Topology and Data

Gunnar Carlsson [*]
Department of Mathematics, Stanford University
Stanford, California 94305

October 2, 2008

1 Introduction

An important feature of modern science and engineering is that data of various kinds is being produced at an unprecedented rate. This is so in part because of new experimental methods, and in part because of the increase in the availability of high powered computing technology. It is also clear that the *nature* of the data we are obtaining is significantly different. For example, it is now often the case that we are given data in the form of very long vectors, where all but a few of the coordinates turn out to be irrelevant to the questions of interest, and further that we don't necessarily know which coordinates are the interesting ones. A related fact is that the data is often very high-dimensional, which severely restricts our ability to visualize it. The data obtained is also often much noisier than in the past, and has more missing information (missing data). This is particularly so in the case of biological data, particularly high throughput data from microarray or other sources. Our ability to analyze this data, both in terms of quantity and the nature of the data, is clearly not keeping pace with the data being produced. In this paper, we will discuss how geometry and topology can be applied to make useful contributions to the analysis of various kinds of data. Geometry and topology are very natural tools to apply in this direction, since geometry can be regarded as the study of distance functions, and what one often works with are distance functions on large finite sets of data. The mathematical formalism which has been developed for incorporating geometric and topological techniques deals with point clouds, i.e. finite sets of points equipped with a distance function. It then adapts tools from the various branches of geometry to the study of point clouds. The point clouds are intended to be thought of as finite samples taken from a geometric object, perhaps with noise. Here are some of the key points which come up when applying these geometric methods to data analysis.

- **Qualitative information is needed:** One important goal of data analysis is to allow the user to obtain *knowledge* about the data, i.e. to understand how it is organized on a large scale. For example, if we imagine that we are looking at a data set constructed somehow from diabetes patients, it would be important to develop the understanding that there are two types of the disease, namely the juvenile and adult onset forms. Once that is established, one of course wants to develop quantitative methods for distinguishing them, but the first insight about the distinct forms of the disease is key.

- **Metrics are not theoretically justified:** In physics, the phenomena studied often support clean explanatory theories which tell one exactly what metric to use. In biological problems, on the other hand, this is much less clear. In the biological context, notions of distance are constructed using some

[*] Research supported in part by DARPA HR 0011-05-1-0007

intuitively attractive measures of similarity (such as BLAST scores or their relatives), but it is far from clear how much significance to attach to the actual distances, particularly at large scales.

- **Coordinates are not natural:** Although we often receive data in the form of vectors of real numbers, it is frequently the case that the coordinates, like the metrics mentioned above, are not natural in any sense, and that therefore we should not restrict ourselves to studying properties of the data which depend on any particular choice of coordinates. Note that the variation of choices of coordinates does not require that the coordinate changes be rigid motions of Euclidean space. It is often a tacit assumption in the study of data that the coordinates carry more intrinsic meaning than they actually do.
- **Summaries are more valuable than individual parameter choices:** One method of clustering a point cloud is the so-called *single linkage clustering*, in which a graph is constructed whose vertex set is the set of points in the cloud, and where two such points are connected by an edge if their distance is $\leq \epsilon$, where ϵ is a parameter. Some work in clustering theory has been done in trying to determine the optimal choice of ϵ, but it is now well understood that it is much more informative to maintain the entire *dendrogram* of the set, which provides a summary of the behavior of clustering under all possible values of the parameter ϵ at once. It is therefore productive to develop other mechanisms in which the behavior of invariants or construction under change of parameters can be effectively summarized.

In this paper, we will discuss methods for dealing with the properties and problems mentioned above. The underlying idea is that methods inspired by topology should address them. For each of the points above, we describe why topological methods are appropriate for dealing with them.

- Topology is exactly that branch of mathematics which deals with qualitative geometric information. This includes the study of what the connected components of a space are, but more generally it is the study of connectivity information, which includes the classification of loops and higher dimensional surfaces within the space. This suggests that extensions of topological methodologies, such as homology, to point clouds should be helpful in studying them qualitatively.
- Topology studies geometric properties in a way which is much less sensitive to the actual choice of metrics than straightforward geometric methods, which involve sensitive geometric properties such as curvature. In fact, topology ignores the quantitative values of the distance functions, and replaces it with the notion of infinite nearness of a point to a subset in the underlying space. This insensitivity to the metric is useful in studying situations where one only believes one understands the metric in a coarse way.
- Topology studies only properties of geometric objects which do not depend on the chosen coordinates, but rather on intrinsic geometric properties of the objects. As such, it is coordinate-free.
- The idea of constructing summaries over whole domains of parameter values involves understanding the relationship between geometric objects constructed from data using various parameter values. The relationships which are useful involve continuous maps between the different geometric objects, and therefore become a manifestation of the notion of *functoriality*, i.e the notion that invariants should be related not just to objects being studied, but also to the maps between these objects. Functoriality is central in algebraic topology, in that the functoriality of homological invariants is what permits one to compute them from local information, and that functoriality is at the heart of most of the interesting applications within mathematics. Moreover, it is understood that most of the information about topological spaces can be obtained through diagrams of discrete sets, via a process of simplicial approximation.

The last point above, concerning functoriality, is critical. In developing methods to address the first two points, we find that we are forced to make functorial geometric constructions and analyze their behavior on maps even to obtain information about single point clouds. Functoriality has proven itself to be a powerful tool in the development of various parts of mathematics, such as Galois theory within algebra, the theory of Fourier series within harmonic analysis, and the applicaton of algebraic topology to fixed point questions in topology. We argue that, as suggested in [46], it has a role to play in the study of point cloud data as well, and we give two illustrations of how this could happen, within the context of clustering.

Informally, clustering refers to the process of partitioning a set of data into a number of parts or clusters, which are recognizably distinguishable from each other. In the context of finite metric spaces, this means roughly that points within the clusters are nearer to each other than they are to points in different clusters. Clustering should be thought of as the statistical counterpart to the geometric construction of the path-connected components of a space, which is the fundamental building block upon which algebraic topology is based. There are many schemes which construct clusterings based on metric information, such as single, average, and complete linkage clustering, k-means clustering, spectral clustering, etc. (see [31]). Although clustering is clearly a very important part of data analysis, the ways in which it is formulated and implemented are fraught with ambiguities. In particular, the arbitrariness of various threshhold choices and lack of robustness are difficulties one confronts. Much of current research efforts are focused in this direction (see e.g. [43] and [39]), and functoriality provides the right general mathematical framework for addressing them. For example, one can construct data sets which have been threshholded at two different values, and the behavior of clusters under the inclusion of the set with tighter threshhold into the one with the looser threshhold is informative about what is happening in the data set. We present two additional examples of how functoriality could be used in analyzing some questions related to clustering.

Example: In the case of very large X, it may often be difficult to apply the clustering algorithm to a full data set, and one may instead find it desirable to cluster subsamples from X. One is then confronted with the task of attempting to verify that the clustering of the subsample is actually representative of a clustering of the full data set X. One way of proceeding is to construct two samples from X, and hoping that they are consistent in an appropriate sense. One version of this idea would be to consider the subsamples X_1 and X_2, together with their union $X_1 \cup X_2$. One could apply the clustering scheme to each of these sets individually, and suppose we denote the set of clusters for the three sets X_1, X_2, and $X_1 \cup X_2$ by $C(X_1)$, $C(X_2)$, and $C(X_1 \cup X_2)$ respectively. If the clustering scheme were functorial, i.e. if inclusions of data sets induced maps of the collections of clusters, then one would have a diagram of sets

$$\begin{tikzcd} & C(X_1 \cup X_2) & \\ C(X_1) \arrow[ur] & & C(X_2) \arrow[ul] \end{tikzcd}$$

If the clusterings are consistent, i.e. if the clusters in $C(X_1)$ and $C(X_2)$ in $C(X_1 \cup X_2)$ correspond well under these maps, one can regard that as evidence that the subsample clusterings actually correspond to clusterings on the full data set X. Of course, what the phrase "correspond well" means is not well defined here. Later in the paper, we will discuss a way to attach more quantitative information to questions of this type.

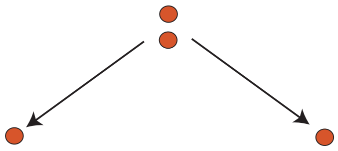
Example: Suppose that we have data X which varies with time. One could then ask for information concerning the behavior of clusterings produced by clusterings over time. Clusters can appear, vanish, merge, or split into separate clusters. The analysis of this behavior can be studied using functoriality. For $t_0 < t_1$, we let $X[t_0, t_1]$ denote the set of points in the data set occurring between times t_0 and t_1. If we have $t_0 < t_1 < t_2 < t_3$, then we have a diagram of point cloud data sets

$$\begin{tikzcd}
& X[t_1, t_2] \arrow[dl] \arrow[dr] & \\
X[t_0, t_2] & & X[t_1, t_3]
\end{tikzcd}$$

If the clustering scheme is functorial in the sense of the preceding example, we obtain the corresponding diagram of sets

$$\begin{tikzcd}
& C(X[t_1, t_2]) \arrow[dl] \arrow[dr] & \\
C(X[t_0, t_2]) & & C(X[t_1, t_3])
\end{tikzcd}$$

This set contains information about the behavior over time of the clusters. For example, the diagram



would correspond to a single cluster at time t_0, which breaks into two clusters in the interval $[t_1, t_2]$, which in turn merge back again in the interval $[t_2, t_3]$.

This paper will deal with a number of methods for thinking about data using topologically inspired methods. We begin with a discussion of persistent homology, which is a mathematical formalism which permits us to infer topological information from a sample of a geometric object, and show how it can be applied to a particular data sets arising from natural image statistics and neuroscience. Next, we show that topological methods can produce a kind of imaging of data sets, not by embedding in Euclidean space but rather by producing a simplicial complex associated to certain initial information about the data set. We then demonstrate that persistence can be generalized in several different directions, providing more structure and information about the data sets in question. We then show that the philosophy of functoriality can be used to reason about the nature of clustering methods, and conclude by speculating about theorems one might hope to prove and discussing how the subject might develop more generally.

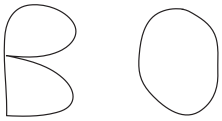
The author is extremely grateful to his collaborators A. Blumberg, A. Collins, V. de Silva, L. Guibas, T. Ishkanov, F. Memoli, M. Nicolau, D. Ringach, G. Sapiro, H. Sexton, G. Singh, and A. Zomorodian. In addition, he is grateful to a number of people for very useful conversations on this subject. They include T. Beke, P. Diaconis, R. Ghrist, S.Holmes, K. Mischaikow, P. Niyogi, S. Oudot, S. Smale, and S. Weinberger.

2 Persistence and homology

2.1 Introduction

In thinking about qualitative properties of spaces X, an obvious one which comes to mind is its decomposition into path connected components. It is a partition of X, and the cardinality of the collection of blocks in this partition is an invariant of X, and surely deserves to be called a qualitative invariant of X. There are other qualitative properties to be considered.

Example: Consider the two spaces in the image below.



We note that both spaces are path-connected, but we can see that they are qualitatively distinct in that the letter "B" on the left has two essentially different loops and the "O" on the right has only one. This property is preserved under continuous deformations, and so if one can formalize it into a precise mathematical statement one can then rigorously distinguish between the two spaces. We refer to information about loops and higher dimensional analogues in a space as "connectivity information". The decomposition into path connected components would be regarded as zeroth level connectivity information, loops as level one connectivity information, and so forth. The mathematical formalism which makes these notions precise is *algebraic topology.* It provides signatures which capture the intuitive notions of essential loops or essential higher dimensional surfaces within a space. We describe the output of the formalism. See [33] for a thorough treatment. We recall that two continuous maps $f, g : X \to Y$ are said to be *homotopic* if there is a continuous map $H : X \times [0,1] \to Y$ so that $H(x,0) = f(x)$ and $H(x,1) = f(1)$.

- **Definition:** For any topological space X, abelian group A, and integer $k \geq 0$, there is assigned a group $H_k(X, A)$.
- **Functoriality:** For any A and k as above, and any continuous map $f : X \to Y$, there an induced homomorphism $H_k(f, A) : H_k(X, A) \to H_k(Y, A)$. One has $H_k(f \circ g, A) = H_k(f, A) \circ H_k(g, A)$ and $H_k(Id_X; A) = Id_{H_k(X,A)}$. These conditions are called collectively *functoriality.* We refer the reader to [44] for a treatment of categories and functors.
- **Homotopy invariance:** If f and g are homotopic, then $H_k(f, A) = H_k(g, A)$. It follows that if X and Y are homotopy equivalent, then $H_k(X, A)$ is isomorphic to $H_k(Y, A)$.
- **Normalization:** $H_0(*, A) \cong A$, where $*$ denotes the one point space.
- **Betti numbers:** For any field F, $H_k(X, F)$ will be a vector space over F. Its dimension, if it is finite dimensional, will be written as $\beta_k(X, F)$, and will be referred to as the k-th *Betti number* with coefficients in F. The k-th Betti number corresponds to an informal notion of number of independent k-dimensional surfaces. If two spaces are homotopy equivalent, then all their Betti numbers are equal.

Example: For any topological space X with a finite number of path components, $\beta_0(X)$ is the number of path components.

Example: The first Betti number β_1 of the letter "B" above is two, and for the letter "O" it is one. In this case, it provides a formalization of the count of the number of loops present in the space.

The actual definition of homology which applies to all topological spaces (singular homology) was introduced in [27]. It relies on the linear algebra of infinitely generated modules over the ring $\mathbb{Z}$ in defining homology groups, and for this reason it is not useful from a computational point of view. Computations can be carried out by hand using a variety of techniques (long exact sequence of a pair, long exact Mayer-Vietoris sequence, excision theorem, spectral sequences), but direct computation from the definition is not feasible for general spaces. However, when one is given a space equipped with particular structures, there are often finite linear algebra problems which produce correct answers, i.e. answers which agree with the singular technique. A particularly nice example of this applies when the space in question is described as a *simplicial complex.*

Definition 2.1 *An* abstract simplicial complex *is a pair (V, Σ), where V is a finite set, and Σ is a family of non-empty subsets of V such that $\sigma \in \Sigma$ and $\tau \subseteq \sigma$ implies that $\tau \in \Sigma$. Associated to a simplicial complex is a topological space $|(V, \Sigma)|$, which may be defined using a bijection $\phi : V \to \{1, 2, \ldots, N\}$ as the subspace of $\mathbb{R}^N$ given by the union $\bigcup_{\sigma \in \Sigma} c(\sigma)$, where $c(\sigma)$ is the convex hull of the set $\{e_{\phi(s)}\}_{s \in \sigma}$, where e_i denotes the $i-th$ standard basis vector.*

Intuitively, a simplicial complex structure on a space is an expression of the space as a union of points, intervals, triangles, and higher dimensional analogues. Simplicial complexes provide a particularly simple combinatorial way to describe certain topological spaces. For this reason, one often attempts to approximate (in various senses) topological spaces by simplicial complexes. The key point for this section, though, is that for simplicial complexes, the homology can be computed using only linear algebra of finitely generated $\mathbb{Z}$-modules. We describe this in detail. Given any simplicial complex $X = (V, \Sigma)$, we write Σ_k for the subset of Σ consisting of all $\sigma \in \Sigma$ for which $\#(\sigma) = k + 1$. Elements of Σ_k are referred to as k-simplices. We define the group of k-chains in X as the group of formal linear combinations of elements in Σ_k, or equivalently the free abelian group on the set Σ_k, and denote it by $C_k(X)$. If we impose a total order on the vertex set V, we define set operators $d_i : \Sigma_k \to \Sigma_{k-1}$, for $0 \leq i \leq k$, by letting $d_i(\sigma) = \sigma - \{s_i\}$, where s_i denotes the i-th element in σ, under the given total ordering. We now define linear operators $\partial_k : C_k(X) \to C_{k-1}(X)$ by

$$\partial_k = \sum_{i=0}^{k} (-1)^i d_i$$

Since the groups $C_k(X)$ are equipped with the bases Σ_k, these operators can be expressed as matrices $D(k)$ whose columns are parametrized by Σ_k, whose rows are parametrized by Σ_{k-1}, and where for $\sigma \in \Sigma_k$ and $\tau \in \Sigma_{k-1}$, the entry $D(k)_{\tau\sigma}$ is $= 0$ if $\tau \not\subseteq \sigma$, and $= (-1)^i$ if $\tau \subseteq \sigma$ and if τ is obtained by removing the i-th member of the subset σ. The key observation is now that $\partial_k \circ \partial_{k+1} \equiv 0$. It follows that $Image(\partial_{k+1}) \subseteq Kernel(\partial_k)$, and that one can therefore define $H_k^{simp}(X, \mathbb{Z})$ by

$$H_k^{simp}(X, \mathbb{Z}) \cong Kernel(\partial_k)/Image(\partial_{k+1})$$

This basis independent version of the definition can be replaced by the result of matrix manipulations on the collection of matrices $\{D(k)\}_{k \geq 0}F$, as in [22]. The end results of these calculations are always the *Smith normal form* of various matrices constructed out of the $D(k)$'s. It turns out that $H_k^{simp}(X, \mathbb{Z})$ is always canonically isomorphic to the singular homology of the space $|X|$. The conclusion is that for simplicial complexes, homology is algorithmically computable.

2.2 Building coverings and complexes

Since the homology of simplicial complexes is algorithmically computable, it is frequently desirable to construct simplicial complexes which compute the homology of an underlying space X, or at least has a strong relationship with it. One way to guarantee that the simplicial complex computes the homology of X is to produce a homotopy equivalence from X to the simplicial complex, or a homotopy equivalence from a space homotopy equivalent to X to the simplicial complex. There are a number of simplicial complexes which can be constructed from X together with additional data attached to X. We begin with the Čech complex. Let X be a topological space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be any covering of X.

Definition 2.2 *The* Čech complex *of $\mathcal{U}$, denoted by $\check{C}(\mathcal{U})$, will be the abstract simplicial complex with vertex set A, and where a family $\{\alpha_0, \ldots, \alpha_k\}$ spans a k-simplex if and only if $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$.*

This is an extremely useful construction in homotopy theory. One reason is that one has the following "nerve theorem" (see [5]), which provides criteria which guarantee that $\check{C}(\mathcal{U})$ is homotopy equivalent to the underlying space X. (Recall that two spaces X and Y are said to be *homotopy equivalent* if there are maps $f : X \to Y$ and $g : Y \to X$ so that $f \circ g$ and $g \circ f$ are homotopic to Id_Y and Id_X respectively. A space which is homotopy equivalent to the one point space is called *contractible*.)

Theorem 2.3 *Suppose that X and $\mathcal{U}$ are as above, and suppose that the covering consists of open sets and is numerable (see [51] for a definition). Suppose further that for all $\emptyset \neq S \subseteq A$, we have that $\bigcap_{s \in S} U_s$ is either contractible or empty. Then $\check{C}(\mathcal{U})$ is homotopy equivalent to X.*

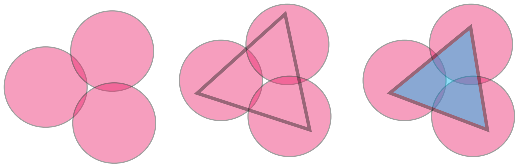
One now needs methods for generating coverings. When the space in question is a metric space, one covering is given by the family $\mathcal{B}_\epsilon(X) = \{B_\epsilon(x)\}_{x \in X}$, for some $\epsilon > 0$. More generally, for any subset $V \subseteq X$ for which $X = \bigcup_{v \in V} B_\epsilon(v)$, one can construct the nerve of the covering $\{B_\epsilon(v)\}_{v \in V}$. This is a useful theoretical construction, in view of the following theorem.

