

Topological approximation by small simplicial complexes

Gunnar Carlsson¹, Vin de Silva^{*,1}

*Department of Mathematics, Building 380, Stanford University
CA 94305-2125, USA.*

Abstract

Given a point-cloud dataset sampled from an underlying geometric space X , it is often desirable to build a simplicial complex \mathcal{S} approximating the geometric or topological structure of X . For example, recent techniques in automatic feature location depend on the ability to estimate topological invariants of X . These calculations can be prohibitively expensive if the number of cells in the approximating complex \mathcal{S} is large. Unfortunately, most existing simplicial approximation algorithms either give too many cells, or involve calculations which are tractable or valid only in low dimensional Euclidean geometry. In this paper we introduce the *combinatorial Delaunay triangulation*, a simplicial complex construction which can be efficiently computed in arbitrary metric spaces, and which gives reliable topological approximations using comparatively few cells.

Key words: point cloud data, topological approximation, computational topology, Delaunay triangulation, persistent homology
1991 MSC: 65D18

1 Topology and point cloud data sets

There are many situations in which a geometrical space is represented by *point cloud data*, that is by a finite set of points sampled from it. It is natural to ask what attributes of the original space can be recovered from this data. For example, a laser scanning device applied to a solid object might return the coordinates of hundreds or thousands of points on the object's 2-dimensional

* Corresponding author.

Email address: `silva@math.stanford.edu` (Vin de Silva).

¹ Both authors supported in part by NSF grant DMS-0101364.

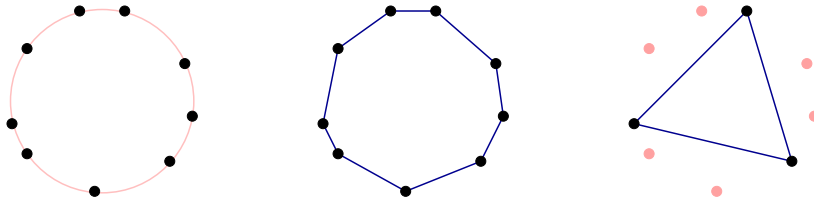


Fig. 1. Data sampled from a circle

surface. Further processing may be necessary to put this information into useable form. One standard operation is to construct a triangulated surface using the data points as vertices. The resulting object may be regarded as an approximation to the original surface, which can be measured or manipulated for certain tasks such as visual rendering.

Such a detailed approximation is entirely appropriate when the goal is to reproduce the contours and other geometric features of an object at fine resolution. On the other hand, when the goal is to measure *coarse* features of the data such as topological invariants, it may be more efficient to construct less detailed (but still topologically faithful) approximations. Figure 1 shows data sampled from a circle embedded in the Euclidean plane (left). A simplicial complex constructed on this data set may be topologically equivalent (middle). However, a much smaller simplicial complex with only three vertices is also topologically equivalent to the circle (right).

In this paper we introduce the *combinatorial Delaunay triangulation* (CDT), a new technique for building a simplicial complex approximation \mathcal{S} to a space $X \subset \mathbb{R}^D$ represented by a finite sample $Z \subset X$. The key idea (compare [1]) is to approximate the intrinsic metric structure of the data by computing shortest paths in a local connectivity graph. The approximation \mathcal{S} is a Delaunay complex defined in terms of this intrinsic metric. The construction has the following characteristics.

- (1) *Landmarking*: The complex \mathcal{S} uses a vertex set $L \subset Z$ considerably smaller than the sample Z itself.
- (2) *Homotopy approximation*: Under favourable circumstances, \mathcal{S} is homotopy equivalent to X , and may be homeomorphic; however it need not be a close geometric approximation.
- (3) *Intrinsic geometry*: The construction estimates and works with the intrinsic geometry of X . As a side effect this largely avoids ‘curse of dimensionality’ effects relating to the embedding dimension n .
- (4) *Efficiency*: the combinatorial Delaunay triangulation is easier to calculate than ‘true’ Delaunay triangulations of the same intrinsic dimension, using Dijkstra-style shortest path computations.
- (5) *Non-redundancy*: \mathcal{S} is not prone to having large redundant clusters of high-dimensional cells such as occur in the Čech construction.

The construction depends on the choice of landmark vertex set $L \subset Z$, and on a local neighbourhood size parameter k or ϵ .

