) f(y)\right]\right)_{t=0} \\ &\approx \frac{1}{t}\left[f(x) - (4\pi t)^{-\frac{m}{2}} \int_{\mathcal{M}} e^{-\frac{||x-y||^2}{4t}} f(y) dy\right] \end{aligned}$$

- For data points $x_1, \ldots, x_k$ on $\mathcal{M}$,

$$\mathcal{L}f(x_i) \approx \frac{1}{t}\left[f(x_i) - \frac{1}{k}(4\pi t)^{-\frac{m}{2}} \sum_{0<||x_i - x_j||<\varepsilon} e^{-\frac{||x_i - x_j||^2}{4t}} f(x_j)\right].$$

- Since the inherent dimension of $\mathcal{M}$ may be unknown, we put $\alpha = \frac{1}{k}(4\pi t)^{-\frac{m}{2}}$; for constant functions, the Laplacian is zero, hence

$$\alpha = \left(\sum_{0<||x_i - x_j||<\varepsilon} e^{-\frac{||x_i - x_j||^2}{4t}}\right)^{-1}.$$

- There are many approximation schemes for the manifold Laplacian; for a positive semidefinite approximation matrix, the graph Laplacian will be weighted by

$$W_{ij} = \begin{cases} e^{-\frac{||x_i - x_j||^2}{4t}} & \text{if } ||x_i - x_j|| < \varepsilon \\ 0 & \text{otherwise} \end{cases}.$$

Examples

Synthetic examples.

- Illustrate the effect of t and N on the low-dimensional representation.
- For very large values of N, it is critical to choose t correctly: 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.

Natural data sets.

- The simplest version of the algorithm is adequate: $W_{ij} \in \{0, 1\}$ or $t = \infty$.
- It does not involve the choice of a parameter.

1. Synthetic Swiss Roll

- 2000 points chosen at random from the swiss roll, a flat two-dimensional submanifold of $\mathbb{R}^3$.
- $t = \infty$ corresponds to the case when the weights are set to 1.
- The algorithm preserves the locality, although not the distances, on the manifold.

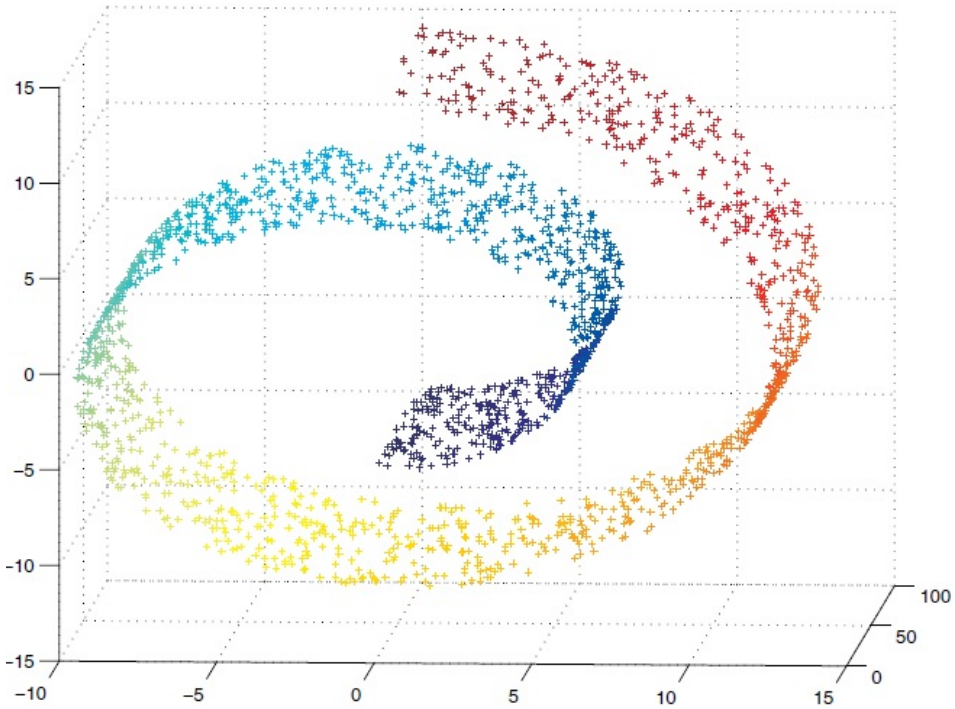

Figure: The swiss roll.

