A TOPOLOGICAL APPROACH TO SPECTRAL CLUSTERING

ANTONIO RIESER

ABSTRACT. We propose two related unsupervised clustering algorithms which, for input, take data assumed to be sampled from a uniform distribution supported on a metric space $X$, and output a clustering of the data based on the selection of a topological model for the connected components of $X$. Both algorithms work by selecting a graph on the samples from a natural one-parameter family of graphs, using a geometric criterion in the first case and an information theoretic criterion in the second. The estimated connected components of $X$ are identified with the kernel of the associated graph Laplacian, which allows the algorithm to work without requiring the number of expected clusters or other auxiliary data as input.

1. INTRODUCTION

The analysis of complex, high-dimensional data is one of the major research challenges in contemporary computer science and statistics. In recent years, geometric and topological approaches to data analysis have been shown to yield important insights into the structure of complex data sets. The common point of departure in these methods is the assumption that data in high-dimensional spaces is often concentrated around a low-dimensional manifold or other topological space.

Geometric techniques, in particular, have proven to be particularly successful. These have largely concentrated on approximating the local geometry of the data as a step towards non-linear dimension reduction. Once an embedding of the data in a lower-dimensional space has been found, standard statistical techniques are then used to analyze the data in the lower-dimension. Methods in this class include ISOMAP [8], Locally Linear Embedding [10], Hessian Eigenmaps [6], Laplacian Eigenmaps [1], and Diffusion Maps [5]. Most of these techniques build a weighted graph to approximate the Laplace-Beltrami operator on a manifold, or else a related Markov chain on a graph, and then use the eigenvalues and eigenvectors of the resulting operator to reduce the dimension of the data, often, in practice, followed by the application of a $k$-means clustering algorithm to perform the clustering of the data [12], [11]. We encounter a more topologically-oriented approach in [3], in which persistent homology [2,7,13] is used to help a statistician identify high-density regions of a distribution function.

This research was supported in part by the grant TOPOSYS FP7-ICT-318493-STREP, Cátedras CONACYT 1076, the US Office of Naval Research Global, and the Southern Office of Aerospace Research and Development of the US Air Force Office of Scientific Research.
In this article, we give an algorithm that directly uses the topological information in the graph Laplacian to produce a clustering of the data, eliminating the need for separate dimension reduction and clustering steps. This process, furthermore, illustrates the utility of considering the data clustering problem from a topological, instead of purely analytic, perspective, i.e. to consider clustering as a problem of estimating the number of connected components of the support of an idealized underlying distribution. While the topological aspect of the clustering problem has been generally acknowledged in the topological data analysis community for some time, this is, to the best of our knowledge, the first completely data-driven clustering algorithm that explicitly exploits this point of view. Additionally, the algorithm produces both the number of clusters and the clusters themselves with no additional information required, unlike popular algorithms such as $k$-means clustering or $k$-nearest-neighbor clustering, in which the number of clusters or other additional input is required. Finally, we propose what we believe to be the first geometric and information-theoretic model selection criteria for choosing a Laplacian (or, more precisely, the associated Markov process) from a family of candidates built from the geometric properties of the sample space alone, instead of from observations of the Markov process.

2. Topology and Clustering

In all that follows, we will assume that there is an ideal unknown probability distribution $P(X)$ supported on a disconnected metric space $X = \sqcup_{i=1}^k X_i$, where $X$ is embedded as a (possibly much lower dimensional) subset of an ambient space $Y$. We will further suppose that the distribution $P(X)$ may be corrupted by noise, and that the coordinates of the sample points are given by sampling from the combined noisy distribution. We will consider a clustering to be correct if it accurately recovers the number of connected components of $X$ and it assigns to each sample point to the nearest connected component.