To give some context, we compare the combinatorial Delaunay triangulation with several other simplicial complex constructions in the literature and in folklore. We present the results of experiments which demonstrate that CDT performs well on a synthetic data set and on a real example derived from natural image data [2]. The question of theoretical analysis will be taken up in future publications.

One significant application is to speed up the calculation of topological invariants such as Betti numbers b^p , which count the p -dimensional holes in a topological space [3]. Rapid calculation or estimation of Betti numbers is central to a new family of techniques for detecting singularities and other geometrical features of spaces represented by point-cloud data [4,5]. The calculation of b^p involves solving matrix equations derived from a pair of matrices of sizes $c^{p-1} \times c^p$ and $c^p \times c^{p+1}$, where c^p denotes the number of p -dimensional cells in a simplicial complex representation of the space. These matrices are extremely sparse, but even when that can be exploited it is still highly desirable for the c^p to be as small as possible. Our claim is that CDT provides a fast, accurate method for estimating Betti numbers from point-cloud data.

The paper is organised as follows. Section 2 begins with general discussion of simplicial complex approximation. This is followed by a review of known constructions. In Section 3 we define the combinatorial Delaunay triangulation, together with its immediate predecessor, the Martinetz–Schulten complex. Both constructions are examples of *witness complexes*. We discuss several variations on the theme, including a framework for defining persistent homology groups. Finally, in Section 4 we discuss the performance these algorithms on synthetic and real data examples.

Remark. From one point of view, the combinatorial Delaunay triangulation is an amalgam of two existing ideas. The first is the Martinetz–Schulten construction [6] described in Section 3.2. The second idea is the use of intrinsic path-length geometry instead of extrinsic Euclidean geometry. In the context of non-linear dimensionality reduction this idea is a critical component in the Isomap algorithm of Tenenbaum et al [1]; it allows Isomap to handle non-linear datasets well beyond the capacity of its (essentially linear) parent algorithm, classical MDS [7,8]. Our experience with test data sets suggests that there is a corresponding gain in flexibility over Martinetz–Schulten when we use the intrinsic path-length metric.

Remark. From another point of view, our construction may be interpreted as a glorified *graph Voronoi diagram* of the kind described and studied by

Erwig [9]. In that construction, a (weighted, directed) graph, together with a set of landmark points taken from the vertices, can be decomposed as a union of disjoint Voronoi cells, one for each landmark point. One difference is that in our case the Voronoi cells are required to overlap, since we make use of their intersection structure. To compute overlapping Voronoi cells efficiently, we can use a suitably adapted version of the parallel Dijkstra algorithm described in [9]. The other point is that CDT is properly regarded as a construction on a point-cloud data set, which makes incidental use of a graph at an intermediate stage. For graph Voronoi diagrams, the primary object of study is the graph itself.

2 Simplicial complex approximations

In this section we develop the notion of simplicial complex approximation, and review some well-known examples: Delaunay triangulations, Čech and Rips complexes, and the α -shape complex.

2.1 Overview of the problem

Let $X \subset \mathbb{R}^D$ be a subspace (such as a manifold or a simplicial complex), and let $Z \subset X$ be a finite sample. When we talk of a *simplicial complex approximation*, we are referring to the following scenario:

- (1) A construction $\mathcal{S} = \mathcal{S}(Z)$ of a simplicial complex depending on Z and possibly on additional parameters, but not depending on X .
- (2) A *similarity theorem* or heuristic demonstrating that $X \cong \mathcal{S}(Z)$ (homeomorphism) or $X \simeq \mathcal{S}(Z)$ (homotopy equivalence) under reasonable conditions on Z as a sample of X , and for some choice of values for the additional parameters.

In this paper our emphasis is on estimating coarse topological features such as the homology groups or Betti numbers of X . These are invariant under homotopy equivalence, so that is the appropriate notion of equivalence for our purposes. In different problems such as surface reconstruction [10], homeomorphism equivalence is the goal.

A typical additional parameter is *feature scale* $R \geq 0$. This can sometimes be interpreted as an amount of blurring applied to Z . The complex $\mathcal{S}(Z, R)$ is expected to capture large-scale geometrical features and ignore small-scale features, where ‘large’ and ‘small’ are defined in terms of R . In many cases R is a *nested parameter*, meaning that there are natural *inclusion maps* $\mathcal{S}(Z, R) \rightarrow$

$\mathcal{S}(Z, R')$ whenever $R \leq R'$. These maps may be genuine inclusion maps, with $\mathcal{S}(Z, R)$ being a subcomplex of $\mathcal{S}(Z, R')$; but they need not be. When there is a nested parameter, the inclusion maps induce corresponding maps between homology groups. We can usefully weaken item 2 above to the following:

- (2') A similarity theorem or heuristic relating $H_k(X)$ to the *persistent homology* group

$$\text{Im}[H_k(Z, R) \rightarrow H_k(Z, R')]$$

under reasonable sampling conditions and for some choice of R, R' and the other parameters.

