UNDERSTANDING PERSISTENT HOMOLOGY AND PLEX USING A NETWORKING DATASET

Laura Balzano[*]

University of Wisconsin, Madison
sunbeam@ece.wisc.edu

Jordan S. Ellenberg[†]

University of Wisconsin, Madison
ellenber@math.wisc.edu

ABSTRACT

We report on a pilot study of the topological structure of data sets using Plex, a software package for computational homology [1]. Plex assigns to a set of points in $\mathbb{R}^n$ a "barcode," which is intended to reveal topological structure in data. In particular, if the data are drawn from a low-dimensional manifold M, the barcode is meant to capture the homology of M. We compared barcodes coming from three sources: byte-count data from the UW-Madison core computer network, Gaussian noise centered at a point, and Gaussian noise convolved with a circle. One of our goals is to understand how "persistent" the result of Plex needs to be in order to distinguish a data set with topological structure from one consisting solely of noise.

Index Terms— Persistent homology, dimension reduction, manifold learning

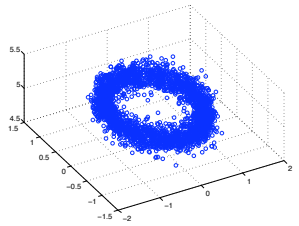
1. INTRODUCTION

Data in high dimensions are essentially impossible to visualize directly, so with the explosion of high-dimensional data collection, many tools for dimensionality reduction and data visualization are being studied. A great number of these tools try in ways to find a mapping of the high-dimensional data to lower dimensions, and they try to make that mapping as true to the structure of the data as possible.

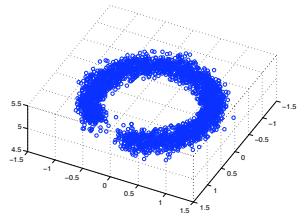
All such methods assume an inherent low-dimensional structure to the data. This is convenient in practice; it helps us visualize relationships in the data and it helps us find parsimonious explanations of a complex system's behavior. But this parsimony may oftentimes be fundamental as well: even very complex systems usually have a small number of factors that are influencing behavior, or a small number of goals or outcomes, etc. By making this assumption, we are saying that nature is simple in some sense; and if it were not we would have little hope of understanding it.

[*]We thank Gunnar Carlsson, Matthew Kahle, and Robert Nowak for useful discussions about the material in this report. The first author's research was supported by the AFOSR grant FA9550-09-1-0140.

[†]The second author's research is partially supported by NSF-CAREER Grant DMS-0448750 and a Romnes Faculty Fellowship.



(a) A noisy circle; a 1-d manifold embedded in 3 dimensions. The Betti numbers of this circle are $\beta_0 = 1$, $\beta_1 = 1$; that is it has one connected component and one hole.



(b) A noisy circle whose ends do not connect. It's topology has $\beta_0 = 1$, but no larger Betti numbers.

We are led to the problem of *dimension reduction*; given a set of data points in a high-dimensional space $\mathbb{R}^N$, find some "natural" smaller manifold M which is close to most of the data points. Traditional methods try to find such an M which is a linear subspace of $\mathbb{R}^N$. More recently, attention has been focused on situations in which M is allowed to be nonlinear. Following the lead of Carlsson et al [3] one might ask about the *topology* of M. Speaking loosely, the topological features of a manifold are those which are preserved by deformations without tearing; so a line and a parabola are topologically identical, while a line and a circle are not.

The method of *persistent homology* described in [3] is meant to reveal topological information about the hidden manifold M. For example, imagine that the data arise from a circle as in Figure 1(a). If we could identify this circle, we could parameterize the space in which the data exist down to a manageable size, in this case one dimension. Identifying this circle from a finite collection of noisy data points is called manifold learning. The difficulty, of course, is that a discrete set of points does not literally have any interesting topology. The customary means of overcoming this problem is to replace each point with a small ball of radius ϵ, which has the effect of "smoothing" between the data points. The result depends strongly on the choice of the parameter ϵ. For instance, imagine that instead of the circle you have a manifold that is nearly a circle but has a break, as in Figure 1(b). If we choose ϵ large enough to bridge the gap in Figure 1(b),

we will have a connected circle as our manifold; if we choose ϵ smaller, we will see simply a curved line.