When $X$ has the homotopy type of a finite CW-complex, the number of connected components of $X$ is given by the dimension of the 0-th real cohomology group $H^0(X; \mathbb{R})$. Furthermore, if $X$ is a manifold, this is equal to the dimension of the kernel of the Laplace-Beltrami operator $\Delta_X$ through a small amount of Hodge theory [9]. We recall that the Laplace-Beltrami operator $\Delta : L^2(X) \to L^2(X)$ may be defined by $\Delta = -d^*d$, where $d$ is the exterior derivative, extended continuously to $L^2(X)$, and $d^*$ is its adjoint in $L^2(X)$. It follows, too, that the functions in the kernel are constant on each connected component of $X$ (Lemma 3.3.5 in [9]).

The following proposition now follows easily.

**Proposition 2.1.** Let $(X, g)$ be a smooth Riemannian manifold, and let $k$ denote the number of connected components of $M$. Let $\{\psi_i\}_{i=1}^k$ be a basis of the kernel of the Laplace-Beltrami operator $\Delta_X$ of $X$, and define the map $\Psi : X \to \mathbb{R}^k$ by

$$\Psi(x) := (\psi_1(x), \ldots, \psi_k(x)) \in \mathbb{R}^k.$$ 

Then the image of $\Psi$ consists of exactly $k$ points in $\mathbb{R}^k$, and the image of each connected component of $X$ is a single point.
Proof. First, we know from [9], Section 3.3, that each basis function \( \psi_i \) is constant on each connected component of \( X \), and it follows that each connected component is sent to a single point. It only remains to show that no two connected components are sent to the same point. Consider the matrix \( A \) defined by \( a_{ij} = \psi_i(x_j) \). Since the \( \psi_i \) are linearly independent, this is a \( k \times k \) matrix of full rank. Suppose now that there are two connected components, \( X_1 \) and \( X_2 \), whose image under \( F \) is the same point \( x \in \mathbb{R}^k \). Then two rows of \( A \) are the same, and the rank \( A < k \), a contradiction. Therefore, all of the connected components of \( X \) are sent to different points in \( \mathbb{R}^k \). \( \square \\

Remark 2.2. We note, too, that Proposition 2.1 also hold for graphs and the graph Laplacian instead of manifolds, with a nearly identical proof.

3. The Graph Laplacian and the Graph Heat Semigroup

Given a noisy sample from a disconnected metric space \( X \), motivated by the above discussion, our primary task in the clustering problem will be to compute an empirical Laplacian \( \hat{\Delta} \) and the corresponding empirical function \( \hat{\Psi} \) so that \( \hat{\Psi} \) is constant on the points sampled from a given connected component of \( X \). We begin by recalling the definition of the graph Laplacian, the associated heat semigroup, and several fundamental results.

Let \( G = (V,E,w) \) be a weighted graph, where the weight function \( w : E \to [0, \infty) \) gives the weights of every edge. We define the matrix \( L_G \) to be

\[
(L_G)_{i,j} = \begin{cases} \frac{w(x_i, x_j)}{i \neq j}, \\ -\sum_{x_j \in V} w(x_i, x_j) & i = j, \end{cases}
\]

where \( x_i, x_j \in V \). We likewise define the corresponding heat operators \( e^{-tL_G} \) by

\[
e^{-tL_G} = \sum_{k=0}^{\infty} \frac{(-tL_G)^k}{k!}\]

for \( t \in [0, \infty) \). Note that \( e^{-tL_G}e^{-sL_G} = e^{-(t+s)L_G} \), so the set of matrices \( \left\{ e^{-tL_G} \right\}_{t \in [0, \infty]} \) forms a semigroup under matrix multiplication, which we call the graph heat semigroup, or the heat semigroup of \( G \). The following result is immediate from the definitions.

Proposition 3.1. The vector \( v \in \mathbb{R}^{|V|} \) is an eigenvector of \( L_G \) with eigenvalue \( \lambda \) iff \( v \) is an eigenvector of \( e^{-tL_G} \) with eigenvalue \( e^{-t\lambda} \).