Theorem 2.4 *Let M be a compact Riemannian manifold. Then there is a positive number e so that $\check{C}(\mathcal{B}_\epsilon(M))$ is homotopy equivalent to M whenever $\epsilon \leq e$. Moreover, for every $\epsilon \leq e$, there is a finite subset $V \subseteq M$ so that the subcomplex of $\check{C}(\mathcal{B}_\epsilon(M))$ on the vertices in V is also homotopy equivalent to M.*

One problem with this construction is that it is computationally expensive, in that it requires storage of simplices of various dimensions. An idea for dealing with that problem is to construct a simplicial complex which can be recovered solely from the edge information. This suggests the following variant of the Čech construction, referred to as the *Vietoris-Rips* complex.

Definition 2.5 *Let X denote a metric space, with metric d. Then the Vietoris-Rips complex for X, attached to the parameter ϵ, denoted by $VR(X, \epsilon)$, will be the simplicial complex whose vertex set is X, and where $\{x_0, x_1, \ldots, x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i, j \leq k$.*

We note that the vertex sets of the two constructions are identical, so they can both be viewed as subcomplexes of the complete simplex on the set X. The following diagram indicates the difference between the complexes.



The leftmost figure shows the covering, the middle the Čech complex, and the rightmost the Vietoris-Rips.

The following comparison between the two complexes is easy to verify. We will see how to make use of it in the next section.

Proposition 2.6 *We have inclusions*

$$\check{C}(X,\epsilon) \subseteq VR(X,2\epsilon) \subseteq \check{C}(X,2\epsilon)$$

Even the Vietoris-Rips complex is computationally expensive, though, due to the fact that its vertex set consists of the entire metric space in question. A solution to this problem which has been used to study subspaces of Euclidean space is the *Voronoi decomposition.* Let X be any metric space, and let $\mathcal{L} \subseteq X$ be a subset, called the set of *landmark points.* Given $\lambda \in \mathcal{L}$, we define the Voronoi cell associated to λ, V_λ, by

$$V_\lambda = \{x \in X | d(x,\lambda) \leq d(x,\lambda')\}$$

for all $\lambda' \in \mathcal{L}$. It is immediate that the Voronoi cells form a covering of X, and we define the Delaunay complex attached to $\mathcal{L}$ to be the nerve of this covering. When the underlying space is Euclidean space, the Voronoi decomposition gives rise to an extremely useful decomposition of the space, and in favorable cases the Delaunay complex gives a triangulation of the convex hull of $\mathcal{L}$, referred to as the *Delaunay triangulation* [21]. For submanifolds of Euclidean space, one may construct the *restricted Delaunay triangulation* as in [25]. The value of this construction is that it produces very small simplicial complexes, whose dimension is often equal to the dimension of the manifold under consideration. Both the the Čech and Vietoris-Rips complexes typically produce simplices in dimensions much higher than the dimension of the space. The definition of the Delaunay complex makes sense for any metric space, in particular for finite metric spaces. However, for finite metric spaces, it generically produces degenerate (i.e. discrete) complexes, with no one dimensional simplices. This is due to the fact that for finite metric spaces, it is generically the case that each value of the distance is taken only for one pair of points, so one does not have any points which are equidistant between a pair of landmarks. In order to make the method useful for finite metric spaces, we therefore modify the definition of the Delaunay complex to accommodate pairs of points which are "almost" (as permitted by the introduction of a parameter ϵ) equidistant from a pair of landmark points. Precisely, we have the following definition from [8].

Definition 2.7 *Let X be any metric space, and suppose we are given a finite set $\mathcal{L}$ of points in X, called the landmark set, and a parameter $\epsilon > 0$. For every point $x \in X$, we let m_x denote the distance from this point to the set $\mathcal{L}$, i.e. the minimum distance from x to any point in the landmark set. Then we define the* strong witness complex *attached to this data to be the complex $W^s(X,\mathcal{L},\epsilon)$ whose vertex set is $\mathcal{L}$, and where a collection $\{l_0,\ldots,l_k\}$ spans a k-simplex if any only if there is a point $x \in X$ (the witness) so that $d(x,l_i) \leq m_x + \epsilon$ for all i. We can also consider the version of this complex in which the 1-simplices are identical to those of $W(X,\mathcal{L},\epsilon)$, but where the family $\{l_0,\ldots,l_k\}$ spans a k-simplex if and only if all the pairs (l_i,l_j) are 1-simplices. We'll denote this by W^s_{VR}*

There is a modified version of this construction, which is quite useful, called the weak witness construction. Suppose we are given a metric space X, and a set of points $\mathcal{L} \subseteq X$. Let $\Lambda = \{l_0,\ldots,l_k\}$ denote a finite subset of a metric space $\mathcal{L}$. We say a point $x \in X$ is a *weak witness* for Λ if $d(x,l) \geq d(x,l_i)$ for all i and all $l \notin \Lambda$. Given $\epsilon \geq 0$, we will also say that x is an ϵ *weak witness* for Λ if $d(x,l) + \epsilon \geq d(x,l_i)$ for all i and all $l \notin \Lambda$.

Definition 2.8 *Let X, $\mathcal{L}$, and ϵ be as above. We construct the* weak witness complex *for the given data, $W^w(X,\mathcal{L},\epsilon)$ by declaring that a family $\Lambda = \{l_0,\ldots,l_k\}$ spans a k-simplex if and only if Λ and all its faces* *admit ϵ weak witnesses. This complex also clearly has a version in which a k-simplex is included as a simplex if and only if all its 1-faces are, and we will denote this version by W^w_{VR}.*

It is clearly the case that if we have $0 \leq \epsilon \leq \epsilon'$, then we have an inclusion

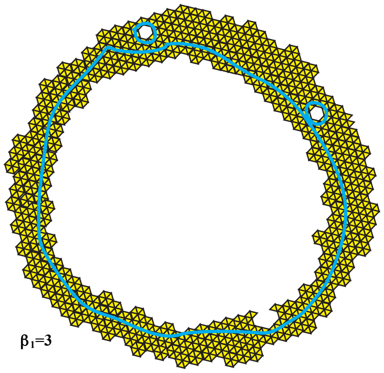
$$W^s(X, \mathcal{L}, \epsilon) \hookrightarrow W^s(X, \mathcal{L}, \epsilon')$$

and similarly for $W^s_{VR}, W^w,$ and W^w_{VR}.

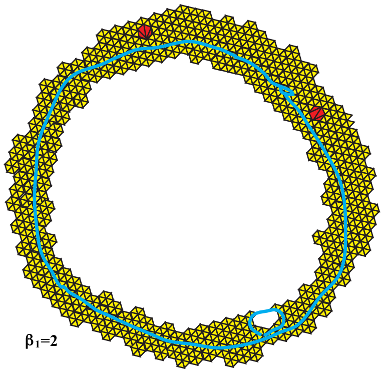
2.3 Persistent homology

Let X be a subspace of $\mathbb{R}^n$. Suppose further that we have a method of sampling points from X, perhaps with noise. By sampling with noise, we will mean that we are sampling points from a probability distribution concentrated near X. Let $\mathbb{X}$ be one sample as described above. An interesting question is to what extent it is possible to infer the Betti numbers of X from $\mathbb{X}$. In general, of course, the answer is no. For example, it will clearly be necessary to assume something about the density of the sampling. Niyogi, Smale, and Weinberger in [52] provide sufficient hypotheses which guarantee that this is possible for Riemannian manifolds. Their method is to prove that the Čech complex associated to a covering by balls of a fixed radius ϵ is homotopy equivalent to the underlying manifold. If one is interested in studying data from experiments, though, one typically cannot assume that the data lies along a submanifold. Further, even if one could assume that the data lies along a manifold, one is usually not in a position to verify that the stringent hypotheses of [52] are satisfied. The key to obtaining the desired homological information is to avoid selecting a fixed value of the threshhold ϵ, and instead obtaining a useful summary of the homological information for all the different values of ϵ at once. This philosophy is referred to as *persistence*, and was first introduced in [24].

We begin with the set $\mathbb{X}$. It is of course a finite metric space, and we may consider the Čech complexes $\check{C}(\mathbb{X}, \epsilon)$, attached to the collection of balls of radius ϵ with centers at the points of $\mathbb{X}$. Note that if the centers actually lie on a submanifold $M \subseteq \mathbb{R}^n$, and the set $\mathbb{X}$ is sufficiently dense in M, then this complex is the Čech complex attached to a covering of M. If further ϵ is sufficiently small, then all the balls will be geodesically convex, and the complex will compute the homology of M correctly. This connection provides heuristic justification for the use of this Čech complex as a method for approximating the homology of M. Now recall that whenever we have $\epsilon \leq \epsilon'$, we have an inclusion of complexes $\check{\mathrm{C}}(\mathbb{X}, \epsilon) \subseteq \check{\mathrm{C}}(\mathbb{X}, \epsilon')$. Consider the picture below, which is that of a Čech complex constructed on a finite collection of points in the Euclidean plane.



We note that the main shape of the set is concentrated around a circle. However, if we compute the homology of this complex, it will yield a first Betti number of three, namely including the large main loop and secondly the two smaller loops corresponding to the two smaller holes in the complex. Intuitively, we regard these two holes as coming from faulty sampling or other errors in the recovery of the data. One could then argue that this comes from an incorrect choice of the parameter ϵ, and that one should simply increase the parameter value to obtain a complex with the correct higher connectivity structure. This would give rise to the following picture.



Note that while the two smaller holes have now been "filled in", a new hole has been introduced in the lower right hand portion of the figure. Consequently, if we computed the homology of this complex, we would obtain a first Betti number of two. The result is incorrect for either of the parameter values. We now observe, though, that there is an inclusion of the upper complex into the lower complex, since the upper one corresponds to a smaller parameter value than the lower one. We can therefore ask about the image of the homology of the upper complex in the homology of the lower complex. The two small cycles in the upper complex vanish in the lower complex, since they are filled in. On the other hand, the small cycle in the lower complex is not in the image of the homology of the upper complex, since the required edge is not filled in in the upper complex. We see therefore that the image consists exactly of the larger cycle, which is what we regard as the "correct" answer in this case. The goal of this section is to make this observation into a systematic computational scheme which will provide the desired summary of the behavior of homology under all choices of values for the scale parameter ϵ.

We begin with the following definition. Again, refer to [44] for material on categories, functors, and natural transformations.

Definition 2.9 *Let $\underline{C}$ be any category, and $\mathcal{P}$ a partially ordered set. We regard $\mathcal{P}$ as a category $\underline{\mathcal{P}}$ in the usual way, i.e. with object set $\mathcal{P}$, and with a unique morphism from x to y whenever $x \leq y$. Then by a $\mathcal{P}$* persistence object *in $\underline{C}$ we mean a functor $\Phi : \underline{\mathcal{P}} \to \underline{C}$ More concretely, it means a family $\{c_x\}_{x\in\mathcal{P}}$ of objects of $\underline{C}$ together with morphisms $\phi_{xy} : c_x \to c_y$ whenever $x \leq y$, such that $\phi_{yz} \circ \phi_{xy} = \phi_{xz}$ whenever $x \leq y \leq z$. Note that the $\mathcal{P}$-persistence objects in $\underline{C}$ form a category in their own right, where a morphism F from Φ to Ψ is a natural transformation. Again, in more concrete terms, a morphism from a family $\{c_x, \phi_{xy}\}$ to a family $\{d_x, \psi_{xy}\}$ is a family of morphisms $\{f_x\}$, with $f_x : c_x \to d_x$, and where the diagrams*

$$\begin{tikzcd}
c_x \arrow[r, "\phi_{xy}"] \arrow[d, "f_n"'] & c_y \arrow[d, "f_y"] \\
d_x \arrow[r, "\psi_{xy}"] & d_y
\end{tikzcd}$$

all commute. We will denote the category of $\mathcal{P}$-persistence objects in $\underline{C}$ by $\mathcal{P}_{pers}(\underline{C})$. We note finally that if $f : \mathcal{P} \to \mathcal{Q}$ is a partial order preserving map, we obtain an evident functor $f^* : \mathcal{Q}_{pers}(\underline{C}) \to \mathcal{P}_{pers}(\underline{C})$ defined by $f^*(\Psi) = \Psi \circ \underline{f}$, where $\underline{f}$ is f regarded as a functor $\underline{\mathcal{P}} \to \underline{\mathcal{Q}}$.

We let $\mathbb{R}$ and $\mathbb{N}$ denote the partially ordered sets of real number and non-negative integers, respectively. We now observe that all of the constructions of the previous subsection (Čech, Vietoris-Rips, witness) yield an $\mathbb{R}$-persistence simplicial complex attached to $\mathbb{X}$. We can now construct the associated chain complexes and homology groups and obtain $\mathbb{R}$-persistence chain complexes and $\mathbb{R}$-persistence groups. What makes homology useful as a discriminator between topological spaces is the fact that there is a classification theorem for finitely generated abelian groups. If one had a classification theorem for $\mathbb{R}$-persistence abelian groups, then it could act as a summary of the behavior of the homology of all the complexes $\check{\mathrm{C}}(X, \epsilon)$. However, we do not have such a theorem. However, it turns out that there is a classification theorem (see [64]) for a subcategory of the category of $\mathbb{N}$-persistence F-vector spaces, where F is a field.

To understand this classification, we observe that $\mathbb{N}$-persistence abelian groups can be identified with a more familiar notion, namely that of a graded module over a graded ring. Let $\{A_n\}$ be any $\mathbb{N}$-persistence abelian group. We will define an associated graded module $\theta(\{A_n\})$ over the graded polynomial ring $\mathbb{Z}[t]$, where t is assigned degree 1, as follows. We set

$$\theta(\{A_n\}) = \bigoplus_{s \geq 0} A_s$$

where the n-th graded part is the group A_n. The action of the polynomial generator t is given by

$$t \cdot \{\alpha_n\} = \{\beta_n\}, \text{ where } \beta_n = \psi_{n-1,n}(\alpha_{n-1})$$

It is readily checked that θ is a functor from $\mathbb{N}_{pers}(\underline{Ab})$ to the category of graded $\mathbb{Z}[t]$-modules, and is in fact an equivalence of categories, since an inverse functor can be given by $M_* \to \{M_n\}$, where the morphisms ψ_{mn} are given by multiplication by t^{n-m}. The conclusion is that the category $\mathbb{N}_{pers}(\underline{Ab})$ is equivalent to the category of non-negatively graded modules over $\mathbb{Z}[t]$. Now, there is still no classification theorem for graded $\mathbb{Z}[t]$-modules. However, if we let F denote any field, then there is a classification theorem for *finitely generated* graded $F[t]$-modules. See [23] for the non-graded case. The graded case is proved in identical fashion.

Theorem 2.10 *Let M_* denote any finitely generated non-negatively graded $F[t]$-module. Then there is an isomorphism*

$$M_* \cong \bigoplus_{s=1}^{m} F[t](i_s) \oplus \bigoplus_{t=1}^{n} (F[t]/(t^{l_t}))(j_t)$$

where for any graded $F[t]$-module N_*, the notation $N_*(s)$ denotes N_* with an upward dimension shift of s. So, $N_*(s)_l = N_{l-s}$. The decomposition is unique up to permutation of factors.

It is therefore a useful question to ask which $\mathbb{N}$-persistence F-vector spaces correspond under θ to finitely generated non-negatively generated $F[t]$-modules. We have the following.

Proposition 2.11 *We say an $\mathbb{N}$-persistence F-vector space $\{V\}_n$ is* tame *if every vector space V_n is finite dimensional, and if $\psi_{n,n+1} : V_n \to V_{n+1}$ is an isomorphism for sufficiently large n. Then we have that $\theta(\{V_n\}_n)$ is a finitely generated $F[t]$-module if and only if $\{V_n\}_n$ is tame.*

We now have an easy translation of the classification result 2.10. For any $0 \leq m \leq n$, we define an $\mathbb{N}$-persistence F-vector space $U(m,n)$ by setting $U(m,n)_t = 0$ for $t < m$ and $t > n$, $U(m,n) = F$ for $m \leq t \leq n$, and $\psi_{s,t} = Id_F$ for $m \leq s \leq t \leq n$. We extend this definition to the value $n = +\infty$ in the evident way.

Proposition 2.12 *Any tame $\mathbb{N}$-persistence F-vector space $\{V_n\}_n$ can be decomposed as*

$$\{V_n\}_n \cong \bigoplus_{i=0}^{N} U(m_i, n_i)$$

where each m_i is a non-negative integer, and n_i is a non-negative integer or $+\infty$. The decomposition is unique in the sense that the collection of pairs $\{(m_i, n_i)\}_i$ is unique up to ordering of factors.

We can reformulate this result as follows. By a *bar code*, we mean a finite set of pairs (m, n), where m is a non-negative integer, and n is a non-negative integer or $+\infty$. We can now restate Proposition 2.11 as the assertion that just as finite dimensional vector spaces are classified up to isomorphism by their dimension, so tame $\mathbb{N}$-persistence vector spaces are classified by associated bar codes.

Returning to the $\mathbb{R}$-persistence simplicial complexes we construct, we may use any partial order preserving map $\mathbb{N} \to \mathbb{R}$ to obtain an $\mathbb{N}$-persistence simplicial complex. There are at least two useful ways to construct such maps. The first would be to choose a small number ε, and define a map $f_\varepsilon : \mathbb{N} \to \mathbb{R}$ by $f_\varepsilon(n) = n\varepsilon$. A second method would be as follows. Given a finite point cloud as above, it is clear that there are only finitely many real values at which there are transitions in the complex. This follows from the nature of the conditions together with the fact that the distance function takes only finitely many values on $\mathbb{X}$. Letting these transition values be enumerated in increasing order as $\{t_0, t_1, \ldots, t_N\}$, we define an order preserving map $g : \mathbb{N} \to \mathbb{R}$ by $g(n) = t_n$ for $n \leq N$, and $g(n) = t_N$ for $n \geq N$. The first construction can be interpreted as sampling values of the persistence parameter from a uniform lattice. Of course, the sampling is finer as ε decreases. The second method more efficient, since it is precisely adapted to the complex at hand. Furthermore, it contains complete information about the original $\mathbb{R}$-vector space.

The methodology we now use to study the homology of the complexes constructed above is now as follows.

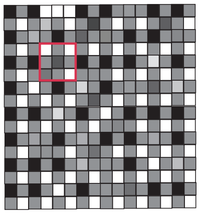
- Construct the $\mathbb{R}$-persistence simplicial complex $\{C_\epsilon\}$ using Čech, Vietoris-Rips, or witness methods. We will denote it by Φ.
- Select a partial order preserving map $f : \mathbb{N} \to \mathbb{R}$.
- Construct the associated $\mathbb{N}$-persistence simplicial complex.
- Construct the associated $\mathbb{N}$-persistence chain complex $\{C_*(n)\}_n$ with coefficients in F. (It is evident from the finiteness hypotheses on $\mathbb{X}$ and the nature of the constructions that the associated $\mathbb{N}$-persistence F-vector spaces are tame.)
- Compute the barcodes associated to the $\mathbb{N}$-persistence vector spaces $\{H_i(C_*(n))\}_n$.

The last step turns out be tractable due to the translation into commutative algebraic terms above. We recall (see [22]) that homology computation can be performed by putting a matrix in Smith normal form. The algorithms for Smith normal form are typically applied for matrices over $\mathbb{Z}$, but they are applicable in the context of any principal ideal domain, of which $F[t]$ is one. The fact that the ring is graded and the boundary matrices are homogeneous makes the application simpler. This is the observation made in [64], where the algorithm for computing persistent homology which we use is described in detail.