The notion persistent homology [3] neatly avoids the necessity of fine-tuning ϵ by keeping track of what happens for *all* values of ϵ. Topological features that "persist" for a wide range of ϵ are then deemed to be actual features of the data set. The topological features in question are "homology groups," which to a very first approximation can be described as a measure of the "number of holes" in the manifold. Technical details can be found in [3] and Plex documentation [1], which can be used to identify persistent features. For example, the homology group H_0 refers to the number of connected components in the dataset; the homology group H_1 refers to the 2-dimensional holes (circular discs) in the manifold, H_2 to the 3-dimensional holes (interiors of spheres), and so on. To give a fuller exposition of homology groups of manifolds is beyond out scope; for the remainder, it suffices to know that

- h_0 and h_1 are two nonnegative integers called *Betti numbers* that can be attached to any manifold;
- $h_0(M)$ is the number of connected components of M;
- A point has $h_0 = 1$ and $h_1 = 0$;
- A circle has $h_0 = h_1 = 1$.

Plex will return a computation of persistent homology of any data set, whether it has hidden topological structure or not. How do we know when the results of Plex indicate significant topological structure in the data? This question is impossible to answer unless we understand the results Plex returns from unstructured data. The goal of the present report is to report on some experimental results obtained from applying Plex to

- a real-world data set acquired from network traffic at the University of Wisconsin;
- a data set engineered to have the topological structure of a circle;
- a data set engineered to have the topological structure of a point (i.e. no interesting structure at all.)

2. OUR DATA

We wanted to explore the capabilities of Plex in identifying persistent homology in real data. First of all we must be clear that we did not clean the data in any way. Plex and Persistent Homology theory are very sensitive to outliers; if the outliers in your data are truly uninteresting points, it makes sense to clean them out, but in our case we had no a priori reason to believe any of our data points were unreliable.

Our data were byte counts collected over the course of ten days in December from 134 network ports in the internal UW-Madison core computer network. Each measurement represents a byte count over a window of approximately five minutes, yielding a total of 2675 time points. These data represent sixty-seven links in the network; each link between routers is represented by two measurement streams, each corresponding to one direction of network traffic. Each link's data is only collected by one of the two connected routers.

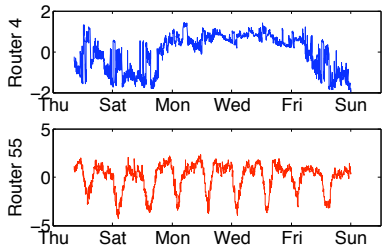


Fig. 1. Example data, logarithmic and normalized. The tick marks are at the midnight point for each day.

In an attempt to normalize the data across ports while maintaining detailed information about low traffic load, we used the logarithm of the traffic measurements, then subtracted the mean and normalized by the variance. Figure 1 shows data on two ports in the network.

3. BARCODE ANALYSIS

The output of Plex is a *barcode*, a collection of intervals for different homology groups. The intervals give the start and end ϵ of a topological feature; longer bars are more "persistent," while shorter bars are transitory and should be ignored. The question of "how long is long enough to be called persistent" is not at all well-understood, and will be a focus of our discussion below.

Consider the following example in order to understand what will happen in a barcode as we vary ϵ. Our data are a set of n equally spaced points around a circle. For small ϵ, the points will not be close enough to each other to be connected, and thus our H_0 barcode will show n bars in this range, denoting n connected components. Our H_1 barcode will show no bars, denoting no holes in the manifold. As soon as ϵ is the distance between the points, the circle will be completely connected and our H_0 barcode will show a single bar, for a single connected component, for the entire rest of the range of ϵ. The H_1 barcode will now show a single bar representing the hole in the middle of the circle. The next transition comes when ϵ reaches the diameter of the circle, at which point all the points will be connected; therefore this ϵ suggests a manifold on which all the data are directly connected with no holes, and so the H_1 bar will disappear.

A central idea of persistent homology is that there should be a large range of ϵ for which the true topology of the data is visible in the barcode. In the example above, this leads to the intuitively natural conclusion that the hypothesis "the data are arrayed along a circle" is more strongly supported when the distance between successive points on the circle is small compared to the diameter of the circle.

4. WHAT COUNTS AS PERSISTENT?

Our goal now was simply to understand better how to differentiate the barcode of a real topological structure from that of random noise.