As in the manifold case, we have the following result.

Theorem 3.2 ([4], Lemma 1.7.iv, and page 3, equation 1.1). The number of connected components of \( G \) is equal to the dimension of the kernel of \( L_G \), and the vectors \( v \in \ker L_G \) are constant on each connected component of \( G \), i.e. if the vertices \( x_i, x_j \in V \) are in the same connected component of \( G \), then for the corresponding coordinates \( v \) we have \( v_i = v_j \).

Remark. Combining Proposition 3.1 with Theorem 3.2, we see that the dimension of the 1-eigenspace of \( e^{-tL_G} \) is also equal to the number of connected components of \( G \), for any \( t \in (0, \infty) \).
4. Graph Models and Model Selection for the Graph Heat Semigroup

We now describe how to combine the above ideas with several additional observations to perform unsupervised distance-based clustering on a data set \( Z \).

**Definition 4.1.** Let \( Z \) be a collection of points in a metric space \( Y \). For each \( r \geq 0 \), we define a graph \( G_r = (V_r, E_r) \) by

\[
V_r = Z, \\
E_r = \{(x_i, x_j) \in Z \times Z \mid d_Y(x_i, x_j) \leq r\}.
\]

That is, the vertices for each graph \( G_r \) are the points in the data set \( Z \), and two vertices in \( G_r \) are joined by an edge iff the distance between them is at most \( r \).

Once we have a collection of graph models \( \{G_r\}_{r \geq 0} \) for the data set \( Z \), we compute their corresponding graph Laplacians \( L_{G_r} \) (which we abbreviate to \( L_r \)), and the heat semigroups \( \{e^{-tL_r}\}_{t \in [0, \infty)} \). We would now like choose a value \( \hat{r} \) from among the \( r \geq 0 \) so that \( \{e^{-tL_{\hat{r}}}\}_{t \in [0, \infty)} \) best represents the heat semigroup \( \{e^{-t\Delta_X}\}_{t \in [0, \infty)} \), where \( X \) is the support of the distribution of the process which generated the points. To solve the clustering problem, it is sufficient to choose a \( \hat{r} \) so that the dimensions of the kernels of \( L_{\hat{r}} \) and \( \Delta_X \) are equal, i.e. so that the graph \( G_{\hat{r}} \) has the same number of connected components as \( X \), and then assign each vertex to the correct connected component. Note that, if we choose \( r \) too small, then there will be too many small connected components, but if we choose \( r \) too large, then there will be too few large ones.

We give two techniques for solving this problem, the first based on a geometric criterion, and the second based on an information-theoretic one. We compare their performance in Section 6.

4.1. Model Selection by Average Relative Neighborhood Volume.
Let \( M \) be a disconnected closed manifold (i.e. without boundary), let \( x \in M \) be a point in \( M \), and denote by \( M_x \subset M \) the connected component of \( M \) containing \( x \). Let \( |A| \) denote the volume of \( A \) for any \( A \subset M \) using the volume form on \( M \). Our first technique for choosing \( r \) is based on the observation that, on any disconnected manifold \( M \), where each connected component is of dimension \( > 0 \), for any point \( x \in M \), the function

\[
R(x) := \lim_{U \in \mathcal{N}(x)} \frac{|U|}{|M_x|} = 0
\]

where the limit is taken over the net defined by the partially ordered set \( \mathcal{N}(x) \) of neighborhoods of \( x \). That is, for any \( x \in M \), the ratio of the volume of a neighborhood of \( x \) to the volume of the connected component containing \( x \) may be made arbitrarily small. It follows that