In interpreting the output, one now finds that long intervals in the output barcode indicates the presence of a homology class which "persists" over a long range of parameter values, while short intervals indicate cycles which are "born" at a given parameter value and then "die" at a nearby parameter value. The pictures and discussion above allow us to formulate the intuition that long intervals correspond to large scale geometric features in the space, and short intervals correspond to noise or inadequate sampling. Of course, what is short and what is long is very problem dependent. Also, in some cases, it may be a false dichotomy, and the more useful point of view is that the barcodes represents the space at various scales, and the whole multiscale version of the space may actually be of interest.

2.4 Example: Natural image statistics

Images taken with a digital camera can be viewed as vectors in a very high dimensional vector space. Each image consists of a number (the gray scale value) attached to each of a large number of pixels, and therefore we may think of the image as lying in $\mathbb{R}^P$, where P is the number of pixels used by the camera. From this point of view, one can ask questions about the nature of the collection of all possible images lying within $\mathbb{R}^P$. For example, can it be modeled as a submanifold or subspace of $\mathbb{R}^P$? If it were, one could conclude on the one hand that it is very high dimensional, since images are capable of expressing such a wide variety of scenes, and on the other hand that it would be a manifold of very high codimension, since random pixel arrays will very rarely approximate an image. David Mumford gave a great deal of thought to questions like this one concerning natural image statistics, and came to the conclusion that although the above argument indicates that whole manifold of images is not accessible in a useful way, a space of small image patches might in fact contain quite useful information. In [41], A. Lee, D. Mumford, and K. Pedersen performed an analysis constructed in this way, and we will summarize the results of that paper. They began with a database of black and white images taken by J. van Hateren and A. van der Schaaf in [34]. The database consisted of images taken around Groningen, Holland, in town and in the surrounding countryside. Within such an image, one can consider 3×3 patches, i.e. square arrays of 9 pixels.



Each such patch can now be regarded as a 9-tuple of real numbers (the gray scale values again), i.e. a vector in $\mathbb{R}^9$. A preliminary observation is that patches which are constant, or rather nearly constant, will predominate among these patches. The reason is that most images have large solid regions, where the gray scale intensity does not change significantly, and these regions will contribute more to the collection of patches than the patches in which some transitions are occurring. These nearly constant patches will be referred to as *low contrast*. Of course, the low contrast patches do not carry interesting structure, so Lee, Mumford, and Pedersen proceeded as follows. They first define the *D-norm* of a 3×3 image patch, as a certain quadratic function of the logs of the gray scale values. It is a way of defining the contrast of an image patch. Then they select 5,000 patches at random from each of the images from [34], and select the top 20% as evaluated by the D-norm. This will now constitute a database of high contrast patches from the patches occurring in the image database from [34]. They then perform two transformations on the data, as follows.

1. Mean center the data. The mean intensity value over all nine pixels is subtracted from each pixel value, to obtain a patch with mean zero. This means that if a patch is obtained from another patch by adding a constant value, i.e. "turning up the brightness knob", then the two patches will be regarded as the same. Note that this transformation puts all points in an 8-dimensional subspace within $\mathbb{R}^9$.

2. Normalize the D-norm. Since all the patches chosen will have D-norm bounded away from zero, one can divide by it to obtain a patch with D-norm $= 1$. This means that if one patch is obtained from another by "turning the contrast knob", then the two patches will be regarded as identical.

The result of this construction is a database $\mathcal{M}$ of c:a 4.5×10^6 points on a 7-dimensional ellipsoid in $\mathbb{R}^8$. The goal of the paper [41] is now to obtain some understanding of how this set sits within S^7, and what can be said about the patches which do appear. A first observation is that the points are scattered throughout the 7-sphere, in the sense that no point on S^7 is very far from the set, but that the density appears to vary a great deal. In particular, in [41] indications were found that the data was concentrated around an annulus. In [11], a systematic study of the topology of the high density portion was carried out, and this work is what we will describe in the remainder of this section.

The first issue to be addressed is what is meant by "high density" portion. Density estimation is a highly developed area within statistics (see for example [58]). We selected a very crude proxy for density, in the interest of minimizing the computational burden. It is defined as follows. Fix a positive integer k, and define the *k-codensity* function δ_k of $x \in \mathbb{X}$, where $\mathbb{X}$ is a set of point cloud data, by

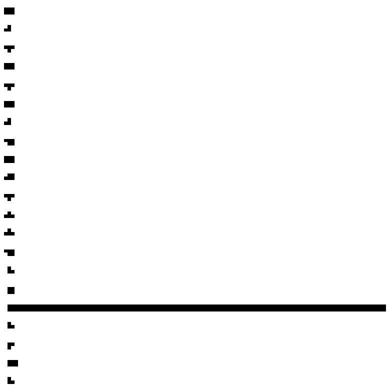
$$\delta_k(x) = d(x, \nu_k(x))$$

where d denotes the distance function in $\mathbb{X}$, and where $\nu_k(x)$ denotes the k-th nearest neighbor of x in $\mathbb{X}$. Note that $\delta_k(-)$ is inversely related with density, since a concentrated region will have smaller distances to the k-th nearest neighbor, so we will be studying subcollections of points for which $\delta_k(-)$ is bounded from above by a threshhold. Secondly, we also note that each δ_k yields a different density estimator. In rough terms, δ_k for large values of k computes density using points in large neighborhoods of x, and for small values uses small neighborhoods. So, δ_k for large k corresponds to a smoothed out notion of density, and for small k corresponds to a version which carries more of the detailed structure of the data set.

For any subset $\mathcal{M}_0 \subseteq \mathcal{M}$, we now define subsets $\mathcal{M}_0[k,T] \subseteq \mathcal{M}_0$, where k is a positive integer, and T is a percentage value, by

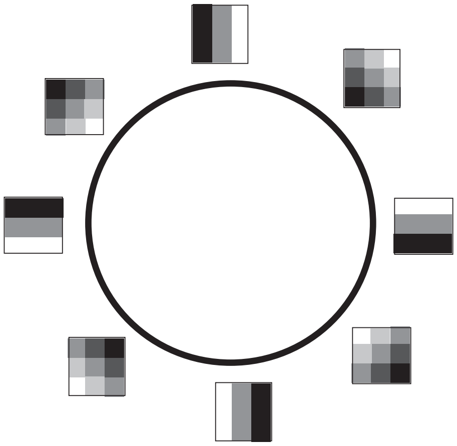
$$\mathcal{M}_0[k,T] = \{x \in \mathcal{M}_0 | \delta_k(x) \text{ lies among the } T\% \text{ lowest values of } \delta_k \text{ in } \mathcal{M}_0\}$$

The goal of the paper [11] is to infer the topology of a putative space underlying the various sets of points $\mathcal{M}[k,T]$. In some cases we will approximate this via a subset $\mathcal{M}_0[l,T]$. It is fairly direct to see that the variable k scales with the size of the set, so that if $\rho = \#(\mathcal{M})/\#(\mathcal{M}_0)$, then $\mathcal{M}_0[k,T]$ is comparable with $\mathcal{M}[\rho k,T]$. We do this by obtaining the data sets $\mathcal{M}_0[k,T]$, then selecting a set of landmark points, and finding various barcodes attached to witness complexes associated to the space. Below is the barcode for $H_1(W\mathcal{M}_0[300,30]))$, where $\mathcal{M}_0$ is a sample of 5×10^4 points from $\mathcal{M}$, and W denotes a witness complex constructed with 50 landmark points.



$$k = 300, T = 30\%$$

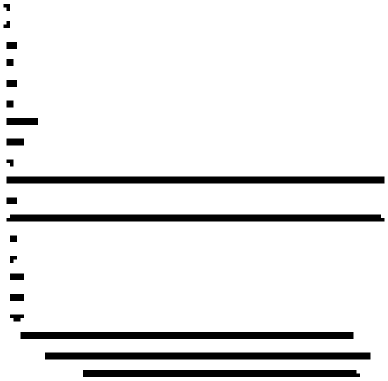
We note that there are a number of short lines, and one long one. According to the philosophy mentioned above, this suggests that the first Betti number should be estimated to be one. The barcode is stable, in the sense that it appears repeatedly when the set of landmark points is varied, and when the sample from the full data set is varied. Therefore, the simplest possible explanation for this barcode is that the underlying space should be a circle. One can then ask if there is a simple explanation involving the data which would yield a circle as the underlying space. The picture below gives such an explanation.



Primary circle

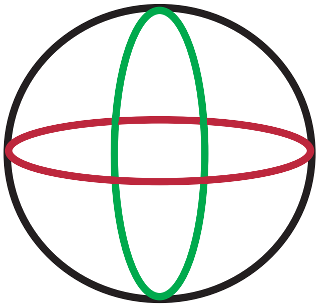
More formally, the picture suggests the explanation that the most frequently occurring patches are those approximating two variable functions which depend only on a linear projection from the two variable space, and so that the function is increasing in that linear projection. This explanation is consistent with an annulus conjectured to represent the densest patches in [41].

The next picture is a barcode for $H_1(W\mathcal{M}_0[15, 30]))$, where $\mathcal{M}_0$ is as above, again with 50 landmark points chosen.



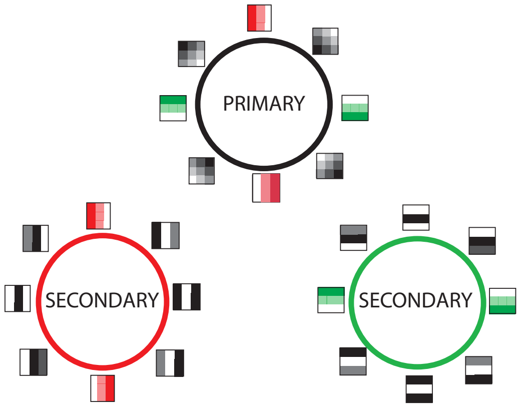
$$k = 15, T = 30\%$$

We note that in this case, there are many short segments and 5 longer segments. It is again the case that this barcode is stable over the choices of landmark points and samples from $\mathcal{M}$. This suggests that the first Betti number of the putative underlying space should be five. It now becomes more difficult to identify the simplest explanation for this result, and a number of such models are possible. The one which ultimately turns out to fit the data is pictured as follows.



Three circle model

This picture is composed of a primary circle, pictured in black, and two secondary circles, pictured in green and red. Although it is perhaps not clear from the image, the intent is that the secondary circles each intersect the primary circle in two points, and do not intersect each other. A space constructed this way can readily be seen to have a first Betti number of five. An explanation for this geometric object in terms of image patches is now the following.

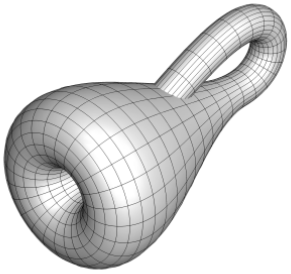


Three circle model in the data

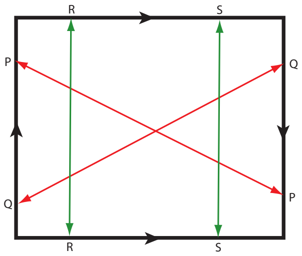
The secondary circles interpolate between functions which are an increasing function of a linear projection to functions which are "bump functions" with an internal local maximum evaluated on the same linear projection. Note that the two secondary circles each intersect the primary circle in two points, as indicated by the coloring, and that they do not intersect each other. We informally confirmed that the indicated patches are the ones which occur in the high density portions of the data set.

Remark: We note that there is a preference within the data for patches which are aligned in vertical and horizontal directions, since one can construct versions of the secondary circles which are not aligned in the vertical direction, and they do not appear. One explanation for this is that patches in natural images are biased in favor of the horizontal and vertical directions because nature has this bias, since for example objects aligned in a vertical direction are more stable than those aligned at a 45 degree angle. Another explanation is that this phenomenon is related to the technology of the camera, since the rectangular pixel arrays in the camera have the potential to bias the patches in favor of the vertical and horizontal directions. We believe that both factors are involved. In [13], we have studied 5×5 patches, and found the three circles model appearing there as well. In that case, one would expect to see less bias in favor of the vertical and horizontal directions, since the pixels give a finer sampling of the image.

One could now ask if there is a larger 2-dimensional space containing the three circle model, occurring with substantial density. We will first ask to find a natural embedding of the theoretical three circle model in a 2-manifold. It turns out that the model embeds naturally in a Klein bottle (Image courtesy of Tom Banchoff).

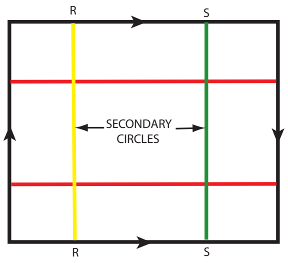


To see this, we first recall that the Klein bottle can be described as an identification space as pictured below.



The colored arrows indicate points being identified using the quotient topology construction (see [33]), which informally means that the top vertical edge is identified with the lower vertical edge in a way which preserves the x-coordinate, and the right hand vertical edge is identified with the left hand vertical edge after a twist which changes the orientation of the edge. It is convenient to represent the Klein bottle in this way, since it does not embed in Euclidean 3-space, and therefore cannot be precisely visualized, although a useful visual representation including self intersections was shown above.

We are interested in finding a sensible embedding of the three circle model in the Klein bottle. Some experimentation with the three circle model results in the following picture



in which the red segments form the primary circle and the yellow and green segments form the secondary circles. The red segments form a single circle since the intersection point of the lower red segment with the right (respectively left) vertical edge is identified with the intersection point of the upper red segment with the left (respectively right) vertical edge, and the yellow and green segments form circles since the intersection of the yellow (respectively green) segment with the upper horizontal edge is identified with the intersection of the yellow (respectively green) segment with the lower horizontal edge. We note that the yellow and green circles intersect the primary circle in two points, and do not intersect each other, so the picture in question produces a natural embedding of the three circle model in the Klein bottle.

Remark: The selection of the Klein bottle was the result of a great deal of mental experimentation with various candidate 2-manifolds, in which we were unable to find similar natural embeddings in other candidate manifolds, such as the torus or projective plane.

In [11], it was demonstrated that the Klein bottle effectively models a space of high contrast patches of high density. To understand the results of that paper, it is necessary to discuss another theoretical version of the Klein bottle. We will be regarding the 3 × 3 patches as obtained by sampling a smooth real valued function on the unit disc at nine grid points, and study subspaces of the space of all such functions which have a rough correspondence with the subspaces of the space of patches we study. We will consider the space $\mathcal{Q}$ of all two variable polynomials of degree 2, i.e. functions

$$f(x,y) = A + Bx + Cy + Dx^2 + Exy + Fy^2$$

The set $\mathcal{Q}$ is a 6 dimensional real vector space. We now consider the subspace $\mathcal{P} \subseteq \mathcal{Q}$ consisting of functions f so that

$$\int_D f = 0 \qquad \text{and} \qquad \int_D f^2 = 1 \tag{2–1}$$

The first condition is the analogue of the mean centering condition imposed on the patches, and the second is the analogue of the normalization condition for the contrast. Imposing only the first condition, which is linear, produces a 5 dimensional vector subspace, and the second, which is quadratic in character, produces a 4 dimensional ellipsoid within this vector space. We now consider the subspace $\mathcal{P}_0 \subseteq \mathcal{P}$, consisting of all functions within $\mathcal{P}$ which are of the form

$$f(x,y) = q(\lambda x + \mu y)$$

where q is a single variable quadratic function, and where $\lambda^2 + \mu^2 = 1$. The space of all functions of this form within $\mathcal{Q}$ is four dimensional (three variables parametrize q, and (λ, μ) lies on the (one-dimensional) unit circle). The two additional constraints in 2–1 above imposed on it now yield the 2-dimensional complex $\mathcal{P}_0$. We claim that $\mathcal{P}_0$ is homeomorphic the Klein bottle. To see this, we let A denote the space of single variable polynomials $q(t) = c_0 + c_1 t + c_2 t^2$ satisfying the two conditions

$$\int_{-1}^{1} q(t) = 0 \qquad \text{and} \qquad \int_{-1}^{1} q(t)^2 = 1$$

It is easy to check that regarded as a subspace of $\mathbb{R}^3$, this subspace is an ellipse and therefore is homeomorphic to a circle. For any unit vector $\vec{v}$ in $\mathbb{R}^2$ and any $q \in A$, we let $q_{\vec{v}} : \mathbb{R}^2 \to \mathbb{R}$ be defined by $q_{\vec{v}}(\vec{w}) = q(\vec{v} \cdot \vec{w})$. It is easy to check that for $q \in A$ and $\vec{v}$ a unit vector, we have that

$$\int_D q_{\vec{v}} = 0 \qquad \text{and} \qquad \int_D q_{\vec{v}}^2 \neq 0$$

and that therefore the formula

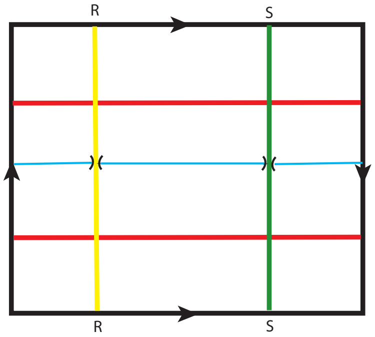
$$(q, \vec{v}) \to \frac{q_{\vec{v}}}{||q_{\vec{v}}||_2}$$

defines a continuous map θ from $A \times S^1$ to $\mathcal{P}_0$. The map θ is however not a homeomorphism, which one can check as follows. Let $\rho : A \to A$ be the involution defined by $\rho(c_0 + c_1 t + c_2 t^2) = c_0 - c_1 t + c_2 t^2$. Then we have the relation

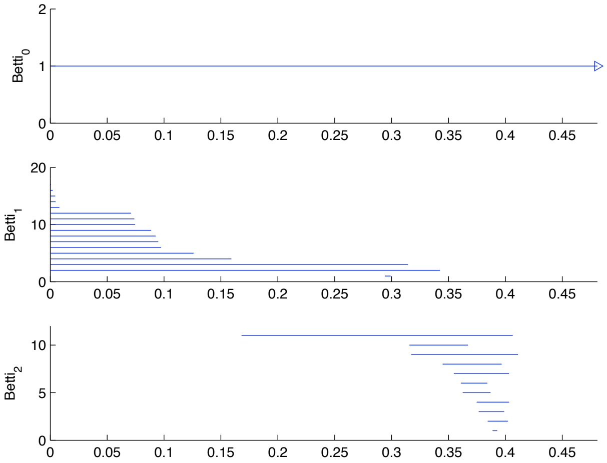
$$\theta(q, \vec{v}) = \theta(\rho(q), -\vec{v})$$

so that the map θ factors through the space of orbits under the involution. It is easy to check (a) that the factorization is a homeomorphism and (b) that the orbit space is homeomorphic to a Klein bottle.

We now ask to what extent we can "see" a Klein bottle in the data. A naive approach to this question would be to simply perform the experiments we did above for $k = 15$ but with a less stringent density threshhold, for example 40 or 50. Performing these experiments do not produce a non-trivial β_2. One might suppose, though, that the set $\mathcal{M}_0$ is not large enough to provide sufficient resolution in the density estimation, and that if one uses the full set $\mathcal{M}$ that one might obtain different answers. With this larger set, one might also vary the estimator parameter k to obtain a finer estimation of density. After some experimentation, one can construct a data set $\mathcal{S}$ by sampling 10^4 points from the data set $\mathcal{M}[100, 10]$. The set $\mathcal{S}$ exhibits the three circle model clearly, but enlarging it still does not exhibit the non-trivial β_2 which would be characteristic of the Klein bottle. In order to begin to understand the situation, one should then ask what the least frequently appearing patches on the Klein bottle would be. In thinking about the polynomial model, one expects that there should be a preference for the linear polynomials, and the experience with the three circle model suggests that there is an additional preference for the patches which are lined up with the vertical directions. This suggests that the least frequently appearing patches would be the pure quadratics composed with the linear functions $(x, y) \to \frac{\sqrt{2}}{2}x + \frac{\sqrt{2}}{2}y$ and $(x, y) \to \frac{\sqrt{2}}{2}x - \frac{\sqrt{2}}{2}y$. Slightly more frequently appearing would be pure quadratics composed with any linear function which is not a projection on the x and y-axes. This set of pure quadratics forms a pair of open one-dimensional arcs within the Klein bottle. These arcs are indicated in blue in the image below, where as before the red circle arcs form the primary circle, and the yellow and green arcs the secondary circles.



