On the Strengthening of Topological Signals in Persistent Homology through Vector Bundle Based Maps

Eric Hanson Francis Motta Chris Peterson Lori Ziegelmeier*

1 Abstract

Persistent homology is a relatively new tool from topological data analysis that has transformed, for many, the way data sets (and the information contained in those sets) are viewed. It is derived directly from techniques in computational homology but has the added feature that it is able to capture structure at multiple scales. One way that this multi-scale information can be presented is through a barcode. A barcode consists of a collection of line segments each representing the range of parameter values over which a generator of a homology group persists. A segment's length relative to the lenght of other segments is an indication of the strength of a corresponding topological signal. In this paper, we consider how vector bundles may be used to re-embed data as a means to improve the topological signal. As an illustrative example, we construct maps of tori to a sequence of Grassmannians of increasing dimension. We equip the Grassmannian with the geodesic metric and observe an improvement in barcode signal strength as the dimension of the Grassmannians increase.

2 Introduction

The need to efficiently extract critical information from large data sets has been growing for decades and is central to a variety of scientific, engineering and mathematical challenges. In many settings, underlying constraints on the data allow it to be considered as a sampling of a topological space. It is a fundamental problem in topological data analysis to develop theory and tools for recovering a topological space from a noisy, discrete sampling. The tools that one might choose to use on a given problem depend on the density, quality, and quantity of the data, on the ambient space from where the sampling is drawn, and on the complexity of the topological space as a sub-object of an ambient space. In this paper, we will focus on data consisting of points sampled from an algebraic variety (the zero locus of a system of polynomials). The data points are obtained using the tools of numerical algebraic geometry. Derived from techniques in homotopy continuation, numerical algebraic geometry allows one to use numerical methods to cheaply sample a large collection of low-noise points from an algebraic set. Persistent homology (PH) allows one to use such a sample to gain insight into the topological structure of the algebraic variety. Implementations of persistent homology are readily available and have been used in a variety of applications, ranging from the analysis of experimental data to analyzing the topology of an algebraic variety. However, as with any algorithm, there are computational limitations. Generally, the time and space required for the persistence computation grows rapidly with the size of the input sample, so the maximum size of a sample is limited. Often, applications of PH start with noisy, real-world data, which may also be limited in size [16]. However, our consideration begins with effectively unlimited, arbitrarily accurate data. Experience shows that as one increases the sample size of a fixed space, the quality of the topological signals produced by PH improves. Since the computational complexity of persistent homology limits the size of a sample, methods of preprocessing data that improve the topological signal, without increasing the sample size, are desirable.

In this paper, we consider how topological re-embeddings affect the topological signal obtained from persistent homology. First, a construction of PH and the inherent challenges of interpreting its output is briefly introduced. Then, we will provide details about the setting in which we have applied this embedding technique, using computational topology to analyze projective algebraic varieties. Lastly, results for a specific example are displayed and interpreted.

3 Background

3.1 Persistent Homology

Beginning with a finite set of data points, which are viewed as a noisy sampling of a topological space, assume one has a way of building the matrix of pairwise distances between points in the data set. From this distance matrix, one constructs a nested sequence of simplicial complexes indexed by a parameter t. Fixing a field $\mathbb{K}$, for each simplicial complex, one builds an associated chain complex of vector spaces over $\mathbb{K}$. The i^{th} homology of the chain complex is a vector space and its dimension corresponds to the i^{th} Betti number, $\beta_i(\mathbb{K})$,

*Colorado State University, Department of Mathematics {hanson, motta, peterson, ziegelme}@math.colostate.edu

of the corresponding topological space. For each pair $t_1 < t_2$, there is a pair of simplicial complexes, S_{t_1} and S_{t_2}, and an inclusion map $j : S_{t_1} \hookrightarrow S_{t_2}$. This inclusion map induces a chain map between the associated chain complexes which further induces a linear map between the corresponding i^{th} homology vector spaces. For each i, the totality of the collection of i^{th} homology vector spaces and induced linear maps can be encoded as a graded $\mathbb{K}[t]$-module known as the persistence module. The i^{th} bar code is a way of presenting the invariant factors of the persistence module. As the invariant factors of the persistence module directly relate to the Betti numbers, from the bar code one can visualize the Betti numbers as a function of the scale, t, and can visualize the number of independent homology classes that persist across a given time interval $[t_i, t_j]$. For foundational material and overviews of computational homology in the setting of persistence, see [8, 21, 12, 6, 9, 20, 15].