\[
R_M := \frac{1}{|M|} \int_M R(x) \, dx = 0
\]
as well.
Now consider the family of graphs $G_r$ built on a fixed, finite sample of points $Z$ taken from a uniform distribution on a metric space $X$ as in Definition 4.1. We say that a subset $U \subset V_r = \{x\}$ is a neighborhood of a vertex $x \in Z$ iff $U$ contains $x$ and all vertices adjacent to $x$, and let the volume of a subset of $V_r$ equal its cardinality. Note that the volume of any neighborhoods $U$ of a point $x$ is necessarily bounded below by 1, and furthermore, if $M_x = \{x\}$, then $R(x) = 1$ as well. It follows that, for $r$ sufficiently small, $R(x) = 1$ for any $x$ (since every vertex is its own connected component in $G_r$), and therefore $R_{G_r} = 1$ as well. On the other extreme, if $r$ is sufficiently large, then $G_r$ will have only one component, and the smallest neighborhood of any vertex $v$ is $V$. In this case, we also have $R_{G_r} = 1$. It is not, however, difficult to find graphs $G$ with $R_G < 1$. For example, consider the circular graph $G_C = (V_C, E_C)$ with

$$V = \{0, \ldots, n-1\}$$

$$E = \{(i \mod n, (i+1) \mod n) \mid i \in V\}.$$ We therefore have that $R_{G_C} = 3/n$, and therefore $R_{G_C} < 1$ for any $n > 3$.

We define

$$\hat{r} = \arg\min R_{G_r},$$

i.e. $\hat{r}$ is the value of $r$ such that the graph $G_r$ has minimal $R_{G_r}$. In the case of a tie, we take the smallest $r$.

In this method, $G_r$ will be our choice of graph model for the data set $Z$, with Laplacian $L_\hat{r}$ and heat semigroup $\{e^{-tL_\hat{r}}\}_{t \in (0, \infty)}$.

### 4.2. Model Selection by Average Relative Entropy.

We now give an information-theoretic criterion for choosing $r$, which we motivate with the following discussion. As in the case of the Average Relative Neighborhood Volume Criterion in Section 4.1, we wish to choose a graph $G_r$ which is sufficiently locally connected to recover the connected components of $X$, but no more. Not that, if $\phi^y_0$ is a delta distribution centered at the point $x \in M$ in a manifold $M$, then the resulting steady state $\phi^y_t$ of the heat flow with initial condition $\phi^y_0$ is constant on the connected component of $M$ containing $x$, and in particular, $\phi^y_t(x) = \frac{1}{|M|}$ for all $x \in M$. Using the fact that $\phi_t$ is a probability distribution for each $t$, we also note that

$$H(\phi^y_0) - H(\phi^y_t) = \int_M \phi^y_t(x) \ln(\phi^y_t(x)) \, dx - \int_M \phi^y_0(x) \ln(\phi^y_0(x)) \, dx$$

$$= \int_M \phi^y_t(x) \ln(\phi^y_t(x)) \, dx - |M| \phi^y_0(x) \ln(\phi^y_0(x))$$

$$= \int_M \phi^y_t(x) \ln(\phi^y_t(x)) \, dx - \ln(\phi^y_t(x))$$

$$= \int_M \phi^y_t(x) \ln(\phi^y_t(x)) \, dx - \ln(\phi^y_0(x)) \int_M \phi^y_t(x) \, dx$$

$$= \int_M \phi^y_t(x) \ln(\phi^y_t(x)) \, dx - \int_M \phi^y_t(x) \ln(\phi^y_0(x)) \, dx$$

$$= d_{KL}(\phi^y_t, \phi^y_0),$$
where $H(\phi) = \int_M \phi(x) \ln(\phi(x)) \, dx$ is the entropy of the distribution $\phi$, $d_{KL}$ is the Kullback-Leibler divergence, or relative entropy, between $\phi_1$ and $\phi_2$.