The idea will be that we wish to include some points corresponding to the blue arcs to the data set, but which turn out not to satisfy the density threshhold. We will find that once these points are added, the set then carries the topology of a Klein bottle. This would then give a reasonable qualitative description for a portion of the density distribution of the high contrast 3×3 image patches, in that it can be said to concentrate around a Klein bottle, but with strongly reduced density around the non-vertical and non-horizontal pure quadratic patches. In order to find such a map, we identify the pixels in the 3×3 array with the points in the set $L = \{-1, 0, 1\} \times \{-1, 0, 1\}$ in the Euclidean plane. For any given function $f \in \mathcal{P}_0$, we define the associated patch by evaluating f at the nine points of L, and then mean centering and D-norm normalizing. This produces a map h from the Klein bottle $\mathcal{P}_0$ to the normalized patch space. We then obtain additional points to add to the data set by selecting 30 points at random from the blue arcs in the Klein bottle, computing h on them, and then for each of the 30 points selecting the points of $\mathcal{M}[100, 10]$ which lie closest to them, and finally adjoining them to the set $\mathcal{S}$ to obtain an enlarged set $\mathcal{S}'$. Witness complexes with 50 landmark points computed for $\mathcal{S}'$ now display the barcodes which would be associated to a Klein bottle. Here is a typical picture.



We note that the β_0 barcode (the upper picture) clearly shows the single component, the β_1 shows two lines from threshhold parameter value .15 to .35, and finally the β_2 barcode shows a single line on roughly the same interval. This gives $\beta_0 = 1$, $\beta_i = 2$, and $\beta_2 = 1$. These are the mod 2 Betti numbers for the Klein bottle.

Remark: There are actually two two-dimensional manifolds with these mod 2 Betti numbers, one is the Klein bottle, and the other is a torus. These two are distinguished by mod 3 homology, and we have performed the computation to show that the mod 3 homology is consistent with the Klein bottle and not with the torus.

Remark: One can also ask if the space we are studying is in fact closely related the theoretical Klein bottle defined above using quadratic polynomials. That it is so is strongly suggested in [11] by a comparison with data sets constructed by adjoining additional points obtained from the whole theoretical Klein bottle to the set $\mathcal{S}$. The resulting space also shows a strong indication of the same Betti numbers, indicating that the spaces represent essentially the same phenomenon.

2.5 Example: Electrode array data from primary visual cortex

The goal of neuroscience is, of course, to obtain as complete as possible an understanding of how the nervous system operates in performing all its tasks, including vision, motor control, higher level cognition, olfactory sensation, etc.. One aspect of this kind of understanding is the analysis of the structure and function of individual neurons, and the creation of an associated taxonomy of individual neurons. Another aspect is the analysis of how families of neurons cooperate to accomplish various tasks, which could be referred to as the study of populations of neurons. The second problem appears to be very amenable to geometric analysis, since it will involve the activities of several neurons at once. In the paper [59], a first attempt at topological analysis of data sets constructed out of the simultaneous activity of several neurons is carried out, with encouraging results, and we will describe the results of that paper in this section.

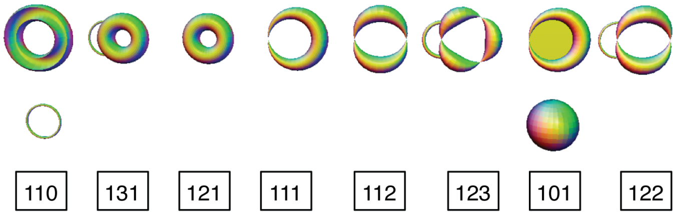
The arrays of neurons studied in [59] are from the *primary visual cortex* or *V1* in Macaque monkeys. The primary cortex is a component in the visual pathway, which begins with the retinal cells in the eye, proceeds through the lateral geniculate locus, then to the primary visual cortex, and then through a number of higher level processing units, such as V2, V3, V4, middle temporal area (MT), and others. See [36] and [63] for useful discussions. It is known that V1 performs low level tasks, such as edge and line detection, and its output is then processed for higher level and larger scale properties further along the visual pathway. However, the mechanism by which it carries out these tasks is not understood. A very interesting series of experiments were conducted in the papers [62] and [38]. These authors study the behavior of the V1 in Macaque monkeys by injecting a voltage sensitive dye in it, and then performing optical imaging of small regions of the cortex. Voltage changes in this portion of the cortex will then give rise to color differences in the imaging. Since the voltages change over time, so will the optical images. These papers study the behavior of the optical images under two separate conditions, one the *evoked* state, in which stimuli are being supplied to the eye of the monkey, and the *unevoked* or *spontaneous* state, in which no stimulus is being supplied. It was observed in [38] that in an informal sense, the images in the different conditions appeared to be quite similar, and a statistical analysis strongly suggested that the behavior of V1 in the spontaneous condition was consistent with a behavior which consisted in moving through a family of evoked images corresponding to responses to angular boundaries, without any particular order.

Another method for studying the behavior of neurons in V1 and other parts of the nervous system is the method of embedded electrode arrays. In this case, arrays of up to c:a 100 regularly spaced electrodes are implanted in the V1 (or whatever other portion of the nervous system one is studying) . The voltage at the electrodes are then recorded simultaneously, so one obtains a voltage level at each of the electrodes at each point in time. Sophisticated signal processing techniques are then used to obtain an array of N (where N is the number of electrodes) *spike trains*, i.e. lists of firing times for N neurons. This experimental setup provides another view into the behavior of the neurons in V1, and the idea of the paper [59] was to attempt to the replicate the results of [38], which were carried out using the voltage sensitive dye technology, in the embedded electrode setting and to attempt to refine the results presented there. We now describe the experiments which were carried out, and the results of our analysis of the data obtained from them.

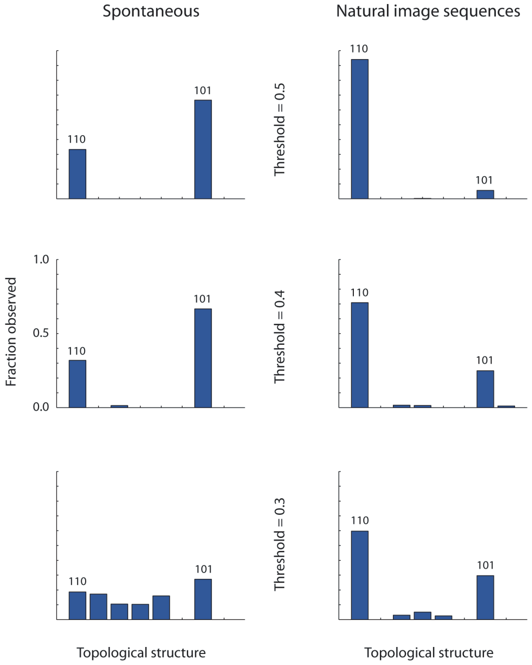
10×10 electrode arrays were used in recording output from the V1 in Macaque monkeys who view a screen.

Two segments of recording, each of roughly 20-30 minutes, were done under two separate conditions. In the first, the eyes of the animal were occluded, so no stimulus was presented to the visual system. In the second, a video sequence obtained by sampling from different movie clips. We refer to the data obtained in the first setting as *spontaneous*, and in the second setting as *evoked*. A signal processing methodology called *spike sorting* was then applied to the data, so that one could identify neurons and firing times for each neuron. Next, the data was broken up into ten second segments in both cases. Each such segment was next divided up into 200 50ms bins, and for each neuron one is able to count the number of firings within each such 50ms bin. The five neurons with the highest firing rate were selected in each 10 second window, and for each bin one can now obtain the 5-vector of number of firings of each of these five neurons. By performing this construction over all 200 bins, one obtains a set of point cloud data consisting of 200 points in $\mathbb{R}^5$. Of course, we have many such data sets, coming from different choice of 10 second segments and from different choice of "regime" (spontaneous or evoked).

Beginning with these point clouds, witness complexes based on 35 landmark points were constructed. The landmark sets were constructed by the "maxmin" procedure, a procedure designed to ensure that the landmark points are well distributed throughout the point cloud. This procedure begins with a seed point, and then constructs the rest of the points deterministically from it. For each data set, we constructed witness complexes from all the possible seed points. In order to derive topological signatures from each such witness complex, one selects a threshhold as a fraction of the covering radius of the point cloud, and then determines the Betti numbers β_0, β_1, and β_2 of the witness complex with this given threshhold value of ϵ. Thus, for each witness complex, one can now obtain a vector or signature of integers $(\beta_0, \beta_1, \beta_2)$. The observed signatures are listed below, with pictures of simple models of possible geometries which they represent.



By far the most frequently occurring signatures were $(1, 1, 0)$ and $(1, 0, 1)$, corresponding to a circle and a sphere, respectively. The picture below shows the distribution of occurrences of these two under various choices of the threshholds.



In order to validate the significance of the results, we ran the identical procedures with data generated at random for firings with a Poisson model. Monte Carlo simulation show that the probability of obtaining segments for β_1 or β_2 longer than 30% of the diameter of the point cloud is $< .005$.

The summary of the results of the experiment are that topological methods clearly distinguish the data in both regimes from a Poisson null hypothesis model. They also suggest that there is a similarity in the spontaneous and evoked regimes, since the same topological signatures occur. Further, though, the statistics of the signatures occurring are also able to distinguish between the two regimes. We do not yet understand the nature of the topological phenomenon, which is something which should be addressed by mapping algorithms, perhaps along the lines of the following section. One aspect we have addressed in this direction, though, is the question of whether a simple periodic phenomenon associated the body's natural rhythms are responsible for the topology. Such a phenomenon would likely create peaks in the amplitude spectrum of the segments of the data we study. No statistically significant peaks of this type were observed.

3 Imaging: Mapper

3.1 Visualization

So far we have discussed the attachment of homological signatures to point clouds in an attempt to obtain geometric understanding of them. Frequently, though, it is possible to find images of various kinds attached to point cloud data which allow one to obtain a qualitative understanding of them through direct visualizaton. One such method is the *projection pursuit* method (see [37]), which uses a statistical measure of information contained in a linear projection to select a particularly good linear projection for data which is embedded in Euclidean space. Another method is *multidimensional scaling*, (see [1]), which begins from an arbitrary point cloud and attempts to embed it isometrically in Euclidean spaces of various dimensions with minimum distortion of the metric. Related developments are the Isomap (see [61]) and locally linear embedding (see [56]). In all cases, the methodologies result in a point cloud in $\mathbb{R}^2$ or $\mathbb{R}^3$, which can then be visualized by the investigator. There are, however, other possible avenues for visualization and qualitative representation of geometric objects. One such possibility is representation as a graph or as a higher dimensional simplicial complex. Such combinatorial representations can lead to useful qualitative understanding in their own right, but graph visualization software such as *Graphviz* (available at http://www.graphviz.org/) can provide useful visualizations. In thinking about how to develop such a representation, it is useful to keep in mind what characteristics would be desirable. Here is a list of some such properties.

- **Insensitivity to metric:** As mentioned in the introduction, metrics used in analyzing many modern data sets are not derived from a particularly refined theory, but instead are constructed as a reasonable quantitative proxy for an intuitive notion of similarity. Therefore, imaging methods should be relatively insensitive to detailed quantitative changes.
- **Understanding sensitivity to parameter changes:** Many algorithms require parameters to be set before an outcome is obtained. Since setting such parameters often involves arbitrary, it is desirable to use methods which provide useful summaries of the behavior under all choices of parameters if possible.
- **Multiscale representations:** It is desirable to understand sets of point cloud at various levels of resolution, and to be able to provide outputs at different levels for comparison. Features which are seen at multiple scales will be viewed as more likely to be actual features as opposed to more transient features which could be viewed as artifacts of the imaging method.

The rest of this section will be devoted to the description of a method which addresses each of these points. We have named it *Mapper*, and it is described in detail in [60].

3.2 A topological method

We begin with a topological construction based on a covering of a topological space X.

Definition 3.1 *Let X be a topological space, and let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ be a finite covering of the space X (so the set A is finite). Let $\Delta[A]$ denote the standard simplex with vertex set A, so $dim(\Delta[A]) = \#(A) - 1$. Further, for any non-empty subset $S \subseteq A$, we let $\Delta[S] \subseteq \Delta[A]$ denote the face spanned by the vertices corresponding to elements of S, and we let $X[S] = \bigcap_{s \in S} U_s \subseteq X$. By the* Mayer-Vietoris blowup *of X associated to $\mathcal{U}$, denoted by $\mathcal{M}(X,\mathcal{U})$, we mean the subspace*

$$\bigcup_{\emptyset \neq S \subseteq A} \Delta[S] \times X[S] \subseteq \Delta[A] \times X$$

We note that there are natural projection maps $f : \mathcal{M}(X,\mathcal{U}) \to X$ and $g : \mathcal{M}(X,\mathcal{U}) \to \Delta[A]$, which have the following properties.

- The map f is a homotopy equivalence when X has the homotopy type of a finite complex and the covering consists of open sets. In fact, using a partition of unity subordinate to the covering $\mathcal{U}$, one can obtain an explicit homotopy inverse $\varphi : X \to \mathcal{M}(X,\mathcal{U})$.

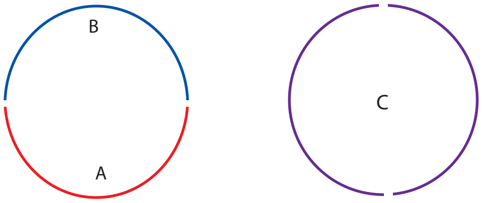
- The map g is a homotopy equivalence onto its image (which is the nerve of the covering, or the Čech complex $\check{C}(\mathcal{U})$) when all the sets $X[S]$ are either empty or contractible. This is how the nerve theorem (Theorem 2.3) is proved.

These two observations demonstrate that one obtains a map from X to $\check{C}(\mathcal{U})$ for any finite complex X. Such a map can be viewed as a kind of coordinatization of the space X. Ordinary coordinatizations provide maps to Euclidean spaces of various dimensions, and they often provide useful insights into the spaces in question. Simplicial complexes, particularly low dimensional ones, can also often be readily visualized, and can therefore also be expected to provide useful information about a space. This is so even if the map is not a homeomorphism, so it does not provide complete information about a space. We will next observe that there is a variant of the $\check{C}(\mathcal{U})$ construction which is a somewhat more sensitive target for this kind of coordinatization map. Let X be any topological space, and let $\mathcal{U}$ be any covering of X. We will now define a simplicial complex $\check{C}^{\pi_0}(\mathcal{U})$ to be the nerve of the covering of X by sets which are path connected components of a set of the form U_α, so the covering is indexed by the set of pairs $\{(\alpha, \xi)\}$, where ξ is a path component of U_α. The set map $(\alpha, \xi) \to \alpha$ yields a map of simplicial complexes $\check{C}^{\pi_0}(\mathcal{U}) \to \check{C}(\mathcal{U})$. It is further clear that we have a map $\mathcal{M}(X,\mathcal{U}) \to \check{C}^{\pi_0}(\mathcal{U})$, which is induced by the projection $U_\alpha \to \pi_0(U_\alpha)$ for each α. Finally, the composite

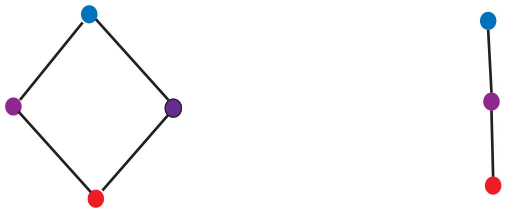
$$\mathcal{M}(X,\mathcal{U}) \to \check{C}^{\pi_0}(\mathcal{U}) \to \check{C}(\mathcal{U})$$

is the earlier defined map g.

Example: Let X denote the unit circle, and let a covering $\mathcal{U}$ of X be given by the three sets $A = \{(x, y)|y < 0\}$, $B = \{(x, y)|y > 0\}$, and $C = \{(x, y)|y \neq \pm 1\}$. We note that $\pi_0(A)$ and $\pi_0(B)$ consist of a single point, and $\pi_0(C)$ consists of two points.



The simplicial complexes $\check{C}^{\pi_0}(\mathcal{U})$ and $\check{C}(\mathcal{U})$ are now given by the picture



Note that $\check{C}^{\pi_0}(\mathcal{U})$ is actually homeomorphic to X, while $\check{C}(\mathcal{U})$ is not. This is an example of the fact that $\check{C}^{\pi_0}$ is more sensitive that $\check{C}$.

In order for this construction to be useful, one must develop methods for constructing coverings of topological spaces. Earlier we have looked at constructions where the sets are open balls in a metric space, and where we have versions of the Voronoi decomposition adapted to general metric spaces. We now suppose that we are given a reference map ρ from our space X to a metric space Z, and further that we are given a finite open or closed covering $\mathcal{U}$ of Z. Then we may consider the covering $\rho^*\mathcal{U}$ given by $\rho^*\mathcal{U} = \{\rho^{-1}U_\alpha\}_{\alpha \in A}$, when $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$. This is clearly a covering of X. Typical examples of useful metric spaces Z include $\mathbb{R}$, $\mathbb{R}^n$, and S^1. These spaces admit natural coverings.

Example: For $X = \mathbb{R}$, let R and e be positive real numbers. Then we may construct the covering $\mathcal{U}[R, e]$ which consists of all intervals of the form $[kR - e, (k + 1)R + e]$. This is a two parameter family of coverings, and as long as $e < \frac{R}{2}$, it has covering dimension 1, in the sense that no non-trivial threefold overlaps are non-empty. Products of these intervals would give a corresponding covering of $\mathbb{R}^n$.

Example: Let $X = S^1$,N be an integer ≥ 2, and $\epsilon > 0$ be a real number. Then we can form a covering $\mathcal{U}[N, \epsilon] = \{U_j\}_{0 \leq j < N}$ of X by setting

$$U_j = \{(cos(x), sin(x))|x \in [\frac{2\pi j}{N} - \epsilon, \frac{2\pi j}{N} + \epsilon]\}$$

whenever $\epsilon > \frac{\pi}{N}$.

Remark: When the reference space is $\mathbb{R}$, our construction is closely related to the *Reeb graph* of a real valued function on a manifold (see [55]). The actual Reeb graph should be viewed as a limiting version of the construction as one studies the coverings $\mathcal{U}[R, \epsilon]$ with R and ϵ tending to zero.

We must now describe a method for transporting this construction from the setting of topological spaces to the setting of point clouds. The notion of a covering makes sense in the point cloud setting, as does the definition of coverings of point clouds using maps from the point cloud to a reference metric space, by "pulling back" a predefined covering of the reference space. The notion which does not make immediate sense is the notion of π_0, i.e. constructing connected components of a point cloud. The notion of *clustering* (see [31]) turns out to be the appropriate analogue. Our main example of such a clustering algorithm will be the so-called single linkage clustering. It is defined by fixing the value of a parameter ϵ, and defining blocks of a partition of our point cloud as the set of equivalence classes under the equivalence relation generated by the relation $\sim_\epsilon$ defined by $x \sim_\epsilon x'$ if and only if $d(x, x') \leq \epsilon$. Note that the set of clusters in this setting is precisely π_0 applied to the Vietoris-Rips complex $VR(X, \epsilon)$, and that each "cluster" corresponds to the set of vertices in a single connected component. Now, our version of the construction $\check{C}^{\pi_0}$ in this context is obtained as follows.