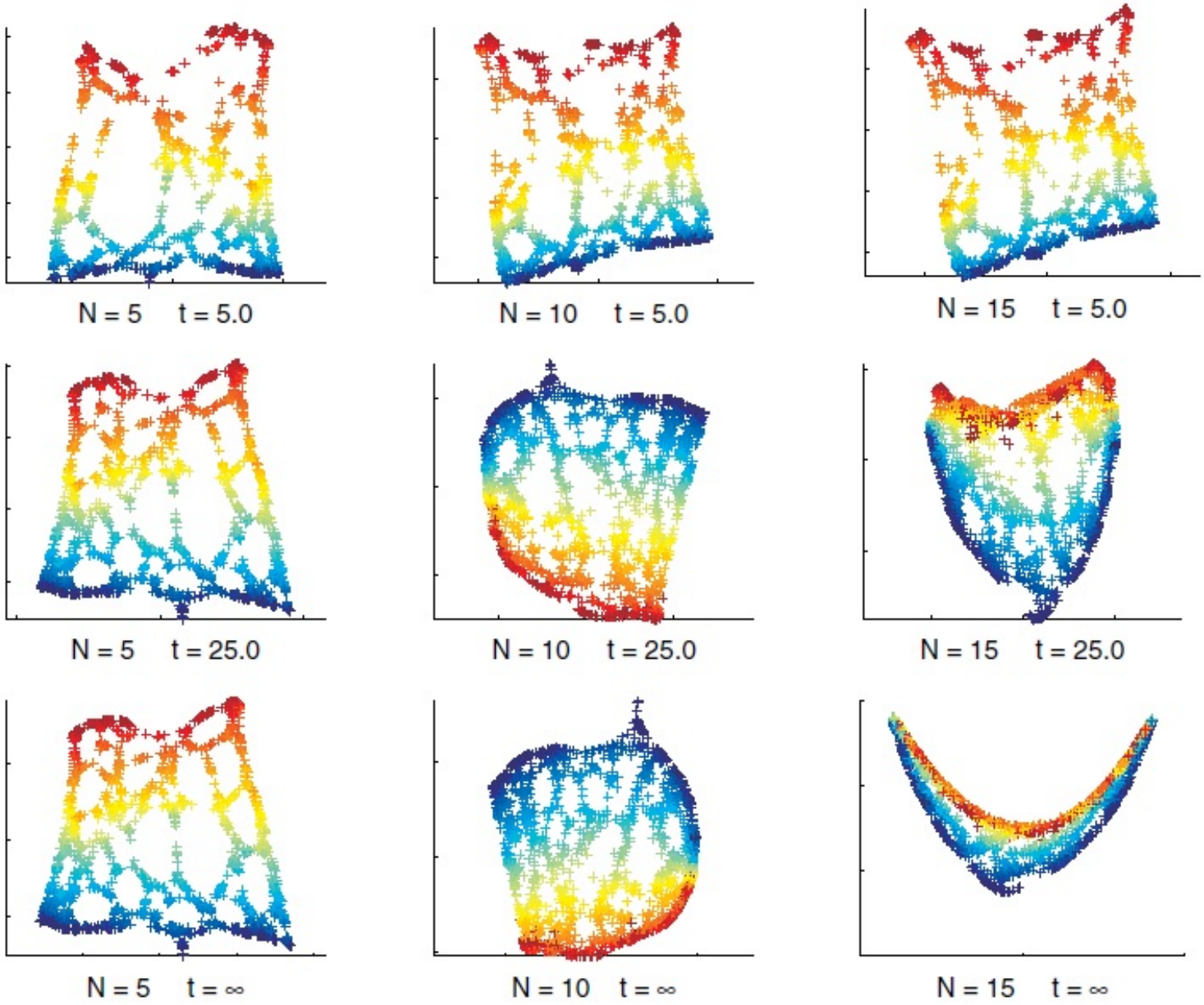

Figure: Two-dimensional representations of the swiss roll data.