For a distribution $\phi_1$ which is concentrated around a point $y$ on a large connected component $M_y$, $d_{KL}(\phi_t, \phi_1^y)$ will be large, and furthermore, if this is true for any $y \in M$, it will also true for the average

$$H_{M,t} := \frac{1}{|M|} \int_M d_{KL}(\dot{\phi}_{t,1}^y, \phi_1^y) \, dy.$$

Now, let $\{G_r\}_{r \geq 0}$ be the family of graphs from Definition 4.1, $\{L_r\}_{r \geq 0}$ be, and we consider the collection of graph heat semigroups $\{e^{-tL_r}\}_{t \in [0, \infty), r \geq 0}$, where the vertices of the $G_r$ are sampled from a uniform distribution on a manifold $X$ in $Y$ with noise. The empirical initial distributions $\hat{\phi}_{r,0}^i$ centered at a point $i \in V$, corresponding to the delta distribution in the manifold case, are given by the standard basis vectors $e_i$, $i \in \{1, \ldots, n\}$, where $n = |V_r|$, the number of vertices in the graph. The solution $\hat{\phi}_{r,t}^i$ at time $t$ of the empirical heat flow with initial condition $\hat{\phi}_{r,0}^i = e_i$ is therefore given by the $i$-th column of $e^{-tL_r}$. The empirical steady state $\hat{\phi}_{r,\ast}^i$, naturally, is the $i$-th column of $\lim_{t \to \infty} e^{-tL_r}$.

Define

$$H_{r,t} := \frac{1}{n} \sum_{i=1}^n d_{KL}(\hat{\phi}_{r,t}^i, \hat{\phi}_{r,\ast}^i).$$

Motivated by the discussion in the paragraphs above, we will choose our preferred scale $\hat{r}$ to be the one which maximizes $H_{r,1}$. This is, we choose $\hat{r}$ to be the value of $r$ at which the empirical average relative entropy at time $t = 1$ takes its maximum, i.e.

$$\hat{r} := \arg\max_{\hat{r}} H_{r,1}. \quad \text{(4.2)}$$

When there are many small clusters $\bar{H}_{r,1}$ will be small, since the average size of the support of $\hat{\phi}_{r,1}^i$ is a large portion of each connected component. This will also be the case when $r$ is large. We therefore see that a kind of "bias-variance" tradeoff is built into the geometry of local neighborhoods vs. connected components, and that this is what powers both methods.

We conclude this section with the remark that, from the properties of the heat equation on $\mathbb{R}^n$, the fundamental solutions have maximal entropy among all distributions which satisfy certain mean and variance constraints. One might expect, based one this, that we should seek to minimize $H(\phi_{r,\ast}) - H(\phi_{r,t})$, but, in fact, this is the opposite of the effective approach, the relevant constraints on the family $\{\phi_{r,t}\}_{r \geq 0}$ being different.

5. Cluster Identification

The results in Section 4 tell us that the points in each connected component of a graph $G$ should be sent to exactly the same point in $\mathbb{R}^k$, where $k$ is the number of connected components of $G$, by the map $\Psi$. In practice, however, there are sometimes small amounts of numerical error in the algorithms for computing eigenvalues and eigenvectors, and this must be accounted for when constructing the final clustering. We do this with a modified version
of Gaussian elimination on the matrix formed by the eigenvectors, which we now describe.

First, note that the \( j \)-th entry in the eigenvector \( f_i \) is the value of the eigenfunction \( f_i \) evaluated on the point \( z_j \). Let \( \Psi \) be the matrix defined by

\[
(\Psi)_{i,j} = (\psi_i)_j = \psi_i(z_j),
\]

We give a modified Gaussian elimination algorithm in Algorithm 1. For what follows, let \( n \) denote the number of points in our sample, and let \( k \) the number of connected components of the graph \( \hat{G}_r \).