In order to focus our attention on a more tractable problem, we restricted our attention to routers 4 and 55. Thus we are examining 2675 points in $\mathbb{R}^2$. The resulting scatterplot can be seen in the upper left quadrant of Figure 3. Note that the scatterplot appears to the eye to have a "hole" – this suggests that Plex might find some persistent H^1. We use a Gaussian cloud as our "null hypothesis," a data set which does not have any hidden topological structure. We quantified "persistence" by recording the lengths of the two longest bars in H_1. The plots hereafter display the length of the longest H_1 bar on the x-axis and the second-longest on the y-axis. If the data set is sampled from a circle, one should expect one very long ("persistent") H_1 bar, with all other H_1 bars small. In particular, the ratio between x and y coordinate will be very large.

We first attempted to run Plex using the Vietoris-Rips complex on the entire dataset of 2675 points in $\mathbb{R}^{134}$. Since the Vietoris-Rips tries to build simplicial complexes using all the data points, matlab ran out of memory for this large dataset. We next attempted to run Plex using the Witness complex. This uses landmark points, a small sample of the 2675 data points which, one hopes, will still carry the topological information pertaining to the whole data set. We found, that the barcode output by Plex depended strongly on the choice of landmark set, as seen in Figure 2(a).

4.1. Landmark Selection

Plex gives two options for how to select landmark points out of the dataset: Random and Max-Min. Random simply selects the landmarks uniformly at random without replacement. Max-Min selects the first landmark randomly, then selects each subsequent landmark to maximize the minimum distance from the new landmark to any already-chosen landmark.

Figure 2(a) has length of the longest bar on the x-axis and length of the second longest bar on the y-axis; therefore the points will by definition be on or below the 45 degree line, which is shown for clarity. The further to the right the points are, the longer is the longest bar; the further away from the 45 degree line, the higher the ratio of the longest to the second longest bar. The different points show 50 different landmark selections on the same two data sets.

For the Gaussian cloud, the presence of a few outliers creates long bars when the max-min landmarks are used. Since we want to use a protocol in which the structureless data do *not* have persistent homology features, we used random landmarks for the remainder of our trials.

4.2. Comparing to Noisy circles and Random data

For our second experiment, we compared four data sets, each consisting of 2675 points in $\mathbb{R}^2$ and normalized to have the same mean and variance. The four data sets are shown in Figure 3: We have our router data, the Gaussian cloud, a circle convolved with a low-amplitude Gaussian noise, and a circle convolved with high-amplitude Gaussian noise.

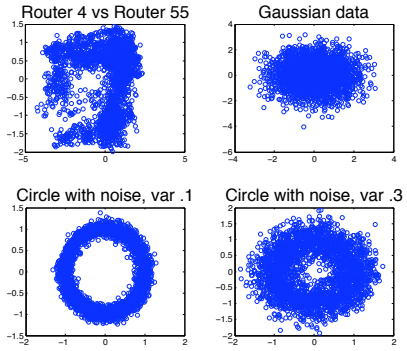


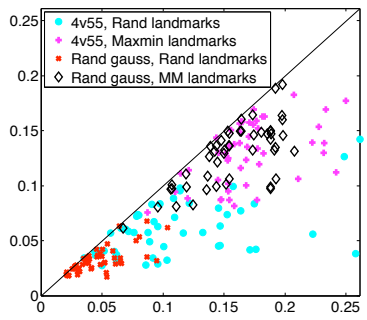
Fig. 3. Scatterplots of four data sets we used for comparison.

We once again looked at the longest bar versus the second longest bar in H_1. As expected, the circle with the small amount of noise gives us the best bars– they are plotted in green in Figure 2(b). The points representing the circle are far to the right, meaning the longest bar is very long, but they are also far from the 45-degree line, meaning the second-longest bar is very short relative to the longest bar. This is the clear topological signature of a circle.

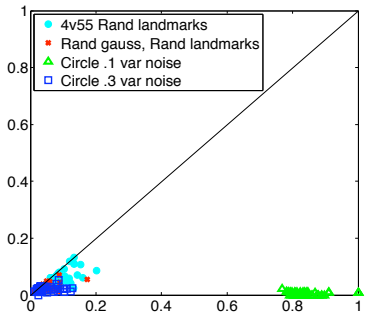
The results for the other three datasets are clustered in the bottom left corner of the plot. Figure 2(c) shows a zoomed-in version of the other plot so that we can see more clearly the relationship of the results to one another. The very noisy circle seems to do better in terms of the ratio between the longest bar and the second longest bar, however our network data does better for the longest bar in H_1.