2. Toy Vision Example

- 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.
- Each image can be represented as a function $f : [0,1] \times [0,1] \to \{0,1\}$ with $f(x) = 0$ if $x \in [0,1] \times [0,1]$ is white and $f(x) = 1$ if the point is black.
- 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 ($L^2([0,1] \times [0,1])$).
- Discrete version: 40×40 grid for each image, which can therefore be represented as a 1600–dimensional binary vector. Choose 1000 images (500 containing vertical bars and 500 containing horizontal bars) at random; $N = 14$ and $t = \infty$.

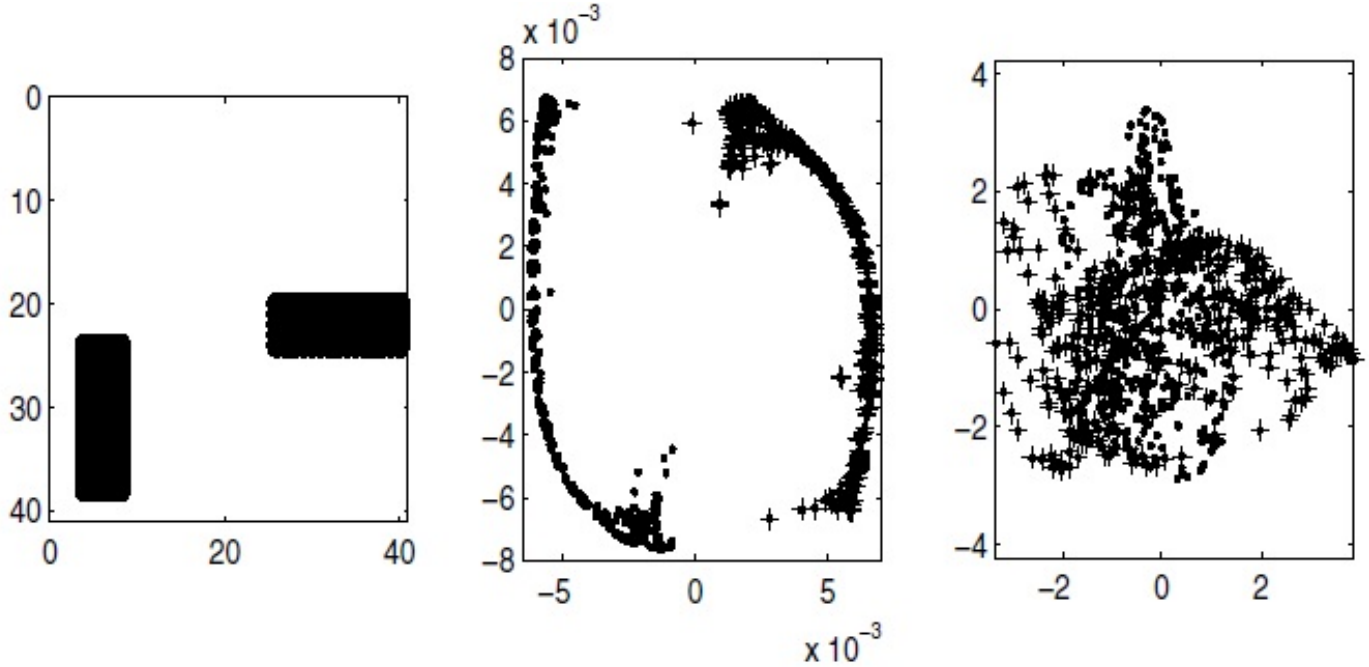

Figure: (Left) A horizontal and a vertical bar. (Middle) A two-dimensional representation of the set of all images using the Laplacian eigenmaps. (Right) The result of PCA using the first two principal directions to represent the data. Blue dots correspond to images of vertical bars, and plus signs correspond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

3. Linguistic Example

- 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); there are 300 vectors in $\mathbb{R}^{600}$.
- There is no a natural low-dimensional manifold structure on these vectors. Nevertheless, it is useful for practical applications to construct low-dimensional representations of this data.