**Algorithm 1** Modified Gaussian elimination on \( \Psi 

1: for \( i = 1 \) to \( k \) do 
2: Reorder columns \( i \) through \( n \) of \( \Psi \) so that \( |(\Psi(i,i))| \) is the maximum of \( |(\Psi(i,j))| \) in row \( i \).
3: Divide row \( i \) by \( (\Psi(i,i)) \).
4: Using elementary row operations, make \( (\Psi(k,i)) = 0 \) for \( k \neq i \).
5: end for 
6: Redefine \( \psi_i := \Psi_{i,*} \), and (abusing notation) using the new \( \psi_i \), redefine the map \( \Psi(z_m) := (\psi_1(z_m), \ldots, \psi_k(z_m)) \).

Note that the algorithm, if there was no estimation error, would send each point in the sample to one of the vectors \( e_i \) in the standard basis of \( \mathbb{R}^k \). Now, however, even given some numerical error, we are able to cluster the sample points according to how close \( \Psi(x) \) are to each of the vectors \( e_i \).

6. Algorithm and Experiments

We now give the complete algorithm and the results of some numerical experiments. We denote the set of points by \( \mathcal{Z} \).

**Algorithm 2** Clustering algorithm

1: For each \( r < \text{Diam}(Z) \), compute \( G_r, L_r, e^{-tL_r} \) and estimate \( \lim_{t \to \infty} e^{-tL_r} \) by \( e^{-t^*L_r} \) for some \( t^* \) large (we use \( t^* = 1000 \)).
2: Using one of the methods in Sections 4.1 or 4.2, compute \( \hat{r} \).
3: Compute the kernel of \( \hat{L}_r, \psi_i, i \in 1 \ldots k \).
4: Using Algorithm 1, create the map \( \Psi : z_m \mapsto (\psi_1(z_m), \ldots, \psi_k(z_m)) \in \mathbb{R}^k \).
5: Compute the distances \( d_i(z_m) = \|\Psi(z_m) - e_i\| \) for each point \( z_m \) in the sample.
6: Assign the vertex \( m \) to the \( i \)-th cluster if \( d_i(z_m) < d_j(z_m) \) for all \( j \neq i \).

The following figures summarize the output of this algorithm on a data set of 500 points sampled with a small amount of Gaussian noise from three interlinked circles embedded in \( \mathbb{R}^3 \). The horizontal circle has radius 1 and center \((0,0,0)\), and the other two have radii 0.5 and 0.4 and centers \((0,-1,0)\) and \((0,1,0)\), respectively. We ran several trials with different levels of Gaussian noise, with standard deviations 0.02 (low noise), 0.03 (medium noise), and 0.045 (high noise). The value of \( \hat{r} \) given by both methods was identical.
for the low noise experiment, and the algorithm correctly assigned all of the points to their respective circles. In the medium noise experiment, the Average Relative Entropy Method continues to classify the circles correctly, but the Average Neighborhood Volume Method simply puts all of the circles into a single cluster. For the high noise experiment, we see that the Average Relative Entropy Method also starts to break down, but nonetheless identifies subclusters of the circles, from which a user would be able to reconstruct the original circles. In contrast, the Average Local Volume Ratio groups all of the points into a single cluster. We therefore see that, while both methods work in ideal conditions, the Average Relative Entropy method is far more robust to noise. Interestingly, too, both the relative entropy curves and the local volume ratio curves exhibit local maxima and minima, respectively, where the smaller circles join the same group as the central circle, behavior reminiscent of the "barcodes" in topological data analysis. We include the figures below.

6.1. **Low Noise Experiment** ($\sigma = 0.02$). In this experiment, both methods successfully recovered the circles. Figure 6.1 illustrates the data set, Figure 6.2 shows the Average Relative Entropy at each scale, Figure 6.3 gives the plot of the Average Local Volume Ratio at each scale, Figure 6.4 shows the image of the map $\Psi$, and, finally, Figure 6.5 shows the classification of the points. Note that in Figures 6.2 and 6.3 the joining of the two smaller circles to the group of the large circle is indicated by the second and third local maxima and minima, respectively.

