Topological Classification of Data Sets without an Explicit Metric

Tim Harrington, Andrew Tausz and Guillaume Troianowski

December 10, 2008

A contemporary problem in data analysis is understanding the nature of high dimensional data sets. Given a set of points in a high-dimensional space, a key question is to determine its topological characteristics. One potential route toward topological classification would be through the standard machine learning methods. However, this approach is too inefficient due to the tremendous variability in the possible examples. Two surfaces that are topologically equivalent may be completely different as subsets of Euclidean space. In other words, the set of point clouds that are topologically equivalent (i.e. homeomorphic) to a given surface is so vast that it would be computationally infeasible to approach the problem in this way. On the other hand, the theory of persistent homology developed by Gunnar Carlsson and his collaborators presents a tractable method for topological classification of point clouds. More information on the subject can be found in [1], and [2]. There is one catch, however, in that the point cloud must have a well-defined metric. In this paper we propose a variation of this method to allow for topological classification of point clouds in spaces where there is only a probabilistic notion of a metric.

1 Background

Recently, data set analysis has become more and more important. To deal with high-dimensional datasets, Carlsson et al. developed a method in computational algebraic topology. Referred to as the theory of persistent homology, this method is used to detect topological features present in point clouds. The idea is to generate something called a simplicial complex, which is an object constructed by gluing together points, line segments, triangles, and higher dimensional analogues of triangles (called n-simplices). From there the simplicial complex is used to compute a sequence of Betti numbers which characterize the estimated topological features present in the data set. Examples of some common shapes with their Betti sequences can be seen in Figure 1.

In persistent homology theory the Betti sequence associated with a complex is also referred to as a *barcode*. Informally, a barcode constitutes a topological signature of the underlying simplicial complex. However, complexes are constructed using an arbitrary notion of proximity given by a parameter ϵ. That is, depending on ϵ the generated complex can exhibit varying topological features. The method constructed by Carlsson et al. handles this problem by considering a sequence of complexes generated by varying ϵ. For each complex an accompanying barcode is generated and only the topological features that persist throughout the sequence are considered representative of the underlying point cloud. For a rigorous definition and theoretical justification of this method see [1, 2]. For more information about the underlying theory see [3].

The tools of persistent homology can therefore generate a reliable topological barcode of the point cloud. However, the theory relies on a clear metric, or notion of distance, between points to construct simplicial complexes. In many data sets distance must be inferred probabilistically. For example, let (X, d) be a metric space consisting of a set of points X and a distance function d. For the purposes of this paper $X = \mathbb{R}^3$ and $d(x, y)$ is Euclidean distance. Let $S = x_1, x_2, x_3, ..., x_n$ be a collection of points drawn from X. For each pair of points x_i, x_j define $d_{ij} = d(x_i, x_j)$. Now define $u(x_i, x_j) = z_{ij}$, where z_{ij} is a random variable drawn independently from a Poisson distribution with mean $(1/\lambda)/d_{ij}$.

Here is the main idea of the project. Suppose that the distances d_{ij} between points in X cannot be obtained, but that the random distances z_{ij} can. This paper presents a way to construct a metric on the point cloud data in such a way that the topological structure of S in (X, d) is preserved.

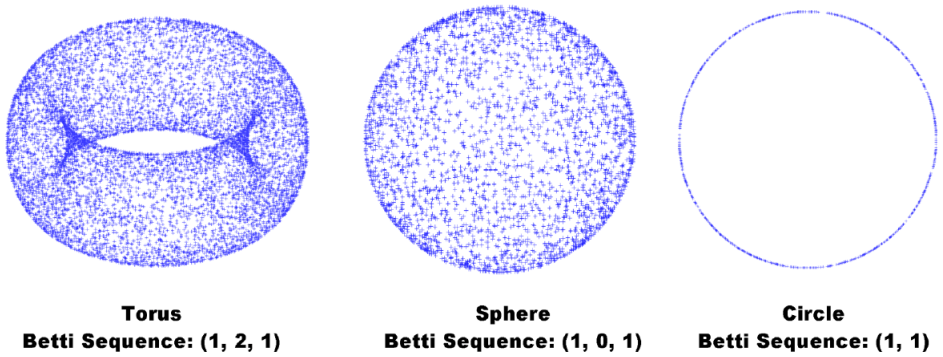

Figure 1: Betti Sequences of Various Shapes

2 Application Area

Using these ideas we wish to create test data sets with only a probabilistically notion of a metric. Informally, suppose we have a set of people distributed on the surface of a sphere, and these people make telephone calls to their friends. The idea is that they call geographically close friends with a much higher probability than they call friends that are far away. We can create a weighted graph from this data by using the number of calls as a weight. In this space we do not have a metric since the only data we are given is the frequency of calls between different people. To deal with this problem we can embed the data into a higher dimensional topological vector space to help us infer more about the data set.