The main thing to draw from Figure 2(c), however, is that it is not so easy for Plex to distinguish between real-world data, a highly noisy circle, and Gaussian noise. There were many trials where the 4 vs. 55 data set and the highly noisy circle yielded barcodes that would be perfectly consistent with the null hypothesis. This is the case even though the eye can easily see the "hole" in the highly noise circle. By the same token, the Gaussian noise sometimes produced barcodes which look to the eye as if they possess persistent H^1. We must conclude that any appropriate protocol for drawing conclusions about real-world data sets from Plex barcodes based on landmarks is likely to be statistical in nature. In order to develop such a protocol, one will need a much more refined theoretical understanding about the persistent homology of random subsets of $\mathbb{R}^N$ than we have at present. For instance:

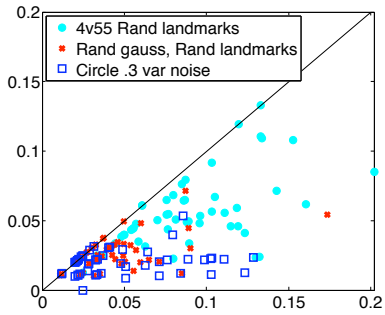
- Let X be the ratio between the longest and second-longest H^1 bar in a set of N points in $\mathbb{R}^2$ drawn from a



(a) Comparison of landmark selection methods. The plot has length of the longest bar on the x-axis and length of the second longest bar on the y-axis; the points will be on or below the 45 degree line, which is shown for clarity.



(b) Longest bar vs. second longest bar length for the four datasets. To build the Witness complex, we used 100 random landmarks per run, and did 40 runs, the result of each of which is shown in the plot.



(c) A zoomed version of Figure 2(b), without the results for the low-noise circle.

Fig. 2. Simulation results on data in $\mathbb{R}^2$ using Plex.

Gaussian distribution. What can we say about the probability distribution on X? In a numerical simulation of 1000 runs of Plex on data from an $\mathcal{N}(0,1)$ distribution, we found that 95% percent of the time, the ratio of longest to shortest bar is at most 3.

- How sensitive is this probability distribution to the choice of model for noise? For instance, what if the N points are drawn from the uniform distribution on a square? Are there universal bounds on the tails of this distribution using only coarse parameters of the distribution from which the points were drawn?

These questions are only beginning to be understood theoretically; see for instance the work of Matthew Kahle [5] and Bubenik-Carlsson-Kim-Luo [2]. It would be interesting and useful to run some larger trials to see, in practice, what kind of barcodes can be expected from random data sets.

Recent work of Carlsson and da Silva also suggests that some of the variation arising from different choices of landmarks can be ameliorated by means of a bootstrapping protocol called "zig-zag," in which the homology features obtained from different choices are considered all at once; this yields a stronger, and potentially more robust notion of persistence [4].

4.3. The meaning (if any) of persistent H^1 in the UW network data

We don't have a convincing explanation for the apparent "circularity" of the scatterplot for router 4 against router 55. Router 4 is attached to a firewall, and it does seem to have traffic increase greatly only during the work week, and it also has less extreme evening dips than most of the data streams do. In order to investigate this we will look at the various flow streams that make up this single data stream and try to identify some subset of flows that are causing this behavior.

5. CONCLUSION

There are three important conclusions to take away from this experiment. The use of Max-Min landmarks should be limited to a situation where you know that outlier data will not affect your results. The results of Plex using landmark points must be viewed in a probabilistic manner; a random Gaussian cloud is capable of producing a barcode that "looks like it indicates topological structure. In order to make a plausible claim of structure, one should require some kind of a "long on average" family of bars in a series of trials. Finally, looking forward, in order to draw principled conclusions from the output of Plex, we need more theoretical results about the barcodes arising from random data sets.

6. REFERENCES

[1] H. Adams et al. Plex: A System for Computational Homology. `http://comptop.stanford.edu/programs/jplex/index.html`. [Online; accessed February 2010].

[2] P. Bubenik, G. Carlsson, P. T. Kim, and Z. Luo. Statistical topology via morse theory, persistence and nonparametric estimation. March 2010. Preprint available at `http://front.math.ucdavis.edu/0908.3668`.

[3] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[4] G. Carlsson and V. de Silva. Zigzag persistence. December 2008. Available at `http://arxiv.org/abs/0812.0197`.

[5] M. Kahle. Random geometric complexes. October 2010. Preprint available at `http://arxiv.org/abs/0910.1649`.