**Figure 6.1.** 500 points sampled from three circles with Gaussian noise ($\sigma = 0.02$)
Figure 6.2. Average Relative Entropy vs. Scale, Low Noise Experiment. Note that local maxima appear where the smaller circles join to a larger cluster.
Figure 6.3. Average Local Volume Ratio vs. Scale. Note that local minima appear where the smaller circles join with the larger circle.
Figure 6.4. The image of $\Psi$, centered around the points $(1,0,0),(0,1,0)$, and $(0,0,1)$
6.2. Medium Noise Experiment ($\sigma = 0.03$). In this experiment, the Average Relative Entropy Method successfully recovered the circles, and the Average Local Volume Ratio Method returned a single cluster of all three circles. Figure 6.6 illustrates the data set, Figure 6.7 shows the Average Relative Entropy at each scale, Figure 6.8 gives the plot of the Average Local Volume Ratio, and Figure 6.9 shows the classification of the points. The image of $\Psi$ is roughly identical to the previous experiment, so we do not repeat the plot here. Note that, in contrast to the previous experiment, the global minimum of the Average Local Volume Ratio occurs at the third local maximum.
Figure 6.6. 500 points sampled from three circles with Gaussian noise ($\sigma = 0.03$)
Figure 6.7. Average Relative Entropy vs. Scale, Medium Noise Experiment. As before, the local maxima appear where the smaller circles join to form a larger cluster.
Figure 6.8. Average Local Volume Ratio vs. Scale. Note that local minima appear where the smaller circles join with the larger circle, and the global minimum is the third local minimum.
6.3. **High Noise Experiment** ($\sigma = 0.045$). In this experiment, we see the manner in which the Average Relative Entropy Method begins to break down. While it successfully identified the two smaller circles, the large, central circle is split into two clusters. While not ideal, the circles could nonetheless be reconstructed from this clustering. As in the medium noise experiment, the Average Local Volume Ratio Method returned a single cluster of all three circles. Figure 6.10 illustrates the data set, Figure 6.11 shows the Average Relative Entropy at each scale, Figure 6.12 gives the plot of the Average Local Volume Ratio, and Figure 6.13 shows the classification of the points by color. The image of $\Psi$ is roughly identical to the previous experiment, so we do not repeat the plot. As in the medium noise experiment, the global minimum of the Average Local Volume Ratio is found at the third local minimum, i.e. after all of the circles have been joined to the same cluster.
Figure 6.10. 500 points sampled from three circles with Gaussian noise ($\sigma = 0.045$)
Figure 6.11. Average Relative Entropy vs. Scale, Medium Noise Experiment. As before, the local maxima appear where the smaller circles join to form a larger cluster.
Figure 6.12. Average Local Volume Ratio vs. Scale. As before, local minima appear where the smaller circles join with the larger circle, and as in the previous experiment, the global minimum is the third local minimum.
A TOPOLOGICAL APPROACH TO SPECTRAL CLUSTERING

Classification of Sample Points via Relative Entropy, Noise SD = 0.045

Figure 6.13. The classification of the points using the Average Relative Entropy method, illustrated by color. The Average Local Volume Method assigned all of the points to a single class in this experiment.

7. Discussion and Future Work

We have presented two novel data clustering algorithms for data sampled from a uniform distribution on a disconnected metric space $X$, possibly corrupted by Gaussian noise. Both algorithms work by identifying a scale $\hat{r}$ with which to build a graph on the points, after which they identify the connected components of the graph using the associated graph Laplacian. Unlike other commonly used clustering algorithms, this technique is completely data driven, and does not require any additional parameters. In particular, the algorithms output the number of clusters as well as the clustering, unlike the popular $k$-means algorithm, which requires the number of clusters as input. Of the two algorithms presented, the Average Relative Entropy Method outperformed the Average Local Volume Ratio Method in terms of robustness to noise.