For the actual simulation we distribute points uniformly on a surface (e.g. a sphere) and create a matrix of pairwise distances between nodes. Then we compute the probabilistic similarity between points as a Poisson random variable with a parameter proportional to the distance between points. The proportion parameter is lambda and it controls the variability of call frequency. Probabilistic distances between nodes are then generated as the reciprocal of call frequency. Next we compute kernel distances by mapping the nodes into a higher dimensional space and computing L^2 norms. At the last step in this process we use the kernel distances to compute the topological characteristics of the data set using PLEX, which is a software library developed by professor Carlsson's group for calculating the persistent homology of data sets in Euclidean space.

Let the set of callers be $C = \{c_1, ..., c_m\}$, where m is the number of callers. In our simulation, we generate m points on a surface. From these points we generate a distance matrix D, where D_{ij} is the distance from point i to the point j. From this physical distance information we then generate a probabilistic similarity matrix. To do this we create a new matrix W such that the i, j-th element is determined by a Poisson random variable with parameter (mean) given by $(1/\lambda)(1/D_{ij})$. For W to be a consistent measure of similarity, we symmetrize the matrix by setting $W \leftarrow 1/2(W + W^T)$. Intuitively, the elements of W represent the number of telephone calls between two points. If the points are far apart the number of phone calls will be low, whereas if they are close the points are more likely to call each other. Note that after this point we completely discard the original physical distances. Our aim is to recover the topology of the original space from the probabilistic call information, without any knowledge of the actual locations of the points.

From the call frequency matrix W we can generate a probabilistic distance matrix P by setting $P_{ij} = 1/W_{ij}$. Although the original points were elements of $\mathbb{R}^n$ for some n, we can no longer use this information. Instead we can consider these to be elements of the high-dimensional space $\mathbb{R}^m$, where m is the number of points on the surface. This can be accomplished via the mapping $\varphi : c_j \mapsto P_j$, where $P = [P_1, ..., P_m]$ (i.e. P_j are the columns of P). In other words, for each point our representation of it in $\mathbb{R}^m$ is given by its corresponding column in the probabilistic distance matrix P. This suggests that we can use the standard metric derived from the 2-norm on $\mathbb{R}^m$ as a distance measure for points on the original unknown manifold. Thus we can use $K(c_i, c_j) = \|\varphi(c_i) - \varphi(c_j)\|_{2,\mathbb{R}^m}$. This mapping into a high dimensional vector space is very similar to the idea of using a kernel. However, instead of using the kernel as a measure of similarity, we are using it as a measure of distance between two points.

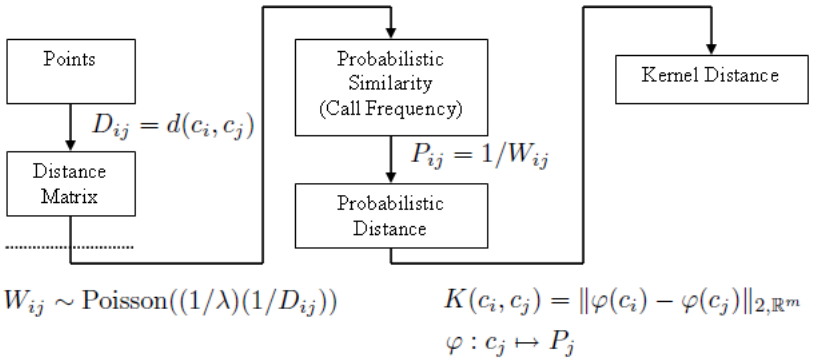

Figure 2: Flow Chart of Calculations

Once we have embedded our points into this high-dimensional vector space, we can compute the distances between each point using our selected metric. Given this pairwise distance information we can construct a filtered sequence of simplicial complexes, each parameterized by $\{\epsilon_n\}$. $\{\epsilon_n\}$ is a monotone sequence that discretizes the interval we wish to compute simplicial complexes over. In practice, these will actually be approximated complexes in order to improve the efficiency of the calculations. Using these complexes it is then possible to calculate the persistent homology to characterize the topological properties of the set C. The expected results of our method are the correct Betti sequences for each tested surface. In our example with the telephone calls on the surface of a sphere, we expect to obtain the same barcode as for a sphere: the sequence of Betti numbers (1, 0, 1, 0, 0 ...). A flow chart of this process can be seen in Figure 2.