Remark. Roughly speaking, persistent homology [11,12] detects features at scales less than R while filtering out features that are no longer visible at scale R' . More generally if the nested parameter R is allowed to vary over a large range, one can study the full system of homology groups $H_k(Z, R)$ together with the induced maps between them. This leads to extremely powerful tools for studying real data, where a single homology group by itself is likely to be highly unstable with respect to parameter settings and noise.

2.2 Abstract versus geometrical simplicial complexes

We wish to distinguish several different notions of ‘simplicial complex’. An *abstract simplicial complex* \mathcal{S} is specified by the following data:

- A vertex set Z .
- A rule specifying when a given ‘ p -simplex’ $\sigma = [z_0 z_1 \dots z_p]$ belongs to \mathcal{S} ; here the vertices z_0, z_1, \dots, z_p of σ are distinct elements of Z , and the order in which they are listed is disregarded.
- Each p -simplex σ has $p + 1$ faces which are $(p - 1)$ -simplices; each face is obtained by deleting one of the vertices z_0, z_1, \dots, z_p . The membership rule has the property that if σ belongs to \mathcal{S} then all of its faces belong to \mathcal{S} .

This is a purely combinatorial notion, but it implicitly defines a topological space $K(\mathcal{S})$ called the *realisation* of \mathcal{S} . This is defined by taking a copy $K(\sigma)$ of a standard topological p -simplex to represent each abstract p -simplex $\sigma \in \mathcal{S}$; and then ‘gluing’ faces together, so whenever τ is a face of σ we identify $K(\tau)$ with the corresponding face of $K(\sigma)$. The standard p -simplex can be taken to be the convex hull of the basis vectors $\vec{e}_1, \vec{e}_2, \vec{e}_{p+1}$ in \mathbb{R}^{p+1} .

A *geometrical simplicial complex* is given by specifying a topological subspace $S \subset \mathbb{R}^D$, an abstract simplicial complex \mathcal{S} , and a homeomorphism $K(\mathcal{S}) \rightarrow S$ which is affine-linear on each simplex. In other words, it is a simplicial complex concretely embedded in Euclidean space.

Remark. A common situation occurs when the vertex set of an abstract simplicial complex \mathcal{S} corresponds to a set of points in \mathbb{R}^D . In that case there is a unique continuous mapping $\eta : K(\mathcal{S}) \rightarrow \mathbb{R}^D$ called the *embodying map*, which sends the vertices to their corresponding points and which is affine-linear on each simplex. If this happens to be an embedding, then the result is a geometrical simplicial complex. For the purpose of using \mathcal{S} to estimate topological invariants, it is unimportant whether η is an embedding or not.

2.3 Delaunay triangulations and Voronoi diagrams

Since Delaunay triangulations and their dual Voronoi diagrams are ubiquitous in this subject, we briefly discuss their properties and relevance here.

The *Voronoi diagram* $\text{Vor}(Z)$ of a finite point set $Z \subset \mathbb{R}^D$ is the tiling of \mathbb{R}^D by Voronoi cells

$$V_z = \{x \in \mathbb{R}^D : |x - z| \leq |x - y| \text{ for all } y \in Z\}$$

defined for $z \in Z$. Each cell is a convex polyhedron and distinct cells have disjoint interiors. If z is on the boundary of the convex hull $\text{conv}(Z)$ then V_z is unbounded; all other Voronoi cells are bounded.

Dual to the Voronoi diagram is the *Delaunay triangulation* $\text{Del}(Z)$, which is a simplicial complex whose vertex set is Z , and which contains the cell $\sigma = [z_0, z_1, \dots, z_p]$ whenever $V_{z_0} \cap V_{z_1} \cap \dots \cap V_{z_p} \neq \emptyset$. Equivalently, $\sigma \in \text{Del}(Z)$ if there is a point x which is equidistant from z_0, z_1, \dots, z_p and which has no nearer neighbour in Z . We say that x is a *witness* to the cell σ .