We remark, however, that the success of these particular algorithms, is highly dependent on the assumption of uniformity of the underlying, non-noisy distribution. This, unfortunately, prevents the current form of these algorithms from producing a correct clustering when applied to many datasets, and the adaptation of these techniques to non-uniform distributions is the subject of ongoing research. We also note that the Average Relative Entropy Method may be seen as a variation of Diffusion Maps [5] and Laplacian Eigenmaps [1], where the choice of free parameter is done in an automatic,
data-driven fashion, which allows for clustering to be achieved directly from the eigenvectors of the operators, instead of after a dimension-reduction step. While our method depends on the kernel function having compact support, which is not the case in either Diffusion Maps or Laplacian Eigenmaps, and in this paper we have concentrated on the clustering problem, it would be interesting to extend the applicability of these model selection methods to dimension-reduction problems as well.

Acknowledgements. We would like to thank Robert Adler, Jacob Abernathy, and Bertrand Michel for useful discussions and comments on an earlier, preliminary version of this work. We would additionally like to thank the organizers of the CIMAT Differential Geometry Seminar, the organizers of the AUS-ICMS Meeting, the XXIst Oporto Meeting on Topology, Geometry, and Physics, and the Toposys Network for the opportunity to present our preliminary results in their conferences and seminars.

References

[1] Mikhail Belkin and Partha Niyogi, Laplacian Eigenmaps for Dimensionality Reduction and Data Representation, Neural Comput. 15 (June 2003), no. 6, 1373–1396.
[2] Gunnar Carlsson, Topology and data, Bull. Amer. Math. Soc. (N.S.) 46 (2009), no. 2, 255–308. MR2476414
[3] F. Chazal, L. J. Guibas, S. Y. Oudot, and P Skraba, Persistence-based clustering in riemannian manifolds, Journal of the ACM 60 (November 2013), no. 6.
[4] Fan R. K. Chung, Spectral graph theory, CBMS Regional Conference Series in Mathematics, vol. 92, Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1997. MR1421568
[5] Ronald R. Coifman and Stéphane Lafon, Diffusion maps, Appl. Comput. Harmon. Anal. 21 (2006), no. 1, 5–30. MR2238665 (2008a:60210)
[6] David L. Donoho and Carrie Grimes, Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data, Proceedings of the National Academy of Sciences 100 (2003), no. 10, 5591–5596, available at http://www.pnas.org/content/100/10/5591.full.pdf+html
[7] Robert Ghrist, Barcodes: the persistent topology of data, Bulletin of the American Mathematical Society 45 (2008), no. 1, 61–75.
[8] Josh Langford Joshua B. Tenenbaum Vin de Silva, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000), 2319.
[9] Jürgen Jost, Riemannian geometry and geometric analysis, Sixth, Universitext, Springer, Heidelberg, 2011. MR2829653
[10] Lawrence K. Saul Sam T. Roweis, Nonlinear dimensionality reduction by local linear embedding, Science 290 (2000), 2323.
[11] Ruben Sanchez-Garcia, Max Fennelly, Sean Norris, Nick Wright, Graham Niblo, Jacek Brodzki, and Janusz Bialek, Hierarchical spectral clustering of power grids, IEEE Transactions on Power Systems (March 2014), 1–9.
[12] Wen-Jun Shen, Hau-San Wong, Quan-Wu Xiao, Xin Guo, and Stephen Smale, Introduction to the peptide binding problem of computational immunology: New results, Foundations of Computational Mathematics 14 (2014), no. 5, 951–984 (English).
[13] Afra Zomorodian and Gunnar Carlsson, Computing persistent homology, Discrete & Computational Geometry 33 (2005), no. 2, 249–274.

CONACYT-CIMAT, CENTRO DE INVESTIGACIÓN EN MATEMÁTICAS, CALLE JALISCO S/N, COLONIA VALENCIANA, C.P 36023 GUANAJUATO, GUANAJUATO, MÉXICO
E-mail address: antonio.rieser@cimat.mx