3 Results

We have completed some simulations and tests of the ideas presented in the previous paragraphs. Our first approach was to simulate random points on a sphere and then create a random adjacency matrix using these points. This produced a graph with random weights given by a Poisson distribution. We represented the weighted graph as an adjacency matrix, then mapped it to a distance matrix with elements $K(c_i, c_j) = \|\varphi(c_i) - \varphi(c_j)\|_{2,\mathbb{R}^m}$.

Several tests were completed which verified whether it was possible to recover the correct Betti sequence of the surface given only the probabilistic information. Table 1 shows the success rates of calculating the Betti sequences at three points along the execution of our algorithm. The sequences were computed in PLEX using the actual physical distances, the probabilistic distances and the kernel distances. Small values of λ amplify the number of calls made. From the table it is evident that the method works better for lower values of lambda and higher values of the number of points

Figure 3 shows histograms and scatter plots for the call frequencies, probabilistic distances and kernel distances in the case of the 2-sphere. The histogram in the top left corner shows that most of the calls are to places that are close. The scatter plot in the lower left demonstrates the inverse relationship between the actual physical distance and the call frequency. Finally, the lower scatter plots in the center and the right show that the probabilistic and kernel distances are good approximations of the actual distance.

4 Discussion

The results in Figure 1 show the accuracy of our method compared to the correct topological characterization of the underlying distance data generated from the circle, the flat sphere and the 2-sphere. The flat sphere is a space that is topologically equivalent to the conventional 2-sphere in $\mathbb{R}^3$. It is a square with all of the edges collapsed to a single point. Thus two points that are close to the edges are actually close together,

Surface	Number of points	λ	Success rate of PLEX on actual physical distances	Success rate of PLEX on probabilistic distances	Success rate of PLEX on kernel distances
Sphere	300	0.1	1	0.2	0
	300	0.01	0.8	0.8	0.7
	600	0.1	1	0.6	0
	600	0.01	1	1	1
Flat Sphere	300	0.1	0.7	0.4	0.2
	300	0.01	0.9	0.7	0.7
	600	0.1	0.9	0.2	0.1
	600	0.01	0.6	0.4	0.2
Circle	300	0.1	1	0.9	0.2
	300	0.01	1	1	1
	600	0.1	1	1	0
	600	0.01	1	1	1

Table 1: Success rates for classifying the Betti sequences based on the actual, probabilistic and kernel distances

even though the interior distance between them may be large. Since our goal is to recover the topology of the data set, the rate of success on the original physical data serves as a reference.

The results show that for values of λ larger than 0.1 the method breaks down. What is interesting is that this limitation does not seem to change when the number of points double. For values of lambda an order of magnitude small the method works with a much higher degree of confidence, recovering the correct Betti sequences for the circle, the 2-sphere and the flat sphere.

We have identified several potential avenues for future work:

- The size of the distance matrices grow proportionally to the square of the number of points, which prevented us from studying the impact of the number of points on the threshold value of λ. This suggests the need to develop more efficient algorithms in order to reduce the spatial complexity and consider much larger data sets.
- Our method was tested with only a small number of surfaces. In order to test the robustness of our method other surfaces could be considered as future candidates.
- In all of our examples the probability distribution we used was a Poisson distribution. Perhaps this method could be extended to utilize other distributions.

5 Acknowledgments

The authors would like to thank Gunnar Carlsson for his guidance throughout this project, Mikael Vejdemo Johansson for assistance with PLEX and Andrew Ng for teaching CS229: Machine Learning.

References

[1] Gunnar Carlsson. Topology and data, preprint, August 2008.

[2] Robert Ghrist. Barcodes: The persistent topology of data, Bull. Amer. Math. Soc. 45 (2008), 61-75.

[3] Allen Hatcher. *Algebraic Topology.* Cambridge University Press (2002).