For a generic set $Z \subset \mathbb{R}^D$, the Delaunay triangulation is a geometric simplicial complex; the embodying map carries $K(\text{Del}(Z))$ homeomorphically onto the convex hull $\text{conv}(Z)$. This assertion is false in the non-generic situation where there is a witness x equidistant from $n+2$ or more nearest neighbours in Z . The simplest example is when Z consists of four points in \mathbb{R}^2 lying at the vertices of a cyclic quadrilateral; in that case $\text{Del}(Z)$ is abstractly a tetrahedron. In applications where this matters, one can ‘perturb by ϵ ’ to return to the generic case.

Now consider a situation where Z is a finite sample taken from a subspace $X \subset \mathbb{R}^D$ such as a submanifold. By itself the Delaunay triangulation $\text{Del}(Z)$ is a contractible simplicial complex, and so its topological invariants carry no information. However, we can define restricted complexes whose structure does reflect the topology of X . The *restricted Voronoi diagram* $\text{Vor}(Z, X)$ is the decomposition of X as a union of cells $\{V_z \cap X : z \in Z\}$. The *restricted Delaunay triangulation* $\text{Del}(Z, X)$ is the dual of $\text{Vor}(Z, X)$; it contains the cell

$\sigma = [z_0, z_1, \dots, z_p]$ whenever $V_{z_0} \cap V_{z_1} \cap \dots \cap V_{z_p} \cap X \neq \emptyset$. In other words, we demand a witness for σ which lies on X itself.

If $Z \subset X$ is a sufficiently fine, generic sample taken from a smooth submanifold $X \subset \mathbb{R}^D$, the simplicial complex $\text{Del}(Z, X)$ can be shown to be homeomorphic to X . Since this construction depends on knowledge of X , it does not qualify as a simplicial complex approximation. However, it is a useful point of reference: the Martinetz–Schulten complex and the combinatorial Delaunay triangulation can both be regarded as approximations to $\text{Del}(Z, X)$.

2.4 The Čech complex

This is perhaps the most natural construction from a mathematical point of view.

- Vertex set: all the data points in Z
- Parameter: $R > 0$, nested
- Definition: the p -simplex $\sigma = [z_0 z_1 \dots z_p]$ belongs to $\check{\text{Cech}}(Z, R)$ iff the closed Euclidean balls $B(z_j, R/2)$, $j = 0, 1, \dots, p$, have non-empty common intersection.

The Čech complex is precisely the *nerve* [13] of the collection of metric balls $\{B(z, R/2) : z \in Z\}$. Since balls are convex, the Čech theorem implies that $\check{\text{Cech}}(Z, R)$ is homotopy equivalent to the union of these balls. As such it has a straightforward geometrical interpretation.

The significant drawback of the Čech construction is its inefficiency. Whenever k points form a cluster of diameter at most R , there is a corresponding $(k-1)$ -dimensional simplex in $\check{\text{Cech}}(Z, R)$. This leads to prohibitively large complexes when R is large, even when the underlying topological information is very simple.

2.5 The Rips complex

The Rips complex is a variant of the Čech complex which is easier to calculate, but slightly more prone to clustering-related inefficiency.

- Vertex set: all the data points in Z
- Parameter: $R > 0$, nested
- Definition: the p -simplex $\sigma = [z_0 z_1 \dots z_p]$ belongs to $\text{Rips}(Z, R)$ iff for every edge $[z_j z_k]$, $0 \leq j < k \leq p$, we have $|z_j - z_k| \leq R$.

The Rips complex $\text{Rips}(Z, R)$ can be characterised as the largest simplicial complex with the same 1-skeleton as $\check{\text{Cech}}(Z, R)$. The definition makes sense for an arbitrary metric structure on Z , and it avoids the calculations needed to determine whether a set of Euclidean balls has nonempty common intersection.

On the other hand the homotopy type of the Rips complex is harder to interpret directly. For Euclidean data, it approximates the Čech complex in the sense that there are inclusions $\check{\text{Cech}}(Z, R) \subset \text{Rips}(Z, R) \subset \check{\text{Cech}}(Z, 2R)$. The constant 2 can be sharpened to $\sqrt{2}$ (see [14]) but in other regards this seems to be the best that can be said.

In a different spirit, if the distances $|z_j - z_k|$ are measured using an ℓ^∞ norm then the Rips complex is equal to the nerve of a collection of hypercubes of side length $2R$. Then $\text{Rips}(Z, R)$ is homotopy equivalent to the union of those hypercubes, giving a clear-cut interpretation in this case.