One commonly used method for building a nested sequence of simplicial complexes from a distance matrix is through a *Vietoris-Rips* complex [12]. This is done by first building the 1-skeleton of the simplicial complex then determining the higher dimensional faces as the clique complex of the 1-skeleton. More precisely, fix $t > 0$, a collection of points X, and a metric, $d(x_i, x_j)$ for x_i, $x_j \in X$. The 1-skeleton of the Vietoris-Rips complex, $C_t(X)$, is defined by including the edge $x_i x_j \in C_t(X)$ if $d(x_i, x_j) \leq t$. A higher dimensional face is included in $C_t(X)$ if all of its lower dimensional sub-faces are in $C_t(X)$. In other words, the abstract k-simplices of $C_t(X)$ are given by unordered $(k+1)$-tuples of sample points whose pairwise distances do not exceed the parameter t.

Given a collection of data points, the resulting Vietoris-Rips complex, and its homology, is highly dependent on the choice of parameter t. To reconcile this ambiguity, persistence exploits that if $t_1 < t_2$ then C_{t_1} is a sub simplicial complex of C_{t_2}. In other words, as t grows so do the Vietoris-Rips complexes, giving an inclusion from earlier complexes to those which appear later. The idea then is to not only consider the homology for a single specified choice of parameter, but rather track topological features through a range of parameters [12]. Those which persist over a large range of values are considered signals of underlying topology, while the short lived features are taken to be noise inherent in approximating a topological space with a finite sample [10].

For clarity, consider 4 points in the plane with distance matrix

$$\begin{bmatrix} 0 & t_2 & t_5 & t_3 \\ t_2 & 0 & t_1 & t_6 \\ t_5 & t_1 & 0 & t_4 \\ t_3 & t_6 & t_4 & 0 \end{bmatrix}.$$

We label the points a, b, c and d and build the sequence of Vietoris-Rips simplicial complexes up to $\mathbf{C}_{t_5}$. Table 1 shows the Betti information (where β_i is the dimension of the i^{th} homology vector space) for the example illustrated in Figure 1 over the range of parameter values $t \geq 0$.[1]

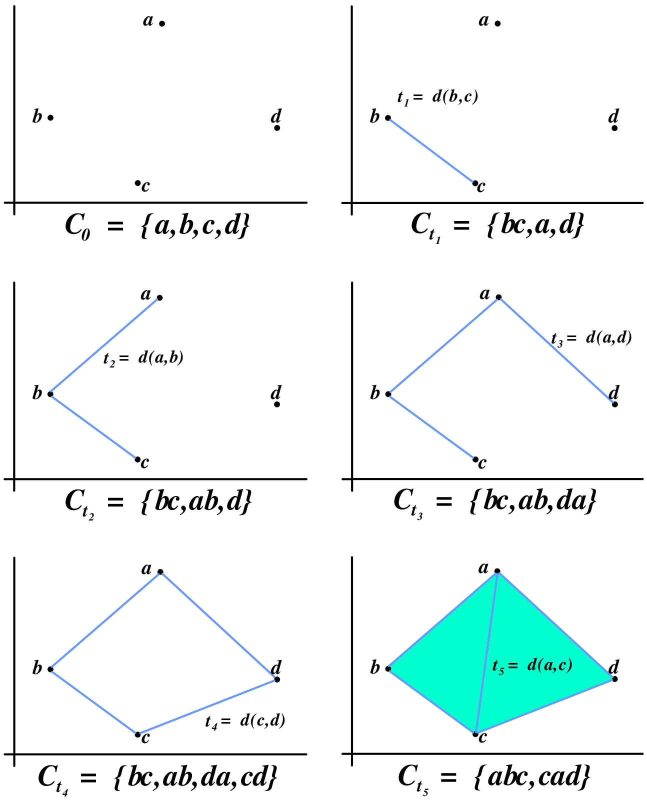


Figure 1: A sequence of Vietoris-Rips simplicial complexes shown geometrically and abstractly along with their maximal faces.