1. Define a reference map $f : X \to Z$, where X is the given point cloud and Z is the reference metric space.
2. Select a covering $\mathcal{U}$ of Z. If $Z = \mathbb{R}$, then $\mathcal{U}$ can be obtained by selecting R and e as above, and constructing the covering $\mathcal{U}[R, e]$.
3. If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, then construct the subsets $X_\alpha = f^{-1}U_\alpha$.
4. Select a value ϵ as input to the single linkage clustering algorithm above, and construct the set of clusters obtained by applying the single linkage algorithm with parameter value ϵ to the sets X_α. At this point, we have a covering of X parametrized by pairs (α, c), where $\alpha \in A$ and c is one of the clusters of X_α.
5. Construct the simplicial complex whose vertex set is the set of all possible such pairs (α, c), and where a family $\{(\alpha_0, c_0), (\alpha_1, c_1), \ldots, (\alpha_k, c_k)\}$ spans a k-simplex if and only if the corresponding clusters have a point in common.

This construction is a plausible analogue of the continuous construction described above. We note that it depends on the reference map, a covering of the reference space, and a value for ϵ. We observe that in fact any clustering algorithm could be used to cluster the sets X_α, and one could still obtain a sensible construction. We note that if the covering $\mathcal{U}$ has covering dimension $\leq d$, i.e. if whenever we are given a family $\{\alpha_0, \alpha_1, \ldots, \alpha_t\}$ of distinct elements of α with $t > d$, then $U_{\alpha_0} \cap \ldots \cap U_{\alpha_t} = \emptyset$, then the dimension of the simplicial complex we construct will be $\leq d$ as well. This follows immediately from the definitions.

We note that this construction readily produces multiresolution or multiscale structure which allows one to distinguish actual features from artifacts. To see this, we begin with the definition of a map of coverings. Let $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ be coverings of a space Z. By a map of coverings from $\mathcal{U}$ to $\mathcal{V}$ we will mean a set map $\theta : A \to B$ so that for all $\alpha \in A$, we have $U_\alpha \subseteq V_{\theta(\alpha)}$.

Example: Consider the coverings $\mathcal{U}[R, e]$ of $\mathbb{R}$ defined above. The indexing set in this case consists of the integers. It is clear from the definition that the identity map from $\mathbb{Z}$ to itself yields a map of coverings $\mathcal{U}[R, e] \to \mathcal{U}[R, e']$ whenever $e \leq e'$. In this case, the map of coverings consists simply of the inclusion of an interval into an interval with the same center but with larger diameter.

Example: The map of integers $k \to \lfloor \frac{k}{2} \rfloor$ defines a map of coverings $\mathcal{U}[R, e] \to \mathcal{U}[2R, e]$, which is two to one in the sense that every interval in $\mathcal{U}[2R, e]$ contains two intervals from $\mathcal{U}$. In order to use these maps to obtain a multiresolution version of the Mapper construction, we need a definition.

Definition 3.2 *A clustering algorithm is said to be* functorial *if whenever one has an inclusion $X \to Y$ of point clouds, i.e. a set map preserving distances, then the image of each cluster constructed in X under f is included in one of the clusters in Y. It follows from the fact that the clustering algorithm produces a partition of the point cloud in question that each cluster is contained in a unique cluster, and therefore that we have an induced map of sets from the clusters in X to the clusters in Y.*

Now suppose we are given data for applying Mapper, i.e. a point cloud X together with a reference map ρ to a metric space Z. Suppose further that we are given two coverings $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ of Z, and a map of coverings $\theta : A \to B$. Since we have a map of coverings, it is clear that we obtain inclusions $\rho^{-1}U_\alpha \subseteq \rho^{-1}V_{\theta(\alpha)}$ for all $\alpha \in A$. If we apply a functorial clustering scheme to each of the sets $\rho^{-1}U_\alpha$ and $\rho^{-1}V_\beta$, it is clear from the definition that we will obtain a map from the set of clusters obtained by applying the clustering algorithm to $\rho^{-1}U_\alpha$ to the set of clusters obtained by applying it to $\rho^{-1}V_\beta$, and therefore a map from the vertex set of $\mathcal{M}(X,\mathcal{U})$ to the vertex set of $\mathcal{M}(X, \mathcal{V})$. One readily checks that it is actually a simplicial map, so we obtain an associated simplicial map $\Theta : \mathcal{M}(X,\mathcal{U}) \to \mathcal{M}(X, \mathcal{V})$. So now, for example, we will always obtain a diagram of simplicial complexes

$$\cdots \to \mathcal{M}(X,\mathcal{U}[R/4, e]) \to \mathcal{M}(X,\mathcal{U}[R/2, e]) \to \mathcal{M}(X,\mathcal{U}[R, e])$$

As one moves to the left, the coverings of $\mathbb{R}$ (and therefore of X) become more refined, and are presumed to give picture with finer resolution of the space in question. Studying the behavior of features under such maps will allow one to get a sense of which observed features are real geometric features of the point cloud, and which are artifacts, since the intuition is that features which appear at several levels in such a multiresolution diagram would be more intrinsic to the data set than those which appear at a single level.

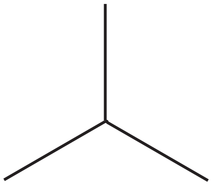
3.3 Filters

An important question, of course, is how to generate useful reference maps ρ. If our reference space Z is actually $\mathbb{R}^n$, then this means simply generating real valued functions on the point cloud. To emphasize the way in which these functions are being used, we refer to them as *filters*. Frequently one has interesting filters, defined by a user, which one wants to study. However, in other cases one simply wants to obtain a geometric picture of the point cloud, and it is important to generate filters directly from the metric which reflect interesting properties of the point cloud. Here are some important examples.

- **(Density):** Consider any density estimator applied a point cloud X. It will produce a non-negative function on X, which reflects useful information about the data set. Often, it is exactly the nature of this function which is of interest.
- **(Data depth):** The notion of *data depth* refers to any attempt to quantify the notion of nearness to the center of a data set. It does not necessarily require the existence of an actual center in any particular sense, although a point which minimizes the quantity in question could perhaps be thought of as a choice for a center. In our group's work, we have referred to quantities of the form

  $$e_p(x) = \frac{1}{\#(\mathbb{X})} \sum_{x' \in \mathbb{X}} d(x, x')^p$$

  (with an obvious generalization to $p = \infty$) as *eccentricity functions*, and have used them as filters. Other notions could equally well be used. The main point is that Mapper output based on such functions can identify qualitative structure of a particular kind. For example, if the space were as pictured below,

  then Mapper would recover the structure of the three flares coming out from the central point.
- **(Eigenvectors of graph Laplacians):** Graph Laplacians are interesting linear operators attached to graphs (see [40]). In particular, their eigenfunctions produce functions on the vertex set of the graph. They can be used, for example, to produce cluster decompositions of data sets when the graph is the 1-skeleton of a Vietoris-Rips complex. We find that these eigenfunctions (again applied to the 1-skeleton of the Vietoris-Rips complex of a point cloud) also can produce useful filters in Mapper analysis of data sets.

3.4 Scale space

The construction from the previous subsection depends on certain inputs, including a parameter ϵ. The decision of how to choose this parameter is in principle a difficult one, for which one has little guidance. Further, it may often be desirable to broaden the definition of the complex to permit choices of ϵ which vary with α, i.e. over the reference space Z. In this section, we discuss a systematic way of considering such varying choices of "scale". We first note that the because the single linkage procedure applied to a point cloud X can be interpreted as computing connected components of $VR(X, \epsilon)$, the persistence barcode for β_0 yields interesting information about the behavior of the components (or clusters) for all values of ϵ. To be explicit about this, we consider the subset $E(X) \subset \mathbb{R}_+ = [0, +\infty)$ consisting of all the endpoints of the intervals occurring in the barcode. $E(X)$ is a finite set on the non-negative real line, and there is consequently a total ordering on it induced from the total ordering on $\mathbb{R}_+$, and we write $E(X) = \{e_1, \ldots, e_t\}$, with $e_i < e_j$ whenever $i < j$. From the definition, it is clear that whenever we have $e_i < \eta < \eta' < e_{i+1}$, the natural map on H_0 induced by the inclusion $VR(X, \eta) \hookrightarrow VR(X, \eta')$ is an isomorphism, and that therefore the inclusion induces a bijection on connected components. For this reason, we call each of the intervals (e_i, e_{i+1}) a *stability interval*, or an *S-interval* for X. We now have the following definition.

Definition 3.3 *Given a point cloud X, a reference map $\rho : X \to Z$ to a metric space, and a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, we define a simplicial complex $SS = SS(X, \rho, \mathcal{U})$ as follows. The vertices of SS are pairs (α, I), where $\alpha \in A$, and where I is a stability interval for the point cloud $X_\alpha = \rho^{-1}(U_\alpha)$. A $(k+1)$-tuple $\{(\alpha_0, I_0), (\alpha_1, I_1), \ldots, (\alpha_k, I_k)\}$ spans a k-simplex in SS if (a) $U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$ and (b) $I_0 \cap \ldots \cap I_k \neq \emptyset$. The vertex map $(\alpha, I) \to \alpha$ induces a map of simplicial complexes $p : SS \to \check{C}(\mathcal{U})$. By a* scale choice *for X and $\mathcal{U}$, we will mean a section of the map p, i.e. a simplicial map $s : \check{C}(\mathcal{U}) \to SS$ such that $p \circ s = Id_{\check{C}(\mathcal{U})}$.*

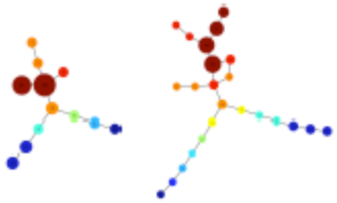
Given any scale choice s for X and $\mathcal{U}$, we set $s(\alpha) = (\alpha, I_\alpha)$. Now, for any scale choice s and $\alpha \in A$, we choose $\epsilon_\alpha \in I_\alpha$. This gives a choice of the scale parameter ϵ varying with α, and we can build a new complex whose vertex set consists of pairs (α, c), where c is a cluster in the single linkage clustering applied to $X_\alpha = \rho^{-1}U_\alpha$ with perimeter value ϵ_α. From the definition of the stability intervals, it is clear that the complex is independent of the choice of $\epsilon_\alpha \in I_\alpha$.

Remark: The intuition behind the definition of scale choice is the following. We wish to permit a choice of scale parameter ϵ which varies with α. Of course, the set of all such choices is too large to contemplate using any kind of exhaustive enumeration of the possible values, and will in any case not be useful since we will not have any criteria to determine which choices are more plausible than others. The definition given above incorporates two different heuristics which permit us to restrict the choices of ϵ_α which we make, as well as to evaluate various choices relative to each other.

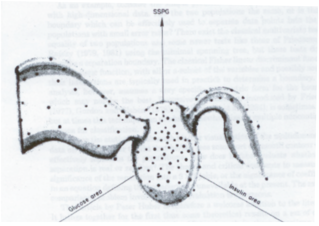
- From the fact that the scale choice s is a simplicial map, it follows that whenever $U_\alpha \cap U_{\alpha'}$, we also have $I_\alpha \cap I_{\alpha'}$. This means that the choices of parameters ϵ_α have a certain kind of continuity in the variable α, which is surely a desirable feature of a varying choice of scales.
- The fact that the stability intervals have a notion of length allows us to evaluate scale choices. The general rule of thumb is that choices of scale which are stable over a large range of parameter values are to be preferred over those with stability over a shorter range. This permits various notions of numerical weights (such as, for example, $\sum_\alpha l(I_\alpha)$, where $l(-)$ denotes length) which allow one to compare scale choices.

3.5 Examples

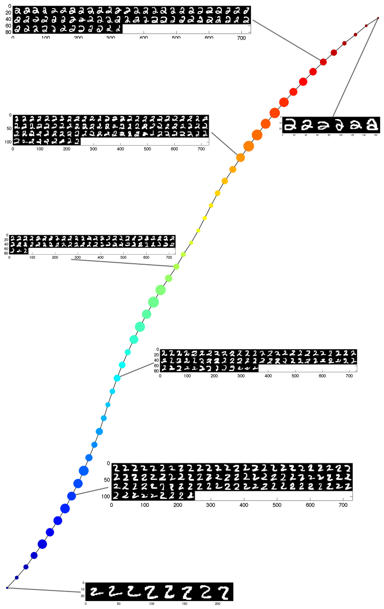
We show the outputs from Mapper applied to various data sets. The first example comes from a six dimensional data set constructed by G.M. Reaven and R.G. Miller from a diabetes study conducted at Stanford University in the 1970's. 145 patients were included. Details of the study and of the construction of the data set can be found in [48]. Below is the output of Mapper applied to this data set, with two different levels of resolution.



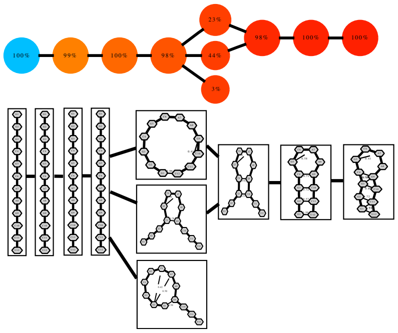
The filter is in this case a density estimator, and high values are indicated in red, and low density values in blue. At both scales, there is a central dense core, and two “flares” consisting of points with low density. The core consists of normal or near-normal patients, and the two flares consist of patients with the two different forms of diabetes. An imaging coming from the projection pursuit method is given below.



A second example comes from the paper [61], and consists of scanned images of handdrawn copies of the digit “two”.



The images are compared using a simple L^2-metric, and Mapper is applied using a density filter. One can observe that the dominant feature which is changing as one moves along the graph is the increasing presence of a loop in the lower left hand corner of the digit. This result is consistent with what was obtained by using ISOMAP in [61].



The picture above is constructed using a data set constructed with the folding@home project (http://folding.stanford.edu/), by simulating the folding of so-called "RNA hairpins". These are relatively small so-called motifs occurring within larger RNA molecules, and are among the most frequently occurring such motifs. The actual data set is obtained from a probabilistic simulation of the dynamics, based on the notion of a *contact map*. The contact map is simply an array of zeroes and ones, whose slots correspond to the residues along the molecule, and where an entry is one if the corresponding pair of residues are in contact with one another, by which we mean that they are within a fixed threshhold of each other. One can impose a Hamming style metric on the set of these contact maps. At this point, given a family of such contact maps generated by simulations, we can employ a filter which is a good proxy for density, and apply Mapper. The output from this application is the colored graph displayed above. One notes that in the middle, one has some slightly complicated behavior among the orange nodes, in particular a loop in the corresponding graph. The contact maps corresponding to members of the clusters corresponding to these nodes are displayed below the Mapper output. The contact maps are displayed by inserting edges between residues which are in contact. We note that given only the Mapper output, one might suspect that the small feature (the array of orange nodes) could simply be an artifact, but examination of the data shows that they correspond to essentially distinct contact maps. Note also that the data is obtained by simply examining the states occurring in the simulation, and that it does not include any dynamic information which would show how the states are traversed in the folding process. This example points out an advantage of the method, in that it is capable of locating small features within a larger data set. The results described above appear in [6].

4 Generalized forms of persistence

4.1 Multidimensional persistence

We have studied families of spaces parametrized by a single parameter ϵ as a way of extracting connectivity information from a point cloud or finite metric space. It turns out that it is often useful to be able to analyze the behavior of increasing families of spaces parametrized by more than one variable.

Example: Given a point cloud X, one often attempts to understand the nature of an underlying probability distribution which may have given rise to it. This was clearly the case in the example of the image patch data described in Section 2.4 above. One way to do this is to estimate the density function using one of many possible density estimators (see [58]). Given such an estimator, one can now construct the family of spaces $X[T]$, where T is a percentage parameter, and $X[T] \subseteq X$ is the subset of points which lie within the T-th percentile of density as measured by the given density estimator. Clearly, if $T \leq T'$, we have an inclusion

$X[T] \subseteq X[T']$, and one is often interested in understanding the geometric evolution of this set as T increases. In the image patch example above, we chose a few possible values of T to locate levels for T in which we saw interesting topological behavior, but if one were to be able to study all the values of T simultaneously, and obtain a summary of the behavior across all values of T, then one would not have to search at random through the different values of T. In this case, we note that when we apply, for example, the Vietoris-Rips construction to these sets, one obtains a two parameter family of simplicial complexes $\{VR(X[T], \epsilon)\}_{\epsilon, T}$. The parameter ϵ is used to introduce geometry into the discrete sets $X[T]$, and the parameter T is the function value we wish to track.

Example: Persistent methods can be used to study qualitative properties of shapes which are not directly topological. For example, if one has a manifold, one can study the filtration on the manifold by the value of scalar curvature, and the evolution of the topology of the sublevel sets of this filtration can reflect interesting properties of the shape, and can provide the basis for methods for discriminating between shapes. In addition, one can build associated spaces to manifolds or other complexes by studying various versions of the tangent bundle or the tangent cone of geometric measure theory, which can also be equipped with interesting filtrations which provide interesting information about the shape, and which can also be used in locating features such as singularities in the space. See [10], [17], [28], and [7] for details of these lines of research. In order to study this kind of persistence for spaces given as point clouds, it is necessary to find discrete versions of the geometric quantities (such as curvature) which are relevant, but it is also necessary to use multiple persistence based on the geometric quantity and the scale parameter ϵ simultaneously. As in the density example above, one needs the ϵ parameter in order to impose some geometry on the discrete point cloud one is given.

Example: Suppose one is interested in studying the qualitative behavior of a real valued function f on $\mathbb{R}^n$, in terms of local maxima, minima, saddle points, etc. An efficient way of doing this is to study the evolution of the topology of the sublevel sets $S_R = \{x \in \mathbb{R}^n | f(x) \leq R\}$, as is done in Morse theory (see [49]). If one does not have the explicit form of f, but only the values of f on some grid or other sample $\mathcal{S}$ of points in $\mathbb{R}^n$, one can approximate the topology of the sublevel sets by the Vietoris-Rips complexes of $\mathcal{S} \cap S_R$, and study their evolution as R increases. Of course, to extract the topology, one also needs the allow the scale parameter ϵ in the Vietoris-Rips complexes to vary, and one obtains a two parameter family of simplicial complexes $\{VR(\mathcal{S} \cap S_R, \epsilon)\}_{\epsilon, R}$.

It is clear from these examples that it very desirable to obtain useful and computable summaries of the evolution of topology in situations where there is more than one persistence parameter. A theory which describes how this can be done is developed in [12]. We now describe this theory.

We recall from Definition 2.9 the notion of a $\mathcal{P}$-persistence object in a category $\underline{C}$, where $\mathcal{P}$ is a partially ordered set, as a functor $\underline{\mathcal{P}} \to \underline{C}$, where $\underline{\mathcal{P}}$ regards $\mathcal{P}$ as a category in the usual way. The morphisms of $\mathcal{P}$-persistence objects are natural transformations of functors. Suppose we are given a family of topological spaces (or simplicial complexes) $\{X_{s,t}\}$, with inclusions $X_{s,t} \to X_{s',t'}$ whenever $s \leq s'$ and $t \leq t'$. By choosing any order preserving map $\mathbb{N} \times \mathbb{N} \to \mathbb{R} \times \mathbb{R}$, we obtain a $\mathbb{N} \times \mathbb{N}$-persistence object in the category of topological spaces (simplicial complexes). Of course, this can be carried out for more than two variables in the obvious way. One can now apply any of the homology functors $H_i(-)$ to obtain an $\mathbb{N} \times \mathbb{N}$-vector space. We recall from Section 2.3 that the category of $\mathbb{N}$-persistence vector spaces is equivalent to the category of non-negatively graded $k[t]$-modules. There is a corresponding statement concerning $\mathbb{N}^s$-persistence vector spaces.