2.6 The α -shape complex

This construction, due to Edelsbrunner [15] and based on the Delaunay triangulation, gives a family of complexes of the same homotopy type as the Čech complexes $\check{\text{Cech}}(Z, R)$, but which are considerably smaller in size.

- Vertex set: all the data points in Z
- Parameter: $R > 0$, nested
- If V_Z is the Voronoi diagram for Z and $V_Z(z)$ is the closed Voronoi cell for the data point z , then define the α -cell for z to be the convex set $\alpha(z, R) = B(z, R) \cap V_Z(z)$.
- Definition: The p -simplex $\sigma = [z_0 z_1 \dots z_p]$ belongs to $A(Z, R)$ iff the α -cells $\alpha(z_j, R/2)$, $j = 0, 1, \dots, p$ have non-empty common intersection.

It is not hard to show that the union of the α -cells $\alpha(z, R/2)$ is equal to the union of the balls $B(z, R/2)$, so by the Čech theorem there is a homotopy equivalence $A(Z, R) \simeq \check{\text{Cech}}(Z, R)$. However the α -complex uses considerably fewer cells; in fact it is always a subcomplex of the Delaunay triangulation.

Standard implementations of the α -shape complex are based on a global calculation of the full Delaunay triangulation of Z . Thus there is ‘curse of dimensionality’ with respect to the dimension D of the ambient space \mathbb{R}^D . However if the parameter R is confined to a range with a small upper bound, we expect that the computation can be significantly localised.

3 Landmark techniques

The combinatorial Delaunay triangulation and its predecessor, the Martinetz–Schulten complex [6], have a vertex set which is a subsample of the full data set. These vertices are referred to as *landmark points* and are generally assumed to be well-distributed over the data.

Both constructions are special cases of *witness complexes*, which are based on the idea that the remaining (non-landmark) data points can be used to determine the edges and higher-dimensional cells of the complex. A common rule is that the edge $[\ell_0\ell_1]$ between two landmark points is included in the complex iff there exists a data point whose two nearest neighbours in the landmark set are ℓ_0 and ℓ_1 . The full data set serves as an approximation to the unknown space X . When the landmark set is fixed but the total sample size N increases, the number of cells increases and the complex converges towards the ideal complex in which every point in X is allowed to be a witness. One needs to use witness rules that can be satisfied with non-zero probability, otherwise this convergence fails. The restricted Delaunay triangulation of Section 2.3 is just such an example where care is necessary.

In this section, we define the Martinetz–Schulten complex and the combinatorial Delaunay triangulation in the framework of witness complexes, and we discuss several of the attendant issues, including computational strategy and choice of landmark points. We finish the section by presenting a parametrised version suitable for defining persistent homology groups.

3.1 Witness complexes

Let D be an $n \times N$ matrix of non-negative entries, regarded as the matrix of distances between a set of n landmarks and N data points. We define the (strict) witness complex $W_\infty(D)$ as follows.

- Vertex set: $\{1, 2, \dots, n\}$ (corresponding to the set of landmarks)
- Definition: the edge $\sigma = [ab]$ belongs to $W_\infty(D)$ iff there exists a witness $i \in \{1, 2, \dots, N\}$ such that D_{ai}, D_{bi} are the two smallest entries in the i -th column of D , in some order.
- Definition (by induction in p): Suppose all the faces of the p -simplex $\sigma = [a_0a_1 \dots a_p]$ belong to $W_\infty(D)$. Then σ itself belongs to $W_\infty(D)$ if and only if there exists a witness $i \in \{1, 2, \dots, N\}$ such that $D_{a_0i}, D_{a_1i}, \dots, D_{a_pi}$ are the $p + 1$ smallest entries in the i -th column of D , in some order.

A convenient alternative definition uses Rips expansion for the higher skeleta. This gives a complex $W_1(D) \supseteq W_\infty(D)$ formally defined as follows.

- Definition: $W_1(D)$ has the same 1-skeleton as $W_\infty(D)$.
- Definition: the p -simplex $\sigma = [a_0 a_1 \dots a_p]$ belongs to $W_1(D)$ iff all of its edges belong to $W_1(D)$.

In other words, $W_1(D)$ is the largest simplicial complex having the same 1-skeleton as $W_\infty(D)$. In practice we seldom use $W_\infty(D)$ since its computation is fussier, and we write $W(D)$ to mean $W_1(D)$.