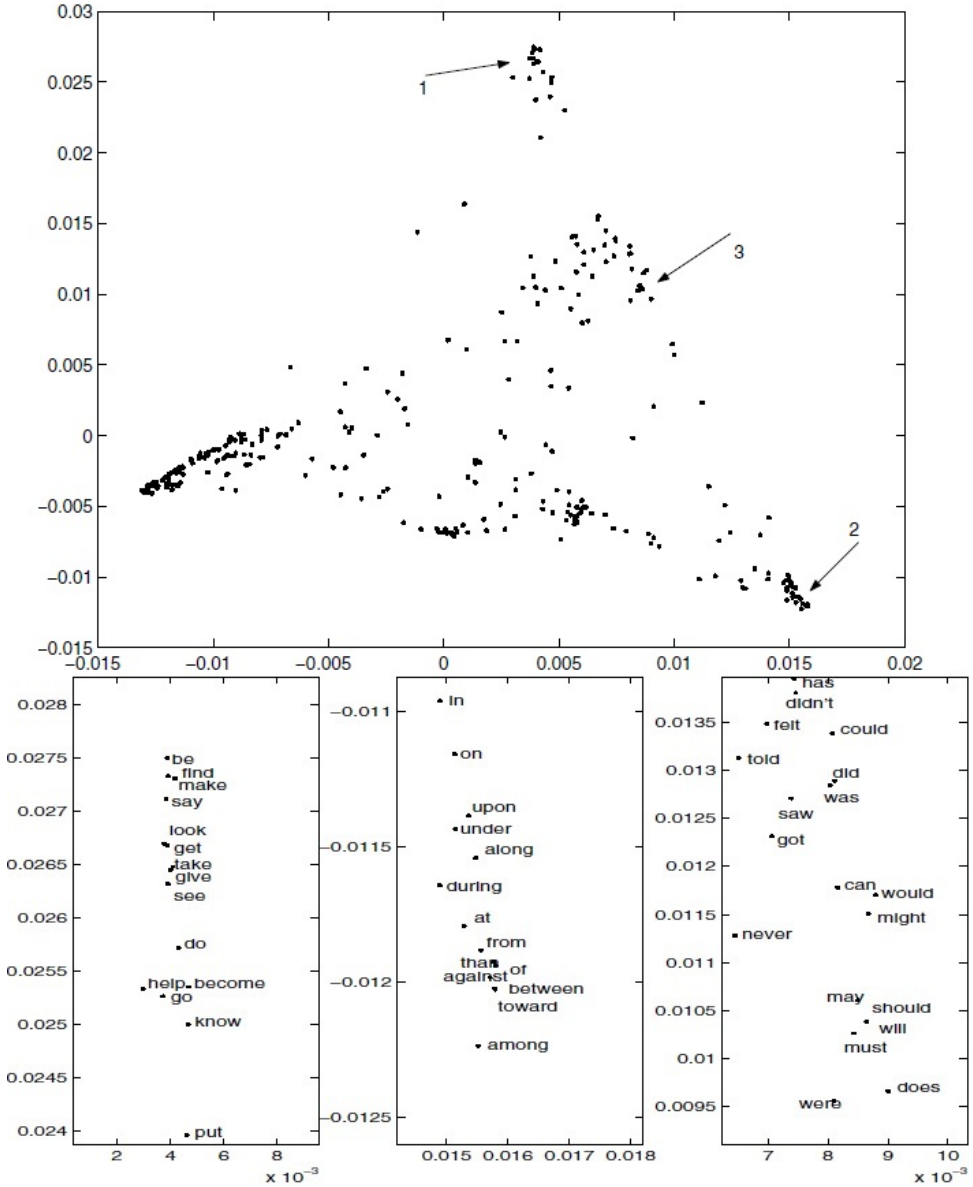

Figure: (Above) The words represented in the spectral domain. (Below) (left) 1 – infinitives of verbs, (middle) 2 – prepositions, (right) 3 – mostly modal and auxiliary verbs.

4. Speech

- 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.
- 685 speech data points were described by vectors of (logarithms of) Fourier coefficients.
- The data points corresponding to the same region have similar phonetic identity, though they arise from occurrences of the same phoneme at different points.