Definition 4.1 *By an n-graded ring, we will mean a ring A together with a direct sum decomposition of abelian groups*

$$A \cong \bigoplus_{t_1, t_2, \ldots, t_k} A_{t_1, t_2, \ldots, t_n}$$

where each of the t_i's varies over $\mathbb{N}$, and where the multiplication in the ring A satisfies the requirement

$$A_{s_1,s_2,\ldots,s_n} \cdot A_{t_1,t_2,\ldots,t_n} \subseteq A_{s_1+t_1,\ldots,s_n+t_n}$$

Similarly, an n-graded module over an n-graded ring A is an A-module M equipped with a direct sum decomposition

$$M \cong \bigoplus_{t_1,t_2,\ldots,t_k} M_{t_i,t_2,\ldots,t_k}$$

so that the requirement

$$A_{s_1,s_2,\ldots,s_n} \cdot M_{t_1,t_2,\ldots,t_n} \subseteq M_{s_1+t_1,\ldots,s_n+t_n}$$

is satisfied. Notions of homomorphism and isomorphism of n-graded rings and modules are defined in the obvious ways, making the collection of n-graded A-modules into a category.

The following proposition from [12] is now a straightforward observation.

Proposition 4.2 *The category of $\mathbb{N}^n$-persistence vector spaces over a field k is equivalent to the category of n-multigraded modules over the polynomial ring $A(n) = k[x_1, x_2, \ldots, x_n]$, where the multigrading structure on $A(n)$ is given by $A(n)_{t_1,t_2,\ldots,t_n} = k \cdot x_1^{t_1} x_2^{t_2} \cdot \cdots \cdot x_n^{t_n}$.*

As is well known to algebraists, the classification of finitely generated modules over multivariable polynomial rings is much more complicated than the corresponding result for single variable polynomial rings, and in fact no reasonable parametrization is known. This situation is also valid in the graded and multigraded cases. The classification of finitely generated graded modules in the single variable case is parametrized by a set which is independent of the field in question, while examples show that in the case of more than one variable, the classification of multigraded modules definitely depends on the field in question. In fact, the classification in the multivariable case is parametrized by points in moduli varieties over the ground field, and we therefore say that the one variable classification is *discrete* while the classification in the multivariable case is *continuous*. This observation is initially disappointing, since it suggests that useful classification results are not likely to be available. However, it turns out that there are useful invariants, even though they are not complete.

Definition 4.3 *Let M be any finitely generated n-graded $A(n)$-module. Then for any vector*

$$\vec{t} = (t_1, t_2, \ldots, t_n) \in \mathbb{N}^n$$

we define $d(\vec{t})$ to be the dimension of the vector space $M_{\vec{t}}$. Similarly, for any pair of vectors $\vec{t}, \vec{t'} \in \mathbb{N}^n$, with $\vec{t} \leq \vec{t'}$ in the sense that $t_i \leq t'_i$ for all i, we define $r(\vec{t}, \vec{t'})$ to be the rank of the multiplication map

$$x_1^{t'_1-t_1} x_2^{t'_2-t_2} \cdots x_n^{t'_n-t_n} \cdot : M_{\vec{t}} \to M_{\vec{t'}}$$

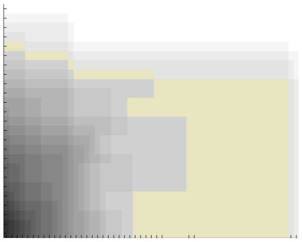
The assignments $\vec{t} \to d(\vec{t})$ and $(\vec{t}, \vec{t'}) \to r(\vec{t}, \vec{t'})$ can be regarded as $\mathbb{N}$-valued functions on the sets $\mathbb{Z}^n$ and $\mathcal{P}_n = \{(\vec{t}, \vec{t'}) \in \mathbb{Z}^n \times \mathbb{Z}^n | \vec{t} \leq \vec{t'}\}$ respectively. These functions are clearly invariants of the isomorphism class of the module M, and we refer to them as the dimension and rank invariants, respectively.

The following result is proved in [12].

Proposition 4.4 *In the case $n = 1$, the rank invariant is a complete invariant of the isomorphism class of a finitely generated graded $F[t]$-module for F a field.*

This suggests that the computation of the rank invariant could be regarded as a suitable generalization of single variable persistence, even though it clearly ignores potentially interesting information. One question one could then ask is if there is an interesting generalization of barcodes, which could be used to understand the rank invariant in the same way as barcodes do for the single variable case. This can be done as follows.

Suppose we are given an n-graded $A(n)$ module M, and we have computed the dimension and rank invariants d_M and r_M. We say two elements $\vec{s}$ and $\vec{t}$ of $\mathcal{P}_n$ are *ρ-related*, and write $\vec{s} \sim_\rho \vec{t}$, if (a) $d_M(\vec{s}) = d_M(\vec{t})$ and (b) $r_M(\vec{s}, \vec{t}) = d_M(\vec{s}) = d_M(\vec{t})$. We then let $\simeq_\rho$ denote the equivalence relation generated by $\sim_\rho$. This equivalence now gives a partition of the set $\mathbb{N}^n$. When we imagine the module as arising from a multidimensional persistence vector space, we can imagine the various nodes in the persistence diagram (i.e. the elements of $\mathbb{N}^n$) as embedded in $\mathbb{R}^n$, and we can assume that they are embedded quite densely, so that adjacent points are very close in the actual Euclidean space. One can now color code the regions in the partition associated to $\simeq_\rho$, and if the dimension is ≤ 3, obtain a kind of image describing the regions of the partition. A typical output might look like the image below.



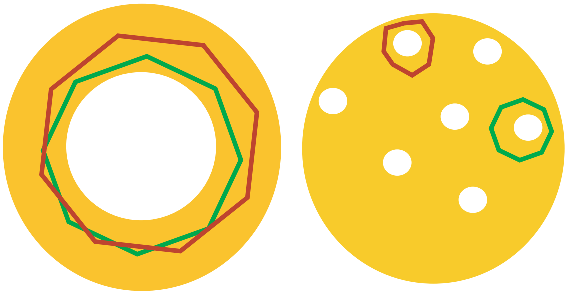
Regions of constant coloring then correspond to regions in which the vector space is constant, i.e. all the members of a single color region have the same dimension, and further can be connected by a sequence of isomorphisms associated to comparable members of $\mathbb{N}^n$. These regions correspond to intervals of constancy in the barcode, i.e. intervals which contain no endpoints of any intervals occurring in the barcode.

A difficulty which must be addressed in working with multidimensional persistence is the computational efficiency, and indeed setting up a viable computational framework. In the case of single variable persistence, we use algorithms developed for computing the Smith normal form of a matrix over a principal ideal domain. This machinery is not available in the multivariable case, but it turns out that it can be replaced by the Gröbner basis methodology (see [18] or [47]). The part of that methodology which is relevant is the multigraded version of the notion of a Gröbner basis for a submodule of a free finintely generated module, the Buchberger algorithm for constructing such a basis, and the algorithm for constructing syzygies attached to homomorphisms of free multigraded modules (Schreyer's algorithm). These results are developed in [14]. The Gröbner basis provides a very compact description which contains all the information about the multidimensional persistence problem. In particular, it permits the reconstruction of the rank invariant, which naively would have to be stored as sets of values for very large sets of inputs.

4.2 Quivers and zigzags

In Section 2.3, we developed the notion of a $\mathcal{P}$-persistence object in a category $\underline{C}$, where $\mathcal{P}$ is a partially ordered set. We then developed the theory of such persistence objects in the case $\mathcal{P} = \mathbb{N}$ to define and analyze persistent homology. In the previous section, we extended this development to the case $\mathcal{P} = \mathbb{N}^n$, and saw that it permitted the study of additional kinds of problems not addressed in the case $n = 1$. In this section, we wish to develop the notion of $\mathcal{P}$-persistence for another class of partially ordered sets, and show that it allows us to address some interesting classes of problems. The results of this section will appear in joint work with V. de Silva. We begin by listing the types of problems we wish to study.

Example: Suppose that we are given a large data set X, and we wish to study its homological invariants by studying the corresponding invariants of subsamples from X. So, for example, if one wanted to estimate the first Betti number of a putative space $\mathbb{X}$ underlying X, one might build a Vietoris-Rips complex with a fixed ϵ for a collection of many samples, and if sufficiently many of them compute the first Betti number to be n, then one might guess that the first Betti number for $\mathbb{X}$ is n. However, the picture below suggests what might go wrong with such an approach. It is a schematic picture of two different data sets, colored in yellow.



Note that in the leftmost data set, the dominant qualitative picture is that of a single loop, and one can expect that with reasonable frequency samples many produce point clouds which capture the circular structure through a barcode computation, in the way illustrated by the green and red loops. In the rightmost data set, though, one sees many different smaller circles, and one can imagine that each of the different samples might compute a first Betti number of one, but where each one corresponds to a different loop, as is again indicated by the green and red loops. One can attempt to distinguish these by insisting that there be some measurable notion of compatibility between the computations. Here is one such notion. Suppose we have a family of samples $S_i \subseteq X$, for $i = 0, 1, \ldots, N$. For each i, with $0 \leq i \leq N-1$, we consider also the sample $T_i = S_i \cup S_{i+1}$, and note that we have inclusions $S_i \hookrightarrow T_i$ and $S_{i+1} \hookrightarrow T_i$. This means that we actually have a diagram of samples from X

$$\begin{tikzcd}
 & T_0 & & T_1 & & \cdots & & T_{N-1} & \\
S_0 \arrow[ur, "i_0"] & & S_1 \arrow[ul, "j_0"'] \arrow[ur, "i_1"] & & \cdots \arrow[ul, "j_1"'] & & S_{N-1} \arrow[ul, "j_{N-2}"'] \arrow[ur, "i_{N-1}"] & & S_N \arrow[ul, "j_{N-1}"']
\end{tikzcd}$$

The value of a diagram of this form is that if we are given elements $x_0 \in H_k(VR(S_0))$ and $x_1 \in H_k(VR(S_1))$, we can obtain information relating the two classes by considering their images $i_0(x_0)$ and $j_0(x_1)$ in the group $H_k(VR(T_0))$. If we find that $i_0(x_0) = j_0(x_1)$, then this acts as confirmation that the two elements correspond to an element arising from the full data set X. Of course, there is no certainty arising from this, but it suggests the likelihood that this occurs. Note that in the example above, we will find this kind of compatibility in the left hand set, and not in the right. We would like to develop a systematic methodology which assesses the frequency of these kind of compatibilities.

Remark: The idea of recovering information about a large data set by studying behavior of various statistics on subsamples is an example of the *bootstrap method* due to B. Efron [26]. In that context, one studies means, variances, and other quantities evaluated on samples, and then assesses how these statistics vary over the samples. This is regarded as more informative than simply evaluating those statistics directly on the full data set. We can view the kind of analysis we are proposing above as a version of the bootstrap idea which is adapted to structural information, such as presence of loops or cluster decompositions, rather than to numerical values.

Example: Suppose that we are given a data set X equipped with a map τ to the real line $\mathbb{R}$. For example, if we have data containing a time component, then a choice of τ would be the time coordinate. For any $t_0 \leq t_1$, we let $X[t_0, t_1]$ denote the set $\{x \in X | t_0 \leq \rho(x) \leq t_1\}$, and let $Z(i) = X[i, i+1] \cup X[i+1, i+2]$. Then we have a diagram of point cloud data sets

$$\begin{tikzcd}
\cdots & & Z(i) & & Z(i+1) & & \cdots \\
& X[i, i+1] \arrow[ul] \arrow[ur] & & X[i+1, i+2] \arrow[ul] \arrow[ur] & & X[i+2, i+3] \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

We note that the "shape" of this diagram is identical to the shape of the diagram in the previous example. The description of the behavior of the diagram of vector spaces obtained by applying homology to the individual terms should be a useful way of tracking how the data behaves dynamically, assuming one can find a useful summary of the nature of the diagram of vector spaces.

Example: Density estimation is an important subject in statistics, and much of what one wants to do in analyzing a data set is to describe in a useful way the behavior of functions which estimate density in some way (see [58]). One way to do this is via kernel density estimators. Suppose that we are given a data set X embedded in Euclidean space $\mathbb{R}^n$. For any point $v \in \mathbb{R}^n$, and positive number ϵ, we let $\gamma_{v,\epsilon}$ denote a spherically symmetric Gaussian distribution with center at v and with variance ϵ. Then one can construct the function

$$\delta_\epsilon = \frac{1}{\#(X)} \sum_{x \in X} \gamma_{x,\epsilon}$$

as an estimate of the density of the distribution from which X arises by sampling. The resulting function depends on the parameter ϵ. For large values of ϵ, one is estimating density in a way which assigns significant weight to points which are far from the given point x, and for smaller values of ϵ, one is estimating density where one weights much more heavily the points in a smaller neighborhood of x. If $\epsilon < \epsilon'$, then $\delta_{\epsilon'}$ is a smoothed out version of δ_ϵ. For a given X, it is an interesting question to determine which choice of ϵ is "correct", and statisticians have developed useful heuristics along these lines. Another approach, though, is to attempt to provide a summary of the behavior of invariants over the full range of values of ϵ at once. Fixing a percentage threshhold T, we define the set $X[T, \epsilon]$ to be the set of points lying in the $T\%$ densest points as measured by the estimator δ_ϵ. If we now fix a sequence of values $\epsilon_0 < \epsilon_1 < \cdots < \epsilon_k$, and set $Z_i = X[T, \epsilon_i] \cup X[T, \epsilon_{i+1}]$, we can now construct the following diagram of data sets.

$$\begin{tikzcd}
\cdots & & Z_i & & Z_{i+1} & & \cdots \\
& X[T, \epsilon_i] \arrow[ul] \arrow[ur] & & X[T, \epsilon_{i+1}] \arrow[ul] \arrow[ur] & & X[T, \epsilon_{i+2}] \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

If we apply the Vietoris-Rips construction with a fixed parameter value to the diagram, and then apply H_j for some j, we will obtain another diagram of vector spaces of the same shape as what we have been looking at in the earlier examples. It should then be helpful in tracking the qualitative behavior of the density estimator with varying ϵ.

Example: We recall the witness complex construction introduced in Section 2.2. Recall that this was a complex $W^w(X, \mathcal{L}, \epsilon)$ constructed on the point cloud X, using a finite "landmark set" $\mathcal{L}$ and a positive parameter ϵ. It would clearly be informative to understand the extent to which homology calculations or clustering depends on the choice of landmark set $\mathcal{L}$, but there is no apparent relationship between the complexes $W^w(X, \mathcal{L}, \epsilon)$, even if the one landmark set is contained in the other. One can, however, proceed as follows. Given a point cloud X and two landmark sets $\mathcal{L}$ and $\mathcal{L}'$, we construct a two-variable version of the witness complexes, denoted $W^w(X, \mathcal{L}, \mathcal{L}', \epsilon)$. Its vertex set is $\mathcal{L} \times \mathcal{L}'$, and we declare that a collection $\{(l_0, l_0'), \ldots, (l_k, l_k')\}$ spans a k-simplex if and only if there exist ϵ weak witnesses x and x' in X for the collections $\{l_0, \ldots, l_k\}$ and $\{l_0', \ldots, l_k'\}$, respectively. It is roughly analogous to the Čech complex of the covering by intersections of Voronoi cells in two different Voronoi decompositions of a metric space. We note that we have projections

$$W^w(X, \mathcal{L}, \mathcal{L}', \epsilon) \to W^w(X, \mathcal{L}, \epsilon) \qquad \text{and} \qquad W^w(X, \mathcal{L}, \mathcal{L}', \epsilon) \to W^w(X, \mathcal{L}', \epsilon)$$

induced by the vertex maps $\mathcal{L} \times \mathcal{L}' \to \mathcal{L}$ and $\mathcal{L} \times \mathcal{L}' \to \mathcal{L}'$. Suppose now that we have a family of landmark sets $\mathcal{L}_i$ for X. We let $W_i = W^w(X, \mathcal{L}, \epsilon)$ and $U_i = W^w(X, \mathcal{L}_i, \mathcal{L}_{i+1})$. Then we have a diagram

$$\begin{tikzcd}
\cdots & & W_i & & W_{i+1} & & \cdots \\
& U_{i-1} \arrow[ul] \arrow[ur] & & U_i \arrow[ul] \arrow[ur] & & U_{i+1} \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

Once again, one can apply homology to the diagram, and the compatibility of classes under the maps in this diagram should be a useful indication that the classes are intrinsic to X and do not depend on the choices of landmark points. This kind of compatibility information would, for example, be extremely useful in interpreting the neuroscience results described in Section 2.5.

The question we now face is how to formalize the notion of compatibility in these diagrams. Here is how one can proceed. We first define a partial ordering on the set of integers $\mathbb{Z}$. We will declare that for every integer k, we have $2k+1 > 2k$ and $2k+1 > 2k+2$. Thus, every odd number is greater than its two adjacent even numbers, and except for identities, there are no other comparabilities.

$$\begin{tikzcd}
\cdots \arrow[dr, ">"] & & 2k-1 \arrow[dl, ">"'] \arrow[dr, ">"] & & 2k+1 \arrow[dl, ">"'] \arrow[dr, ">"] & & 2k+3 \arrow[dl, ">"'] \arrow[dr, ">"] & & \cdots \arrow[dl, ">"'] \\
& 2k-2 & & 2k & & 2k+2 & & 2k+4 &
\end{tikzcd}$$

We will denote this partially ordered set by $\mathcal{Z}$, and we will denote by $\mathcal{Z}[m, n]$ the subset of all integers i with $m \leq i \leq n$, where m and n are integers with $m \leq n$. We note that all the examples above are either $\mathcal{Z}$-persistence sets or $\mathcal{Z}[m, n]$-persistence sets, and when one applies a Vietoris-Rips construction one obtains corresponding $\mathcal{Z}$ or $\mathcal{M}(m, n)$-persistence simplicial complexes. Applying H_i for some i to each of these diagrams then gives $\mathcal{Z}$ or $\mathcal{Z}[m, n]$-persistence vector spaces. The key ingredient in ordinary persistence is the observation that there is a classification of $\mathbb{N}$-persistence vector spaces, and it turns out that there is a classification of the isomorphism classes of $\mathcal{Z}[m, n]$-persistence vector spaces, which is proved and discussed in [29]. We will describe the structure of this classification.

Fix $m \leq n$. For any $m_0 \leq n_0$, with $m \leq m_0 \leq n_0 \leq n$, we will define an elementary object $E(m_0, n_0)$ as follows. In order to specify a $\mathcal{Z}[m,n]$-persistence vector spaces, it is sufficient to specify (a) a vector space V_i for all $m \leq i \leq n$ and (b) linear transformations $V_{2k+1} \rightarrow V_{2k}$ and $V_{2k+1} \rightarrow V_{2k+2}$ whenever the two vector spaces V_i involved in the linear transformation are defined. To define $E = E(m_0, n_0)$, we set $E_i = k$ for $m_0 \leq i \leq n_0$, and $E_i = \{0\}$ otherwise, and we declare that all transformations are the identity if they can be. If they cannot, then they involve a vector space which contains only the zero vector, and are therefore of necessity equal to zero . See below for a picture of $E(1,4)$.

$$\begin{tikzcd}
\{0\} \arrow[dr] & & \{0\} \arrow[dl] \arrow[dr] & & k \arrow[dl] \arrow[dr, "id"] & & k \arrow[dl, "id"'] \arrow[dr, "id"] & & \{0\} \arrow[dl] \arrow[dr] & & \{0\} \arrow[dl] \\
& \cdots & & \{0\} & & k & & k & & \cdots &
\end{tikzcd}$$

The upper row consists of the vector spaces for the odd integers, and the lower row of the ones for the even integers. The three dots indicate an array of zero vector spaces. Now, as in Section 2.3, we have a notion of morphism of $\mathcal{Z}[m,n]$-persistent vector spaces, and we can therefore ask for the classification up to isomorphism of then. Here is the theorem, which can be extracted from [29].