3.2 The Martinetz–Schulten and CDT complexes

The essential idea for the Martinetz–Schulten complex was introduced in [6], in the context of a dynamically adaptive learning algorithm for representing a point cloud data set by a network of weighted edges. Leaving aside the dynamic learning, the construction by itself gives a useful approximation to a restricted Delaunay triangulation for the space X .

Suppose Z is a set of N points in Euclidean space (or indeed any metric space) and $L \subset Z$ is a designated set of n landmark points. Let D be the $n \times N$ matrix of distances between landmark points (labelled $1, 2, \dots, n$) and data points (labelled $1, 2, \dots, N$). Martinetz–Schulten complexes are defined by:

$$MS_\infty(L, Z) = W_\infty(D), \quad MS_1(L, Z) = W_1(D)$$

In practice it is more convenient to use $MS_1(L, Z)$; we write $MS(L, Z) = MS_1(L, Z)$ for simplicity.

The combinatorial Delaunay triangulation is an adaptation of the Martinetz–Schulten complex which has greater tolerance of nonlinearity and curvature in the data. The key idea, borrowed from the Isomap algorithm of Tenenbaum et al. [1], is to use intrinsic geodesic distance for the entries of D .

CDT depends on the same input data as MS—a set Z of data points in a metric space and a designated subset $L \subset Z$ of landmark points—and additionally a neighbourhood size parameter, which is a positive integer k or real number $\epsilon > 0$. Either way, this determines a *neighbourhood graph* G on the vertex set Z as follows: $[xy]$ is an edge of G iff y is one of the k nearest neighbours of x and vice versa; or $[xy]$ is an edge iff $|x - y| < \epsilon$.

The metric distance $d(x, y)$ is assigned as the weight of each edge $[xy]$. If G is connected, this defines a new metric by considering connecting paths of least weight (‘shortest paths’) in the graph. Let D_G denote the $n \times N$ matrix of shortest-path distances between landmark points and data points. Then we define:

$$CDT_\infty(L, Z; G) = W_\infty(D_G), \quad CDT_1(L, Z; G) = W_1(D_G)$$

Again we more commonly use the second definition, so we write $\text{CDT}(L, Z; G) = \text{CDT}_1(L, Z; G)$ for convenience.

Remark. The neighbourhood graph and shortest-paths computation introduce a degree of tolerance to nonlinear transformations which preserve the local topology of the data while distorting the extrinsic global geometry [1].

3.3 Strong and weak witnesses

The strict witness complex $W_\infty(D)$ can be motivated by comparing it with the the Delaunay triangulation in Euclidean space. A theorem is necessary to make the motivation complete.

Suppose $L \subset \mathbb{R}^D$ is a collection of points. Recall that the Delaunay triangulation $\text{Del}(L)$ contains the p -simplex $\sigma = [\ell_0 \ell_1 \dots \ell_p]$ precisely when there exists a point $x \in \mathbb{R}^D$ such that x is equidistant from the points $\ell_0, \ell_1, \dots, \ell_p$ and has no nearer neighbour in L . We call x a *strong witness* to the existence of σ , with respect to L .

When the set of allowed witnesses is discrete, there is no point looking for strong witnesses because they exist with probability 0. We say that $x \in \mathbb{R}^D$ is a *weak witness* for σ with respect to L iff $|x - \ell_i| \leq |x - \ell|$ for all $i = 0, 1, \dots, p$ and $\ell \in L \setminus \{\ell_0, \ell_1, \dots, \ell_p\}$; in other words, if the $p + 1$ nearest neighbours of x in L are $\ell_0, \ell_1, \dots, \ell_p$ (in a sense that tolerates equality). Our definition of $W_\infty(D)$ can be formulated in terms of weak witnesses: σ is a p -simplex of $W_\infty(D)$ iff it has a weak witness and all of its cells have weak witnesses.

Theorem 1 *Suppose $L \subset \mathbb{R}^D$ is a finite collection of points, and $\ell_0, \ell_1, \dots, \ell_p \in L$. Then $\sigma = [\ell_0 \ell_1 \dots \ell_p]$ has a strong witness with respect to L iff σ and all its cells have weak witnesses with respect to L .*

In the light of this theorem, the definition of $W_\infty(D)$ seems quite natural. The case $p = 1$ was discussed by Martinetz and Schulten in [6], justifying the definition of the graph which forms the 1-skeleton of the complex $\text{MS}_\infty(L, Z)$. A proof of the full result is given in [16]. We note that the argument is quite general, and equally valid in hyperbolic geometry.