filtration times (t)	β_0	β_1
$0 \leq t < t_1$	4	0
$t_1 \leq t < t_2$	3	0
$t_2 \leq t < t_3$	2	0
$t_3 \leq t < t_4$	1	0
$t_4 \leq t < t_5$	1	1
$t_5 \leq t$	1	0

Table 1: Persistent homology data

Even in this simple example, the amount of information created by the persistent homology computation is non-trivial. Furthermore, an effective rendering of the complexes in Figure 1 is only possible because there are very few points in the example. In the 4-point example, at time t_6 the simplicial complex $\mathbf{C}_{t_6}$ becomes three-

[1] For finite data there will only be finitely many parameter values where the simplicial complex changes.

dimensional. As the vertex set or the dimension of the ambient space grows, visualizing the sequence of complexes is not practical.

The *barcode* is a visual method for presenting some of the homological information in a sequence of chain maps. In particular, it displays the structure of the invariant factors of the i^{th} persistence module. Figure 2 is the barcode corresponding to the example of the four points in the plane described in Figure 1.

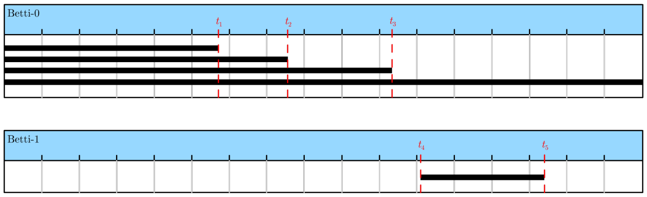


Figure 2: Barcodes corresponding to Figure 1

The computational requirements of the persistence computation is related to the sample size. It is often the case that computing the persistent homology using the Rips filtration is impractical. There is an alternative construction, introduced by Carlsson and de Silva, called the *witness complex* [7, 13]. Starting with a large sample set X, one picks a distinguished subset $L \subset X$ of *landmark points*. The witness complex is a family of simplicial complexes built on L using information from the entire set X.

To build the witness complex, first use the landmark set to assign to each point $x \in X$ the numbers $m_k(x)$ corresponding to the distance from x to its $(k+1)$-th nearest landmark point. For each integer k $(0 < k < |X|)$ and vertices $\{l_{j_i} | 0 \leq i \leq k\} \subset L$, include the k-simplex $[l_{j_0} l_{j_1} ... l_{j_k}]$ in the complex (at time t) if there exists a point $x \in X$ such that $\max\{d(l_{j_i}, x) | 0 \leq i \leq k\} \leq t + m_k(x)$, and if all of its faces are in the complex [1].

The output of the witness filtration is sensitive to the choice of landmark set. One technique for choosing a landmark set, called sequential maxmin, is implemented in the freely distributed persistent homology software package JPlex [17]. The procedure for using sequential maxmin is to first pick a point $l_0 \in X$ then inductively choose the i-th landmark point from X by choosing the point furthest from the set of $(i-1)$ points already chosen. In practice, this seems to produce a stronger topological signal than choosing L randomly, so it is the method we will utilize.

3.2 Algebraic Varieties and Numerical Algebraic Geometry

A motivating problem for this paper is the computation of the Betti numbers of a complex projective algebraic variety from numerically obtained sample points. The method we use to obtain sample points derive from several algorithms in numerical algebraic geometry.

The term *numerical algebraic geometry* is often used to describe a wide ranging set of numerical methods to extract algebraic and geometric information from polynomial systems. The field includes a diverse collection of algorithms (both numeric and numeric-symbolic). The class of numerical algorithms that we use are rooted in homotopy continuation. The idea of homotopy continuation is to link a pair of polynomial systems through a deformation and to relate features of the two systems through this deformation. For example, one can track known, isolated, complex solutions of one polynomial system to unknown, complex solutions of a second polynomial system through a deformation of system parameters.

Let Z be the complex algebraic variety associated to an ideal in $\mathbb{C}[z_1, \ldots, z_N]$. With numerical homotopy continuation methods combined with monodromy breakup, it is practical to produce sets of numerical data points which numerically lie on each of the irreducible components of Z [19, 18].