Theorem 4.5 *Let M denote any $\mathcal{Z}[m,n]$-persistent vector space, so that every vector space V_i is finite dimensional. Then there is an isomorphism*

$$M \cong \bigoplus_{j=1}^{t} E(m_j, n_j)$$

for some t and a family of pairs of integers (m_j, n_j), such that $m \leq m_j \leq n_j \leq n$ for all j. Moreover, this decomposition is essentially unique, in the sense that t is the same for all such decompositions, and further that the pairs (m_j, n_j) are also unique up to a reordering of elements.

One can use this result to obtain the following straightforward consequence.

Corollary 4.6 *We say a $\mathcal{Z}$-persistence vector space M is* finite *if each M_i is finite dimensional and there exists an integer N such that $M_i = \{0\}$ if $|i| \geq N$. Any finite $\mathcal{Z}$-persistence vector space M can be decomposed as*

$$M \cong \bigoplus_{j=1}^{t} E(m_j, n_j)$$

for some t and a family of pairs of integers (m_j, n_j), such that $m_j \leq n_j$. The decomposition is again essentially unique.

Remark: $\mathcal{Z}[m,n]$-persistence vector spaces are examples of *quiver representations*, a highly developed area of algebra. See [29] for a complete description.

The families of intervals in each of these decompositions can be interpreted as barcodes with integer valued endpoints. We argue that the analysis of these diagrams should be useful in the analysis of the kinds of problems we have discussed above. We give intuitive illustrations of how this might work. We suspect that there are theorems which could be proved in this direction, and we hope that that will be the subject of future work.

Example: In the context of clustering samples as above, supposing that one obtains a β_0-barcode with two long lines and a family of shorter lines. This outcome would suggest the possibility that one is really looking at two essential clusters in the full data set, with others arising out of the sampling. We view this idea as a potential contributor to the study of consistency or stability of clustering [4], [43].

Example: In the context of variable landmark sets as above, suppose one obtains a β_1-barcode with one long line. This would be confirming evidence toward the hypothesis that the original data set has a β_1 of one, and that what one is seeing in each of the witness complexes is a reflection of that qualitative feature.

Example: In considering dynamic data, the barcodes obtained should be useful in understanding the topological transitions occurring in a changing data set. They should, for example, give a guide to the behavior of clusters over time.

Remark: It is interesting to note that the problems mentioned here are interesting and difficult even in the analysis of β_0, so that the methods we propose should give interesting new information about the behavior of clustering.

4.3 Tree based persistence

As we have seen above, algebraic topology is capable of producing signatures which indicate the presence of topological features within a space. As it stands, however, it is not capable of describing the source of the feature, i.e. where in the space the hole or other feature is located. By using persistence, one is able to develop a systematic way of addressing such questions. That methodology has been described in [65]. In this section, we describe what was done there, and also suggest another persistence framework into which it might fit.

We suppose that we are given a simplicial complex Σ and a covering $\mathcal{S} = \{\Sigma_\alpha\}_{\alpha \in A}$ of Σ by subcomplexes. For example, if we suppose that we are given a set of point cloud data X and a covering $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ of X, one could obtain a covering of $VR(X, \epsilon)$ by the full subcomplexes of $VR(X, \epsilon)$ on the vertex sets U_α. The use of filters as described in Section 3 above can provide coverings of X, or one could cover using balls with centers distributed through the space. One might also cover X by versions of Voronoi cells, as in the discussion of witness complexes in Section 2.2.

Definition 4.7 *Let $x \in H_i(\Sigma)$. We say x is $\mathcal{S}$-*small *if $x \in im(i_\alpha : H_i(\Sigma_\alpha) \to H_i(\Sigma))$ for some α.*

Once one determines that a class is $\mathcal{S}$-small, and for which α the class x is in the image of i_α, one has in effect shown that the feature at least can be represented within the given subset of Σ, and so has information about the source of the class. We recall the Mayer-Vietoris blowup construction $\mathcal{M}(|\Sigma|, \mathcal{S})$ from Section 3.2. This is a space equipped with projection maps $\pi_\Delta : \mathcal{M}(|\Sigma|, \mathcal{S}) \to \Delta[A]$, where A is the indexing set for the covering $\mathcal{S}$, and $p_\Sigma : \mathcal{M}(|\Sigma|, \mathcal{S}) \to |\Sigma|$. It can be shown that p_Σ is a homotopy equivalence (see [65]). One can consider the *skeletal filtration* $\{\Delta[A]^{(k)}\}$ on $\Delta[A]$, where $\Delta[A]^{(k)}$ is the subspace consisting of the union of all faces of dimension $\leq k$, and the corresponding filtration $\pi_\Delta^{-1}(\Delta[A]^{(k)})$ on $\mathcal{M}(|\Sigma|, \mathcal{S})$. The following propositions are now easy to verify.

Proposition 4.8 *A homology class $x \in H_i(|\Sigma|)$ is $\mathcal{S}$-small if and only if x is in the image of the homomorphism $H_i(\pi_\Delta^{-1}(\Delta[A]^{(0)})) \to H_i(\mathcal{M}(|\Sigma|, \mathcal{S}))$.*

Proposition 4.9 *If a homology class $x \in H_i(|\Sigma|)$ is in the image of*

$$H_i(|\Sigma_{\alpha_0}| \cup |\Sigma_{\alpha_1}| \cup \cdots \cup |\Sigma_{\alpha_k}|) \to H_i(|\Sigma|)$$

for some $(k+1)$-fold collection of elements of $\mathcal{S}$, then x is in the image of the homomorphism

$$H_i(\pi_\Delta^{-1}(\Delta[A]^{(k)})) \to H_i(\mathcal{M}(|\Sigma|, \mathcal{S}))$$

Note that this means that we have an $\{0, 1, \ldots, \#(A)-1\}$-persistence vector space $\{H_i((\pi_\Delta^{-1}(\Delta[A]^{(j)}))\}_j$, which contains information about the where the homology classes in Σ arise from. If we are asking whether the class arises from an individual set in $\mathcal{S}$, then the persistence vector space gives us a complete answer to this question. If the question is instead whether or not an element arises from a union of k elements of $\mathcal{S}$, then we do not obtain complete information this way, but we do obtain partial information in that we can preclude the possibility that a class arises from such a union. Examples of the application of this approach are given in [65].

We will also suggest the possibility of another approach to the question of determining the origin of homology classes. Consider first any rooted tree (T, v_0), where v_0 is the root. The vertex set of T can now be given a partial ordering $\leq_T$ defined by $v_1 \leq_T v_2$ if and only if the shortest path from v_1 to v_0 contains v_2. The properties of trees guarantee that $\leq_T$ is a partial ordering. Next, suppose that we have a simplicial complex Σ with a family of coverings $\mathcal{S}_i = \{\Sigma_\alpha\}_{\alpha \in A_i}$ by subcomplexes, equipped with functions $\theta_i : A_i \to A_{i+1}$ such that for any simplex $\sigma \in \Sigma$ and $\alpha \in A_i$, we have that $\sigma \in \Sigma_\alpha$ implies that $\sigma \in \Sigma_{\theta_i(\alpha)}$. We suppose that there is an integer N so that A_N consists of a single element, and that therefore the covering $\mathcal{S}_N$ consists only of Σ itself. This covering data gives rise to a rooted tree whose vertex set is $\coprod_{i=1}^N A_i$, and where the edges are all of the form $(\alpha, \theta_i \alpha)$, for some $\alpha \in A_i$. The single element in A_N is a root for the tree. This kind of family of coverings can arise in natural ways from certain coverings of complexes or spaces.

Example: Consider the unit interval $I = [0,1]$, and fix N. Let $\mathcal{S}_i$ denote the covering of I given by the family of intervals $[\frac{k}{2^{N-i}}, \frac{k+1}{2^{N-i}}]$, where $0 \leq k \leq 2^{N-i} - 1$ is an integer. Let A_i denote $\{k \in \mathbb{Z} | 0 \leq k \leq 2^{N-i} - 1\}$, and define $\theta_i : A_i \to A_{i+1}$ to be the function $k \to \lfloor \frac{k}{2} \rfloor$. We now have a family of coverings, which become increasingly "coarse" as i increases. The associated tree is a binary tree with 2^{N-1} leaves.

Example: If our space is I^n instead, we may cover by products of the sets in the coverings $\mathcal{S}_i$.

Let $\mathcal{S}_i = \{U_\alpha\}_{\alpha \in A_i}$ be a family of coverings as above, and $\theta_i : A_i \to A_{i+1}$ be maps of coverings as above. Let (T, v_0) denote the associated rooted tree, and V_T its vertex set equipped with the partial ordering $\leq_T$. Then we define an associated $(V_T, \leq_T)$-persistence vector space $\{W_t\}_{t \in V_T}$ as follows. Given $\alpha \in A_i \subseteq V_T$, we set $W_\alpha = H_i(U_\alpha)$, and whenever $\alpha \leq_T \alpha'$, then the associated linear transformation from $H(U_\alpha)$ to $H(U_{\alpha'})$ is the map induced from the inclusion $U_\alpha \hookrightarrow U_{\alpha'} = U_{\theta_i(\alpha)}$. The idea of studying the source of homology classes should now be rephrased in terms of invariants of (V_T, v_0)-persistent vector spaces. This situation is a bit like the situation encountered in Section 4.1 in that the classification will involve points on positive dimensional varieties over the ground field. Nevertheless, it appears plausible that one can construct useful invariants, such as the rank invariant discussed there.

5 Reasoning about clustering

As we have noted above, clustering algorithms are methods which take as input a finite metric space (X, d) and produce as output a partition $\Pi(X, d)$ of the underlying set X. In [39], J. Kleinberg proves a non-existence theorem for clustering algorithms satisfying certain properties, in a spirit similar to that of the Arrow impossibility theorem. We will enumerate three properties which may be satisfied by clustering algorithms.

1. **Scale invariance:** Given a finite metric space (X, d), the partitions of X associated to (X, d) and (X, rd), where $r > 0$, are identical.

2. **Richness:** Any partition of a finite set X can be realized as $\Pi(X, d)$ for some metric d on X.

3. **Consistency:** Suppose that d and d' are two metrics on a finite set X, and suppose that (a) for any x, x' contained in one of the clusters attached to d (i.e. blocks in the partition $\Pi(X, d)$) we have $d'(x, x') \leq d(x, x')$ and (b) for any x, x' which belong to distinct clusters attached to d, we have $d'(x, x') \geq d(x, x')$. Then $\Pi(X, d) = \Pi(X, d')$.

Kleinberg's theorem is now

Theorem 5.1 (Kleinberg, [39]) *There are no clustering algorithms which satisfy scale invariance, richness, and consistency.*

This interesting result is disappointing in that it does not give guidance concerning the choices of clustering algorithms, but rather points out deficiencies from which all clustering algorithms must suffer. It is therefore interesting to identify situations in which one can prove existence, and perhaps existence and uniqueness given certain properties. As Kleinberg points out, it is possible to do so in a number of different ways by specifying cost functions on particular clusterings, and prove a uniqueness results for optimal choices of clusterings with respect to this cost function, but that this is perhaps less interesting in that cost functions can often be defined which will isolate a particular algorithm. In [9], such a context is developed which is not dependent on the choice of a cost function but is rather on "structural" criteria which have a great deal in common with Kleinberg's requirements. We now describe the main result of [9].

We begin with the informal observation that clustering for finite metric spaces can be thought of as the statistical version of the geometric notion of forming the set $\pi_0(X)$ of connected components of a topological space X. We note that the correspondence $X \to \pi_0(X)$ can actually be viewed as a functor (see [44]) from the category of topological spaces to the category of sets, in the sense that a continuous map $f : X \to Y$ induces a map of sets $\pi_0(f) : \pi_0(X) \to \pi_0(Y)$, satisfying certain obvious conditions on composite maps and identity maps. This observation is much more than a curiosity. It is the basis for many comparison theorems in topology, and in fact underlies the combinatorization of topology obtained via simplicial sets [45], [19]. It is also what underlies the theoretical constructions underlying the Mapper algorithm discussed in Section 3. Finally, it is also the basis for *etale homotopy theory*, which adapts topological methods to the study of number theoretic problems. The naturality of this condition as well as its utility in many other mathematical contexts suggests that it is very natural to formulate such a condition for clustering algorithms as well. We will therefore attempt to describe clustering algorithms as functors between two categories, where the domain category has as its objects the collection of finite metric spaces. One must therefore first define a notion of morphism of finite metric spaces. We define several such notions.

1. **Isometry:** An *isometry* from a finite metric space (X, d_X) to another finite metric space (Y, d_Y) is a bijective map of sets $f : X \to Y$ so that

$$d_Y(f(x), f(x')) = d_X(x, x') \text{ for all } x, x' \in X$$

.

2. **Embeddings:** An *embedding* from a finite metric space (X, d_X) to another finite metric space (Y, d_Y) is a map of sets $f : X \to Y$ so that

$$d_Y(f(x), f(x')) = d_X(x, x') \text{ for all } x, x' \in X$$

.

3. **Monomorphisms:** A monomorphism from a metric space X to another metric space Y is a monic set map $f : X \to Y$ so that for all $x, x' \in X$, we have $d_Y(f(x), f(x')) \leq d(x, x')$.

4. **Distance non increasing maps:** A distance non increasing (dni) map from a finite metric space (X, d_X) to another finite metric space (Y, d_Y) is a map of sets $f : X \to Y$ so that $d_Y(f(x), f(x')) \leq d_X(x, x')$ for all $x, x' \in X$.

Each of these notions creates the structure of a category whose set of objects are the finite metric spaces, since one can readily observe that each of the classes of morphisms is closed under composition and contains the identity. We denote each of these categories by $\underline{\mathcal{M}}^{iso}$, $\underline{\mathcal{M}}^{emb}$, $\underline{\mathcal{M}}^{mon}$, and $\underline{\mathcal{M}}^{gen}$, respectively. One could initially hope to study clustering algorithms as functors from each of these categories to sets. Thinking in these terms, though, it is first clear that not every functor on one of these categories deserves the name "clustering functor". To see this, we return to the geometric notion of connected components. Not only is the correspondence $\pi_0(-)$ a functor from spaces to sets, it is equipped with a natural surjective map of sets $\eta_X : X \to \pi_0(X)$. Further, these surjective maps have the property that for every continuous map $f : X \to Y$, the diagram of sets

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\eta_X"'] & Y \arrow[d, "\eta_Y"] \\
\pi_0(X) \arrow[r, "\pi_0(f)"] & \pi_0(Y)
\end{tikzcd}$$

commutes.

Remark: In formal terms, the maps η_X, as X runs over all topological spaces, form a *natural transformation* from $\mathcal{S}$ to $\pi_0(X)$, where $\mathcal{S}$ denotes the "underlying set" functor from the category of topological spaces to the category of sets.

By arguing with the analogy with the connected component construction, we will begin with a provisional definition of a $\underline{C}$-functorial clustering algorithm, where $\underline{C}$ is one of the above mentioned categories. It will be a functor χ from $\underline{C}$ to the category of sets together with a family of surjective maps of sets $\eta_{(X,d)} : X \to \chi(X, d)$ so that the diagrams

$$\begin{tikzcd}
X \arrow[r, "f"] \arrow[d, "\eta_X"'] & Y \arrow[d, "\eta_Y"] \\
\chi(X, d_X) \arrow[r, "\chi(f)"] & \chi(Y, d_Y)
\end{tikzcd}$$

commute for every morphism $f : (X, d_X) \to (Y, d_Y)$ in $\underline{C}$.

It is clear that a $\underline{C}$-functorial clustering algorithm is a clustering algorithm in the sense of Kleinberg, since the surjective map of sets from X to $\chi(X, d_X)$ yields a partition of X, namely the partition whose blocks are the sets $\eta_X^{-1}(z)$, as z ranges over all elements of $\chi(X, d_X)$. This means that Kleinberg's conditions also make sense in this context. We examine the possible clustering functors in two of these cases.

Example: $\underline{\mathcal{M}}^{iso}$-functorial clustering algorithms are very simple to describe. Let $\mathcal{I}$ denote the collection of isometry classes of finite metric spaces, and for each ι let (X_ι, d_ι) denote an element of the isomorphism class ι. Let Let G_ι denote the automorphism group of (X_ι, d_ι), and let $\mathcal{P}_\iota$ denote the set of all possible partitions on X_i. Clearly the group G_ι acts on the set $\mathcal{P}_\iota$, and we let $\mathcal{P}_\iota^{G_\iota}$ denote the fixed point set of that action. A $\underline{\mathcal{M}}^{iso}$-functorial clustering algorithm determines a choice $p_\iota \in \mathcal{P}_\iota^{G_\iota}$ for every ι, and conversely an arbitrary choice of such p_ι's determines a $\underline{\mathcal{M}}^{iso}$-functorial clustering algorithm. If we impose Kleinberg's scaling condition, we must instead decompose the set of all finite metric spaces into equivalence classes, where two metric spaces are in the same equivalence class if and only if one is isometric to a rescaling of the other. The classification is now determined exactly as above, except one needs only to make a choice for each of these new equivalence classes.

Example: The study of general $\underline{\mathcal{M}}^{gen}$-functorial and $\underline{\mathcal{M}}^{mon}$ clustering algorithms is more subtle, and we do not yet have a general classification. We will instead give some examples.

- Fix a threshhold parameter ϵ, and for a finite metric space (X, d_X), we let $\sim_\epsilon$ denote the equivalence relation generated by the relation R_ϵ on X defined by $xR_\epsilon x'$ if and only if $d(x, x') \leq \epsilon$. Then the clustering algorithm which assigns to each (X, d_X) the partition associated to $\sim_\epsilon$ is clearly $\underline{\mathcal{M}}^{gen}$-functorial. This example corresponds to single linkage clustering with a fixed parameter value ϵ.
- Consider the finite metric space $\Delta[n]_\epsilon$ whose elements are $\{1, 2, \ldots, n\}$, and where $d(i, j) = \epsilon$ for all pairs $1 \leq i < j \leq n$. For any finite metric space (X, d_X) we define a new relation $\mathcal{R}_\epsilon$ on X by the requirement that $x\mathcal{R}_\epsilon x'$ if and only if there is a distance non-increasing inclusion $j : \Delta[n]_\epsilon \hookrightarrow (X, d_X)$ such that $j(1) = x$ and $j(2) = x'$. We then let E_ϵ denote the equivalence relation generated by $\mathcal{R}_\epsilon$. The clustering algorithm which assigns to each (X, d_X) the partition of X into the blocks of the equivalence relation E_ϵ is now clearly $\underline{\mathcal{M}}^{mon}$-functorial. This notion of clustering is closely related to clique clustering algorithms in network clustering [53].
- More generally, for any family Φ of finite metric spaces, one can define clustering algorithms attached to Φ by analogy with the previous example, where the relation involves injective maps from elements of the family or possible arbitrary maps of finite metric spaces from the family.
- For any finite metric space (X, d_X), define $\mu(X)$ to be the minimal non-zero distance between distinct points of X. The clustering algorithm which assigns to each metric space (X, d_X) the clustering associated to the equivalence relation generated by $R_{\frac{1}{\mu(X)}}$ is readily checked to be $\underline{\mathcal{M}}^{mon}$-functorial.