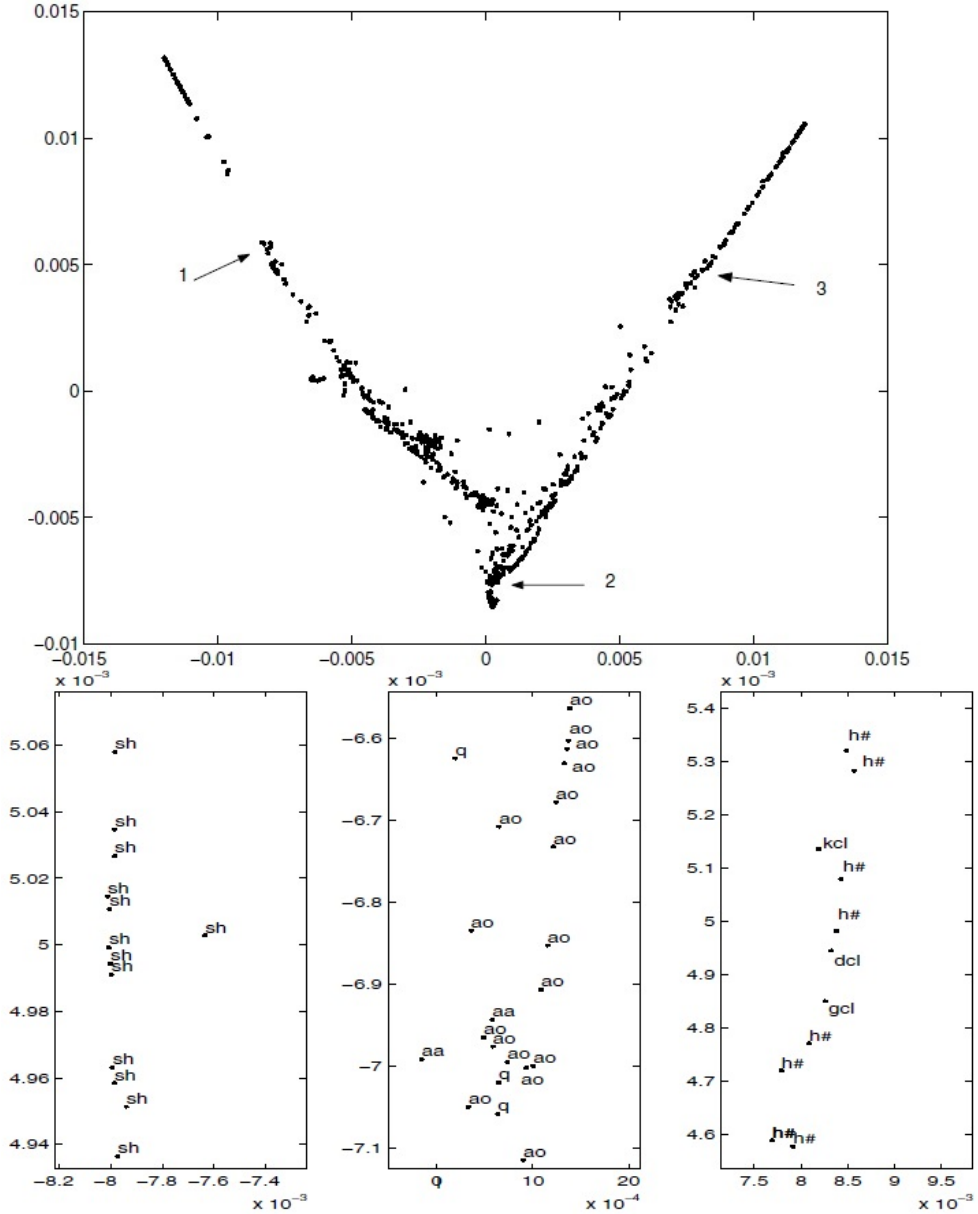

Figure: (Above) The speech data points plotted in the two-dimensional Laplacian spectral representation. (Below) A blowup of the three selected regions corresponding to the arrows.

Conclusions

- *Weak point*: the embedding is not isometric. Still, it is not clear what properties of an embedding make it desirable for pattern recognition and data representation problems.
- *Problem*: how to estimate reliably even such a simple invariant as the intrinsic dimension of the manifold.
- *Issues*: how the algorithm behaves when the manifold in question has a boundary; the effect of the choice of ε and N on the behavior of the embeddings; the convergence of the finite sample estimates of the embedding maps.
- *Main question*: while the notion of manifold structure in natural data is a very appealing one, it is not clear how often and in which particular empirical contexts the manifold properties are crucial to account for the phenomena at hand.