There are several important features of the methods of numerical algebraic geometry that are worth highlighting. The first feature is the ability to refine sample points to arbitrarily high precision via Newton's method. A second feature is the ability to produce an arbitrary number of sample points on any given component. A third feature is the parallelizability of these numerical methods. For instance, 10,000 processors could be used in parallel to track 10,000 paths and could be used in parallel to refine the accuracy of each sample point to arbitrarily high precision. The basic algorithms of numerical algebraic geometry (including monodromy breakup) are implemented in the freely available software package, Bertini [4].

It is important to note that sampling is computationally inexpensive, so obtaining large sample sets does not pose a significant challenge. However, it is not clear that this sampling technique will provide points that are well distributed for the purpose of persistent homology computations.

4 Main Idea

4.1 Theory

By its very nature, persistent homology characterizes intrinsic topological features which should be relatively insensitive to the metric used to build a pairwise distance matrix. However, experiments show that the signal *strength* is impacted by the choice of metric. In our experience, even if the topological features remain the same, the ability to correctly interpret information from a barcode depends on the strength of the signal. We will consider the barcode signal strength of mappings of an algebraic variety into various Grassmannians.

The Grassmannian $Gr(n, k)$ is a manifold parametrizing all k dimensional subspaces of a fixed n dimensional vector space. The Grassmann manifold $Gr(n+1, 1)$ is the projective space $\mathbb{P}^n$, and from this vantage point, Grassmannians can be viewed as generalizations of projective spaces. These manifolds can be given a topological structure, a differential structure and even the structure of a projective variety (e.g. via the Plucker embedding).

Points in an n-dimensional projective space correspond to 1-dimensional subspaces of a fixed $(n+1)$-dimensional vector space. A natural notion of distance is given by the smallest angle between the subspaces. We would like to define the distance between points on other Grassmannians by extending this definition. As a starting point, it can be shown that any unitarily invariant metric on a Grassmannian can be written in terms of the principal angles between the corresponding subspaces. The principal angles between a pair of subspaces A, B in $\mathbb{C}^n$ can be determined as follows. First, determine matrices M and N whose columns form orthonormal bases for A and B. Next, determine the singular value decomposition $M^*N = U\Sigma V^*$. The singular values of M^*N are the diagonal entries of Σ. These singular values are the cosines of the principal angles between A and B (see [5]). If A and B are k-dimensional, then there will be principal angles $\Theta(A, B) = (\theta_1, \theta_2, \ldots, \theta_k)$ with $0 \leq \theta_1 \leq \theta_2 \leq \cdots \leq \theta_k \leq \pi/2$. There are many common metrics computed as functions of the principal angles [2]. For instance, the Fubini-Study metric induced by the Plucker embedding is $d_{FS}(A, B) = \cos^{-1}\left(\prod_{i=1}^{k} \cos\theta_i\right)$. We have found that the Fubini-Study metric does not, in general, yield a strong signal, and instead, we restrict our attention to the geodesic distance

$$d(A, B) = \sqrt{\theta_1^2 + \ldots + \theta_k^2}.$$

Since we wish to compare the effect of considering a sample in various Grassmannian embeddings, it remains to define what we mean by relative topological signal strength. Imagine we know that our sample was taken from a topological space whose i^{th} Betti number is b_i. Assuming that the b_i longest segments in the barcode represent these topological features, we will measure signal strength as the ratio of the sum of the lengths of the b_i longest segments to the sum of the total length of all the segments in the i^{th} Betti barcode, including noise. Note that noise consisting of many segments of total length m and noise consisting of a single segment of length m cannot be distinguished by this statistic. To cope with this limitation we also consider the ratio of the length of the b_i^{th} longest segment to the $(b_i+1)^{th}$ longest segment in the barcode.