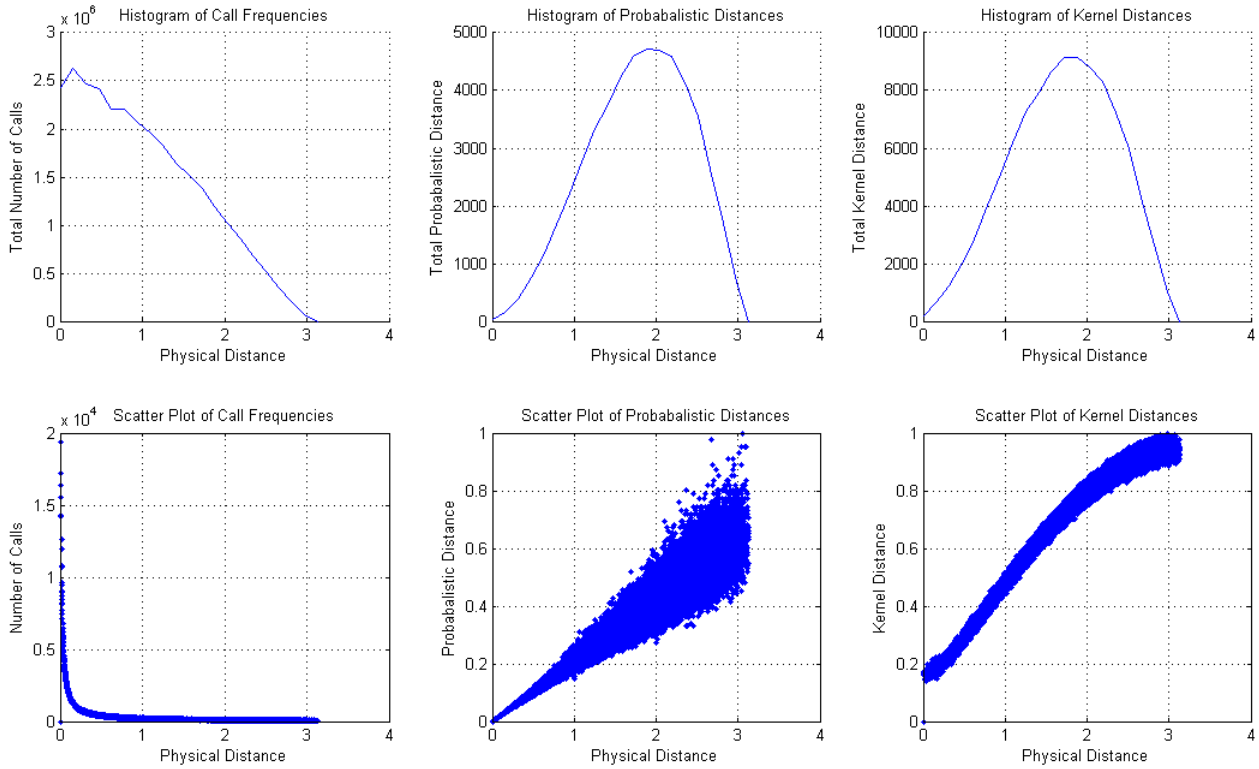

Figure 3: Histograms and Scatter Plots of Physical, Probabilistic and Kernel Distances

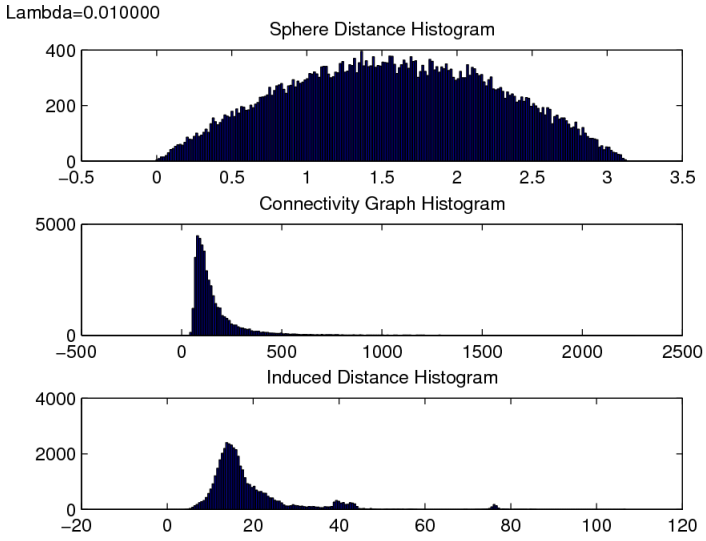

Figure 4: Histograms of Connectivity and Induced Kernel Distances