It is now easy to show that if one imposes the scale invariance condition of Kleinberg, one finds that there are no non-trivial $\underline{\mathcal{M}}^{gen}$-functorial claustering algorithms, where the trivial ones are understood to mean the discrete one (with one element clusters) and the indiscrete one (in which X forms the single cluster for every (X, d_X)).

Non-existence results are of course interesting, but more useful from the point of view of applying and using clustering algorithms are situations where existence and uniqueness can be proved. The non-existence result mentioned in the previous example suggests that one should look for a more relaxed framework. In order to do this, we will change the target category for clustering algorithms from the category of sets to the category of $\mathbb{R}_+$-persistent sets, where $\mathbb{R}_+$ denotes the non-negative real numbers. This is not an unreasonable thing to do in view of the fact that hierarchical clustering algorithms do not report single partitions but rather *dendrograms*, which are roughly speaking $\mathbb{R}_+$-persistent sets. We have already defined the notion of morphisms of $\mathbb{R}_+$-persistent objects in any category. A morphism of persistent sets is said to be *surjective* if each of the individual morphisms which make it up is surjective. For any finite metric space (X, d_X), we associate to it the $\mathbb{R}_+$-persistent set $\alpha(X) = \{\alpha(X)_r\}_r$ for which $\alpha(X)_r = X$ for all $r \in \mathbb{R}_+$, and so that all the morphisms $\alpha(X)_r \to \alpha(X)_{r'}$, for $r \leq r'$, are the identity morphisms on X. Then by a persistent clustering algorithm we will mean an assignment to every finite metric space (X, d_X) a persistent set $\xi(X, d_X)$ and a surjective morphism $\alpha(X) \to \xi(X, d_X)$ of persistent sets for every finite metric space (X, d_X). Letting $\underline{C}$ be any of the category structures on the collection of finite metric spaces given above, one can now define a corresponding notion of $\underline{C}$-functorial persistent clustering algorithms as follows. Such a clustering algorithm is a functor ξ from $\underline{C}$ to the category of persistent sets, equipped with a surjective morphism of persistent sets $\eta_X : \alpha(X) \to \xi(X, d_X)$ for every (X, d_X), so that the diagrams

$$\begin{tikzcd}
\alpha(X) \arrow[r, "\alpha(f)"] \arrow[d, "\eta_X"'] & \alpha(Y) \arrow[d, "\eta_Y"] \\
\xi(X, d_X) \arrow[r, "\xi(f)"] & \xi(Y, d_Y)
\end{tikzcd}$$

commute for every morphism f in $\underline{C}$. In this context, Kleinberg's scale invariance condition takes a different form. For any $\mathbb{R}_+$-persistent set $\mathcal{U} = \{\mathcal{U}_r\}$ and any positive real number t, we define the persistent set $t\mathcal{U}$ by $t\mathcal{U}_r = \mathcal{U}_{\frac{r}{t}}$, and by requiring that for any $r \leq r'$, the homomorphism $t\mathcal{U}_r \to t\mathcal{U}_{r'}$ be identified with the corresponding homomorphism $\mathcal{U}_{\frac{r}{t}} \to \mathcal{U}_{\frac{r'}{t}}$. The correspondence $\mathcal{U} \to t\mathcal{U}$ is clearly a functor from the category of persistent sets to itself, which we will write as δ_r. We also have the rescaling functor ρ_r from the category $\underline{C}$ to itself, which simply multiplies all distances by r. We now say that a persistent clustering algorithm ξ is *scale invariant* if (a) $\chi(\rho_r(X, d_X)) = \delta_r(\chi(X, d_X))$ and (b) $\eta_{\rho_r(X,d_X)} = \delta_r(\eta_{(X,d_X)})$. In [9], the following result is proved.

Theorem 5.2 *Let E denote a metric space with two points, and so that the distance between those two points is $= 1$. Let P denote the persistent set for which P_r consists of two points for $r < 1$, of one point for $r \geq 1$, and so that all the maps $P_r \to P_{r'}$ are surjective for all $r \leq r'$. There is a unique $\underline{\mathcal{M}}^{gen}$-functorial persistent clustering algorithm Ξ which is scale invariant and which satisfies the requirement that $\Xi(E) = P$. The algorithm Ξ is the algorithm which associates to a finite metric space (X, d_X) the persistent set $\{\pi_0(VR(X, \epsilon))\}_{\epsilon \geq 0}$.*

The proof is not difficult. This result is in the spirit of Kleinberg's result in that the requirements which define it are structural rather than requiring the optimization of some cost function, but yields an existence and uniqueness result. This theorem therefore precludes a number of interesting algorithms such as average linkage and complete linkage clustering from being functorial. Users of clustering algorithms often assert that average linkage and complete linkage clustering are to be favored over single linkage clustering because the clusters tend to be more "compact" in an intuitive sense. We believe that their observation should be interpreted as saying that in clustering one needs to take into account more than just the metric as a geometric object, but in addition some notion of density. This suggests the possibility of including density into notions of functoriality, which we will explore in a future paper.

6 What should the theorems be?

We have presented some examples of how topological methods can be applied to study data sets. The methods provide signatures which yield information about the shape of the data sets being studied, and can also provide useful methods for visualizing data sets. However, there are many outstanding questions of a statistical nature. One situation which illustrates some of these questions was that which occurred in the discussion of the neuroscience data in Section 2.5 above. In that case, one found that the data exhibited barcodes which appeared to indicate the presence of non-zero β_1 and β_2 in the data sets. While the observation is of course suggestive, to prove that it indicates structure in the data set one must show that segments of the indicated length cannot occur in a random model associated to a null hypothesis. In the paper [59], this was shown by simulation, which is of course always an option. However, it is clear that if it were the case that such results could actually be proved, one would avoid the time and effort spent in simulation, and it would also provide the basis for thinking more systematically about significance questions for the barcodes arising out of persistent homology. We outline how one might begin formulating such results.

Suppose we are given a metric space X and a probability measure on X. Then we can consider an experiment consisting of selecting n points $\{x_1, x_2, \ldots, x_n\}$ i.i.d. from X. For each time this experiment is carried out, we can then form the $\mathbb{R}_+$-persistent simplicial complex $\{VR(\{x_1, \ldots, x_n\}, \epsilon)\}_{\epsilon \geq 0}$, the corresponding homology vector spaces $\{H_i(VR(\{x_1, \ldots, x_n\}, \epsilon))\}_{\epsilon \geq 0}$, and finally therefore a barcode. We note that the barcode is simply a family of intervals with endpoints in $\mathbb{R}_+$. Each such interval can be considered as a point in the set $D = \{(x, y) \in \mathbb{R}^2 | x \leq y\}$, and the output is therefore a finite subset of the set D. As such, we now have a *finite spatial point process* (see [20], [3]). Roughly, a finite spatial point process with values in a locally compact second countable Hausdorff space S is a probability measure on a σ-algebra constructed on the collection of finite counting measures on S. The σ-algebra is the minimal one which makes all the counting maps Φ_B measurable, where B is in the Borel σ-algebra associated to S, and Φ_B evaluates the counting measure on B. Point processes are a heavily studied area within statistics.

We believe that theorems which describe these point processes will be very useful in applying algebraic topology to the qualitative analysis in many areas of science. Here are some suggestions which would be interesting.

- Determine the distributions of various statistics, such as maximum and average distance to the diagonal, on the point processes attached to the Vietoris-Rips complexes obtained by sampling sets of points from various probability distributions on Euclidean space. Gaussian distributions are a good place to start.
- Similarly, determine distributions of various statistics on point processes obtained by selecting sets of landmark points using various strategies from probability distributions on Euclidean space and computing the persistent witness complexes.
- Study consistency of clustering problems by studying the distributions of various statistics on the zig-zag barcodes obtained by sampling from a larger data set.
- Develop a method for studying the output of multidimensional persistent homology probabilistically.

We point out that in the context of ordinary homology (i.e. not persistent homology), M. Penrose and others have developed a theory of “geometric random graphs”, and have proved various results concerning the Betti numbers of complexes attached to such graphs (see [54]). Also, results about barcodes under the general heading have now begun to appear (see [15] and [16]). These results should be an excellent starting point for the development of theorems of the type mentioned above.

References

[1] H. Abdi, *Metric multidimensional scaling*, in **Encyclopedia of Measurement and Statistics**, Sage, Thousand Oaks (CA), (2007), pp. 598-605

[2] H. Adams and G. Carlsson, *On the non-linear statistics of range image patches*, preprint, (2007), available at http://comptop.stanford.edu/preprints/.

[3] A. Baddeley, *Spatial point processes and their applications*, appears in A. Baddeley, I. Brny, R. Schneider, and W. Weil, editors, **Stochastic Geometry: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13-18, 2004**, Lecture Notes in Mathematics 1892, Springer. ISBN 3-540-38174-0, pp. 1–75

[4] S. Ben-David, U.von Luxburg, D. Pal, *A sober look on clustering stability*, in G. Lugosi and H. Simon, editors, Proceedings of the 19th Annual Conference on Learning Theory (COLT), pages 5 - 19, Springer, 2006.

[5] A. Björner, *Topological methods*, appears in **Handbook of Combinatorics**, Vol. 1, 2, 1819–1872, Elsevier, Amsterdam, 1995.

[6] G. R. Bowman, X. Huang,Y. Yao, J. Sun, G. Carlsson, L. J. Guibas, and V.S. Pande, *Structural insight into RNA hairpin folding intermediates*, Journal of the American Chemical Society Communications, July, 2008.

[7] E. Carlsson, G. Carlsson, and V. de Silva, *An algebraic topological method for feature identification*, International Journal of Computational Geometry and Applications, 16 (2006), no. 4, pp. 291-314

[8] G. Carlsson and V. de Silva, *Topological estimation using witness complexes*, Symposium on Point-Based Graphics, ETH, Zürich, Switzerland, June 2-4, 2004.

[9] G. Carlsson and F. Memoli, *Persistent Clustering and a Theorem of J. Kleinberg* , Preprint, March 2008.

[10] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, *Persistence barcodes for shapes*, International Journal of Shape Modeling, 11 (2005), pp. 149-187.

[11] G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian, *On the local behavior of spaces of natural images*, International Journal of Computer Vision, (76), 1, 2008, pp. 1-12.

[12] G. Carlsson and A. Zomorodian, *The theory of multidimensional persistence*, 23rd ACM Symposium on Computational Geometry, Gyeongju, South Korea, June 6-7, 2007

[13] G. Carlsson and T. Ishkanov, *Local structure of spaces of natural images*, preprint, (2007), available at http://comptop.stanford.edu/preprints/

[14] G. Carlsson, G. Singh, and A. Zomorodian, *Computing multidimensional persistence*, in preparation.

[15] D. Cohen-Steiner, H. Edelsbrunner and J.Harer, *Stability of persistence diagrams* Discrete Comput. Geom., **37** (2007), 103–120.

[16] D. Cohen-Steiner, H. Edelsbrunner, J. Harer and Y. Mileyko, *Lipschitz functions have L_p-stable persistence*, Found. Comput. Math., to appear.

[17] A. Collins, A. Zomorodian, G. Carlsson, and L. Guibas, *A barcode shape descriptor for curve point cloud data* Computers and Graphics, Volume 28, 2004, pp.881–894.

[18] D. Cox, J. Little, and Donal O'Shea, **Using Algebraic Geometry**, Graduate Texts in Mathematics, Springer Verlag, 1998, xii + 499 pages, ISBN 0-387-98492-5.

[19] E. Curtis, *Simplicial homotopy theory*, Advances in Math. 6 1971 107-209 (1971).

[20] D. Daley and D. Vere-Jones, **An Introduction to the Theory of Point Processes**, two volumes, Second Edition, Springer Verlag, 2003, ISBN 0-387-95541-0

[21] B. Delaunay, *Sur la sphere vide*, Izvestia Akademii Nauk SSSR, Otdelenie Matematicheskikh i Estestvennykh Nauk, 7:793-800, (1934)

[22] J.G. Dumas, F. Heckenbach, B.D. Saunders, and V. Welker, *Computing simplicial homology based on efficient Smith normal form algorithms*, In *Algebra, Geometry, and Software Systems* (2003), 177-207.

[23] D. Dummit and R. Foote, *Abstract Algebra. Third edition*, John Wiley & Sons, Inc., Hoboken, NJ, 2004. xii+932 pp. ISBN: 0-471-43334-9 00-01 (16-01 20-01)

[24] H. Edelsbrunner, D. Letscher, and A. Zomorodian, *Topological persistence and simplification*, Discrete and Computational Geometry **28**, 2002, 511-533

[25] H. Edelsbrunner and N.R. Shah, *Triangulating topological spaces*, Tenth Annual ACM Symposium on Computational Geometry (Stony Brook, NY, 1994). Internat. J. Comput. Geom. Appl. **7** (1997), no. 4, 365–378.

[26] B. Efron, *Bootstrap methods: another look at the jackknife*, Ann. Statist. 7 (1979), no. 1, pp. 1–26.

[27] S. Eilenberg, *Singular homology theory*, Ann. of Math. (2) 45, (1944). 407–447.

[28] P. Frosini and C. Landi, *Size theory as a topological tool for computer vision*, Pattern Recognition And Image Analysis, vol. 9 (4) (1999), pp. 596-603.

[29] P. Gabriel and A. Roiter, **Representations of Finite-Dimensional Algebras**. Translated from the Russian. With a chapter by B. Keller. Reprint of the 1992 English translation. Springer-Verlag, Berlin, 1997. iv+177 pp. ISBN: 3-540-62990-4

[30] P.G. Goerss, and J.F. Jardine, **Simplicial homotopy theory**, Progress in Mathematics, 174. Birkhuser Verlag, Basel, 1999. xvi+510 pp. ISBN: 3-7643-6064-X

[31] J. Hartigan, **Clustering Algorithms**, Wiley, New York.

[32] T. Hastie, R. Tibshirani, and J. Friedman, **The Elements of Statistical Learning**, Springer, New York, 2001, ISBN: 0-387-95284-5

[33] A. Hatcher, **Algebraic Topology**, Cambridge University Press, Cambridge, 2002. xii+544 pp. ISBN: 0-521-79160-X; 0-521-79540-0

[34] J.H. van Hateren and A. van der Schaaf, 1998. *Independent component filters of natural images compared with simple cells in primary visual cortex*, Proc. R. Soc. Lond.vol. B 265,(1998), pp. 359- 366.

[35] J. Headd, Y.-H. A. Ban, H. Edelsbrunner, M. Vaidya and J. Rudolph. *Protein-protein interfaces: properties, preferences, and projections*, Protein Research, to appear, 2007.

[36] D. Hubel, **Eye, Brain, and Vision**, Scientific American Library, W. H. Freeman, New York, 1995.viii+242pp. ISBN: 0-716-76009-6

[37] P.J. Huber, *Projection pursuit*, Ann. Statistics (13), 2, (1985), pp. 435-525, with discussion.

[38] T. Kenet, D. Bibitchkov, M. Tsodyks, A. Grinvald, and A. Arieli, *Spontaneously emerging cortical representations of visual attributes*, Nature 425, 954 - 956 (2003)

[39] J.M. Kleinberg, *An impossibility theorem for clustering*, NIPS 2002: 446-453

[40] S. Lafon and A.B. Lee, *Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parametrization*, IEEE Transactions on Pattern Analysis and Machine Intelligence **28**, 9 (2006), pp. 1393-1403.

[41] A.B. Lee, K.S. Pedersen, and D. Mumford, *The nonlinear statistics of high-contrast patches in natural images*, International Journal of Computer Vision (54), No. 1-3, August 2003, pp. 83-103.

[42] R.Y.Liu, J.M. Parelius and K. Singh, *Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh)*, Ann. Statist. Volume 27, Number 3 (1999), 783-858.

[43] U. von Luxburg, M. Belkin, and O. Bousquet., *Consistency of spectral clustering*, Annals of Statistics, 36 (2), 555-586, 2008

[44] S. Mac Lane, **Categories for the Working Mathematician**, Second edition. Graduate Texts in Mathematics, 5. Springer-Verlag, New York, 1998. xii+314 pp. ISBN: 0-387-98403-8

[45] J.P. May, **Simplicial objects in algebraic topology**, Reprint of the 1967 original. Chicago Lectures in Mathematics. University of Chicago Press, Chicago, IL, 1992. viii+161 pp. ISBN: 0-226-51181-2

[46] P. McCullagh, *What is a statistical model?* With comments and a rejoinder by the author. Ann. Statist. 30 (2002), no. 5, 1225–1310.

[47] Miller, E., and Sturmfels, B. **Combinatorial Commutative Algebra** Graduate Texts in Mathematics, 227. Springer-Verlag, New York, 2005. xiv+417 pp. ISBN: 0-387-22356-8

[48] R. Miller, *Discussion - projection pursuit*, Ann. Statistics 13,2, 1985, pp. 510-513. (With discussion)

[49] J. Milnor, **Morse Theory**, Based on lecture notes by M. Spivak and R. Wells. Annals of Mathematics Studies, No. 51 Princeton University Press, Princeton, N.J. 1963 vi+153 pp.

[50] D. Mumford,*The dawning of the age of stochasticity*, appears in **Mathematics: Frontiers and Perspectives**, Amer. Math. Soc., Providence, RI, 2000, 197–218.

[51] J. Munkres, **Topology: a First Course**, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975. xvi+413 pp.

[52] P. Niyogi, S. Smale, and S. Weinberger, *Finding the homology of submanifolds with high confidence from random samples*, Discrete and Computational Geometry, vol. 39, nos. 1-3, (2008).

[53] G. Palla, I. Derènyi, I. Farkas, and T. Vicsek, *Uncovering the overlapping community structure of complex networks in nature and society*, Nature, Volume 435, 9 June 2005, pp. 814-818.

[54] M. Penrose, **Random Geometric Graphs**, Oxford Studies in Probability, 5. Oxford University Press, Oxford, 2003. xiv+330 pp. ISBN: 0-19-850626-0

[55] G. Reeb, *Sur les points singuliers d'une forme de Pfaff complètement intégrable ou d'une fonction numèrique*, C.R. Acad. Sc. Paris 222 (1946), pp. 847-849.

[56] S.T. Roweis and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science 290 (2000) (December), pp. 2323-2326.

[57] V. de Silva, R. Ghrist, *Coverage in sensor networks via persistent homology*, Algebraic and Geometric Topology, **7**, 2007, pp.339 - 358.

[58] B.W. Silverman, **Density Estimation for Statistics and Data Analysis**, Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. x+175 pp. ISBN: 0-412-24620-1

[59] G. Singh, F. Memoli, T. Ishkhanov, G. Carlsson, G. Sapiro and D. Ringach, *Topological Structure of Population Activity in Primary Visual Cortex*, Journal of Vision, Volume 8, Number 8, Article 11, pp. 1-18, 2008.

[60] G. Singh, F. Memoli and G. Carlsson, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, Point Based Graphics 2007, Prague, September 2007

[61] J.B. Tenenbaum, V. de Silva and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science 290 (2000) (December), pp. 2319-2323

[62] M. Tsodyks, T. Kenet, A. Grinvald, and A. Arieli, *Linking spontaneous activity of single cortical neurons and the underlying functional architecture*, Science 286, (1999), pp. 1943-1996.

[63] B. Wandell, **Foundations of Vision**, Sinauer Associates, Sunderland, Mass., 1995.xvi+476pp., ISBN:0-878-93853-2

[64] A. Zomorodian and G. Carlsson, *Computing persistent homology*, Discrete and Computational Geometry, **33** (2), 2005, pp. 247-274

[65] A. Zomorodian and G. Carlsson, *Localized homology*, to appear, Computational Geometry: Theory and Applications, 2008.