4.2 Embeddings into the Grassmannian

Consider a complex projective curve, $C \subset \mathbb{P}^2$, defined by the zero locus of a homogenous polynomial $F(x, y, z)$. When we think of the zero set as a projective variety, then each point, $[x : y : z]$ on C, corresponds to a 1-dimensional subspace of $\mathbb{C}^3$ (note that the homogeneity of the equation leads to the conclusion that if (x, y, z) is a solution then so is (cx, cy, cz) for any $c \in \mathbb{C}$). Thus, points on a projective variety correspond to one-dimensional subspaces of $\mathbb{C}^3$ constrained to lie on the vanishing locus of a homogeneous polynomial. From this point of view, C is a sub-object of $\mathbb{P}^2 = Gr(3, 1)$. We can sample random points on C with several different methods. If we wish to build a distance matrix from these points, then we should consider the distance between a pair of points as the principal angle between the one dimensional spaces to which they correspond.

Consider the matrix

$$E(x, y, z) := \begin{bmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{bmatrix},$$

and observe that for any point $(x, y, z) \neq (0, 0, 0)$, the rank of $E(x, y, z)$ is 2. This can be seen by observing that the determinant of $E(x, y, z)$ is identically zero and that the locus of conditions such that all 2×2 minors of $E(x, y, z)$ are zero force $x = y = z = 0$. For each value of (x, y, z), we consider the row space of $E(x, y, z)$. Note also that

$$\begin{bmatrix} 0 & z & -y \\ -z & 0 & x \\ y & -x & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

As a consequence, the row space of $E(x, y, z)$ is the same as the row space of $E(cx, cy, cz)$, and $E(x, y, z)$ can be viewed as a rule for attaching a smoothly varying 2 dimensional subspace to each point of $\mathbb{P}^2$. In other words, $E(x, y, z)$ determines a rank two vector bundle on $\mathbb{P}^2$. For each one dimensional subspace of $\mathbb{C}^3$, we can determine a 2-dimensional subspace of $\mathbb{C}^3$ by mapping it to the row space of $E([x : y : z])$. If $\Phi_0 : \mathbb{P}^2 \to Gr(3, 2)$ denotes the image of this map, then by restriction this gives a map $\phi_0 : C \to Gr(3, 2)$.

For each integer $k > 0$, consider the set of monomials in x, y, z of degree k. We construct new matrices, $E_k(x, y, z)$, by concatenating matrices of the form $m_i \cdot E(x, y, z)$ for each degree k monomial m_i. For example, $E_1(x, y, z)$ is the matrix

$$\begin{bmatrix} 0 & xz & -xy & 0 & yz & -y^2 & 0 & z^2 & -zy \\ -xz & 0 & x^2 & -yz & 0 & yx & -z^2 & 0 & zx \\ xy & -x^2 & 0 & y^2 & -yx & 0 & zy & -zx & 0 \end{bmatrix}.$$

For each k, $E_k(x, y, z)$ has constant rank 2 on $\mathbb{P}^2$ and can be used to define a map $\phi_k : C \to Gr(N_k, 2)$ (where

N_k is the number of columns of $E_k(x, y, z)$). Geometrically, the columns of $E_k(x, y, z)$ corresponds to a "spanning set for the space of sections of the twisted tangent bundle, $\mathbf{T}_{\mathbb{P}^2}(k-1)$". In this way, we can consider images of C, in increasingly large Grassmannians via the maps $\phi_0, \phi_1, \phi_2, \ldots$. It can be shown that for each k, Φ_k embeds $\mathbb{P}^2$ into $Gr(N_k, 2)$ and that ϕ_k embeds C into $Gr(N_k, 2)$.

4.3 Example

Consider the complex projective elliptic curve, $C \subset \mathbb{P}^2$ defined by the equation

$$x^2y + y^2z + z^2x = 0. \tag{1}$$

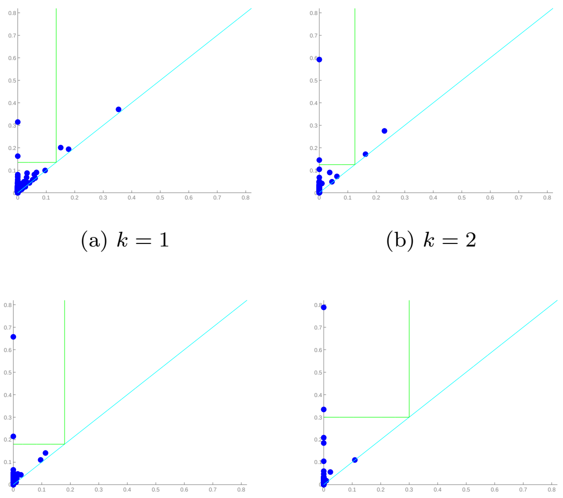
Topologically, C is a torus whose Betti numbers are $\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$.

Using Bertini, we sampled 10,000 points satisfying Equation 1. We mapped each point to $Gr(N_k, 2)$ using ϕ_k, for $k = 1, \ldots, 10$. From these 10,000 points we fixed 100 landmark sets, L_i (of size 200) using the sequential maxmin algorithm with random initial points $l_{0,i}$ for $i = 1, \ldots, 100$. For each embedding $\phi_k(C)$ and for each of the fixed landmark sets, we compute the persistent homology barcodes for the zeroth and first Betti numbers using the witness complex construction.

Using the geodesic distance to measure distances between points, Figure 3 shows prototypical Betti-1 barcodes for the images of the 10,000 points in $Gr(N_k, 2)$. In the figure, each segment in the barcode is plotted as a point in the (x, y)-plane with the x-coordinate corresponding to the starting parameter and the y-coordinate corresponding to the ending parameter. Short segments (i.e. topological noise) appear near the $y = x$ line. Notice that as we move the elliptic curve to Grassmannians of higher degree, the two longest segments in the barcode grow in length while the number and lengths of the other segments decrease.

In Figure 4, we plot the relative signal strength of the Betti-1 barcodes, as measured by the ratio of the sum of the length of the two longest segments to the total sum of lengths of all segments, averaged over all landmark sets for each embedding. We observe an increase from approximately 10% to 55% of the total length of the barcodes being accounted for in the longest two segments. We also observe that the improvement of the relative signal strength levels-off after $k = 5$.

Figure 5 compares the second longest segment in the Betti-1 barcode (corresponding to a topological circle) to the third longest segment (representing topological noise). We notice a sharp increase in this measure of signal strength followed by a similarly steep decrease. Together with the content of Figure 4 this indicates that after $k = 5$, the relative length of the longest Betti-1 segment remains somewhat unchanged while the disparity in the lengths of the second and third longest segments is diminished.



(a) $k = 1$

(b) $k = 2$

(c) $k = 3$

(d) $k = 4$

Figure 3: Betti-1 barcodes for each of the four specified embeddings.

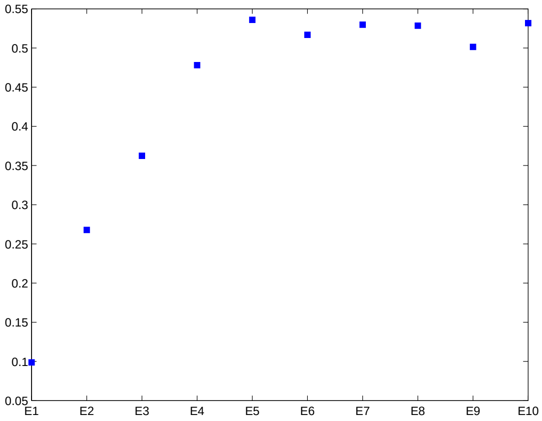


Figure 4: Average ratio of the sum of the longest two barcode lengths to the sum of the lengths of all barcodes for $k = 1, \ldots, 10$.

It is worth noting that there is a diminished signal strength from the projective variety to the embeddings in the Grassmannian. Our aim, however, is to compare topological signal strength improvement across successive Grassmannian embeddings.

5 Conclusion

Using the techniques of numerical algebraic geometry, we can sample arbitrarily many points, to an arbitrary degree of accuracy, on any prescribed component of an

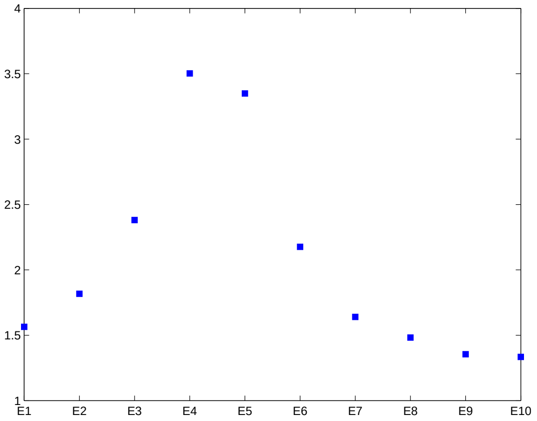


Figure 5: Average ratio of the second longest barcode to the third longest barcode for $k = 1, \ldots, 10$.

algebraic set. Using twists of the tangent bundle to projective space, we can map these points to a sequence of Grassmann manifolds of increasing dimension. With techniques of computational homology, we can build the persistence module and decompose the module into its invariant factors. A visual plot of the starting and ending points of the invariant factors aids in the understanding of the underlying variety as a topological space. Higher embeddings of the data seem to strengthen the topological signal.

For further research, we intend to develop improved sampling techniques for algebraic varieties. We will also conduct experiments to determine if alternate vector bundles or alternate metrics on the Grassmannian can be used to strengthen topological signals.

References

[1] H. Adams, JPlex with Matlab Tutorial. (2011) comptop.stanford.edu/programs/jplex/files/

[2] A. Barg and D.Yu. Nogin, Bounds on packings of spheres in the Grassmann manifold. IEEE Trans. on Info. Theory. 48 (2002), 2450-2454.

[3] D.J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler, Adaptive multiprecision path tracking. *SIAM J. Numer. Anal.* 46 (2008), 722-746.

[4] D.J. Bates, J. D. Hauenstein, A. J. Sommese, and C. W. Wampler, Bertini: Software for numerical algebraic geometry. http://www.nd.edu/∼sommese/bertini.

[5] A. Bjorck and G. Golub, Numerical methods for computing angles between linear subspaces. *Mathematics of Computation* 27, (1973), no 123, 579-594.

[6] G. Carlsson, Topology and data. Bulletin of the American Mathematical Society, Vol. 46 (2009), no. 2, 255308.

[7] V. de Silva and G. Carlsson, Topological estimation using witness complexes. *SPBG '04 Symposium on Point-Based Graphics* (2004), 157-166.

[8] H. Edelsbrunner and J. Harer, Persistent homology - a survey. *Contemporary Math* 453 (2008), 257-282.

[9] H. Edelsbrunner and J. Harer, Computational Topology: An Introduction. American Mathematical Society, Providence, RI, 2010. xii+241 pp.

[10] H. Edelsbrunner, D. Letscher, and A. Zomorodian, Topological persistence and simplification. Discrete Computational Geometry, 28:4 (2002), 511-533.

[11] A. Galantai and Cs. J. Hegedus, Jordan's principal angles in complex vector spaces. Numer. Linear Algebra Appl. 13 (2006), 589-598.

[12] R. Ghrist, Barcodes: The persistent topology of data. Bulletin of the American Mathematical Society, Vol. 45 (2008), pp. 61-75.

[13] L. Guibas and S. Oudot, Reconstruction using witness complexes. *Proc. 18th ACM-SIAM Sympos. on Discrete Algorithms* (2007).

[14] A. Hatcher, Algebraic Topology, Cambridge University Press (2002).

[15] T. Kaczynski, K. Mischaikow, and M. Konstantin, Computational Homology. Applied Mathematical Sciences 157. Springer-Verlag, New York, 2004. 480 pp.

[16] D. Mumford, A. Lee, and K. Pederson, The non-linear statistics of high-contrast patches in natural images. Itnl. J. Computer Vision. 54 (2003), 83-103.

[17] H. Sexton and M. Vejdemo-Johansson, JPlex simplicial complex library. http://comptop.standord.edu/programs/jplex/.

[18] A.J. Sommese and C.W. Wampler, The Numerical Solution of Systems of Polynomials. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.

[19] A.J. Sommese, J. Verschelde, and C.W. Wampler, Numerical decomposition of the solution sets of polynomials into irreducible components. *SIAM J. Numer. Anal.* 38 (2001), 2022-2046.

[20] A. Zomorodian, Topology for computing. Cambridge Monographs on Applied and Computational Mathematics, 16. Cambridge University Press, Cambridge, 2005. xiv+243 pp.

[21] A. Zomorodian and G. Carlsson, Computing persistent homology. *Discrete Comput. Geom.* 33 (2005), no. 2, 249274