3.4 Computation of MS and CDT

The computation of Betti numbers using either of the two landmark techniques can be broken into four stages.

- (1) Compute the $n \times N$ matrix of landmark-to-point distances.
- (2) Determine the 1-skeleton of the complex, as a list of vertex-pairs.
- (3) Expand to a $(j + 1)$ -complex by determining all compatible simplices of dimension up to $j + 1$.
- (4) Compute Betti numbers b^i , where $0 \leq i \leq j$.

For Martinetz–Schulten, Step 1 is a straightforward $O(dnN)$ calculation for Euclidean distances in \mathbb{R}^d . Step 2 involves identifying the indices of the smallest two entries in each column (that is, the closest two landmarks for each data point), and forming a list of unique, unordered index pairs. This is an $O(nN + n^2) = O(nN)$ operation.

For CDT, the lazy option is to use Dijkstra’s algorithm to calculate the single-source shortest-path distances for each landmark in turn, at a total cost of $O(knN \log N)$, where k is the degree of the neighbourhood graph. This dominates the $O(nN)$ cost of Step 2.

A more efficient but cumbersome algorithm combines both Step 1 and Step 2 into a single ‘parallel Dijkstra’ calculation similar to what is described in [9]. The idea is to compute only those distances $d(\ell, x)$ where ℓ is one of the two nearest landmark points to x . We set up a heap H in which each entry corresponds to a data point x , an associated landmark label ℓ , and a distance d provisionally regarded as $d(\ell, x)$. Any point x can occur more than once with different landmark labels. The initial configuration has n landmark points on the heap, at distance 0, labelled by themselves. In each iteration the closest point is pulled off the heap and its graph neighbours are added to the heap with the same landmark label and appropriate distances. Once a data point has been pulled off the heap twice (labelled by two different landmark points), all of its remaining copies are removed from consideration. This achieves the goal of avoiding much unnecessary calculation. We leave out the implementation details, since the simple approach is usually adequate when n is small or N is not enormous.

Step 3 is built into the Plex software library, a collection of MATLAB routines developed and used by the authors for handling simplicial complexes. We represent the 1-skeleton of the complex as a list of index pairs $A = (a_0, a_1)$ with $a_0 < a_1$, and the i -skeleton as a list of tuples $B = (b_0, b_1, \dots, b_i)$ with $b_0 < b_1 < \dots < b_i$. We then search for pairs A, B with $b_i = a_0$ and then write down the $(i + 1)$ -simplex (b_0, \dots, b_i, a_1) if the tuple $(b_0, \dots, b_{i-1}, a_1)$ belongs to the i -skeleton. Thus we inductively construct the skeleta of the complex.

Step 4 can be carried out in many different ways, once the complex has been constructed as far as the $(j + 1)$ -skeleton. For example see [11].

3.5 Choosing the landmark points

Each landmark point effectively determines a Voronoi cell in the graph metric. In an ideal reconstruction, these cells correspond to convex, convexly intersecting regions of the underlying space X . The average cell size in the discrete approximation is given by the ratio $\lambda = |Z|/|L|$. If the landmark points are well spread out and do not bunch together, then each Voronoi cell will have approximately λ points in it. It is well to be aware of the value of λ in practice.

We suggest obtaining the landmark set in one of two ways: by random choice, or by *sequential maxmin*. Both methods seem to give reasonable results for the MS and CDT algorithms. Sequential maxmin is the following inductive procedure:

- (1) Select $\ell_1 \in Z$ randomly.
- (2) When $\ell_1, \ell_2, \dots, \ell_{i-1}$ have been chosen, select $\ell_i \in Z \setminus \{\ell_1, \ell_2, \dots, \ell_{i-1}\}$ to be the data point which maximises the function:

$$z \mapsto \min\{d(z, \ell_1), d(z, \ell_2), \dots, d(z, \ell_{i-1})\}$$

- (3) Repeat Step 2 until a set $L = \{\ell_1, \ell_2, \dots, \ell_n\}$ of n landmark points has been chosen.

The metric d is taken to be the Euclidean metric, or in the case of CDT it may also be the shortest-paths metric. This can be implemented with no loss of efficiency: Dijkstra's algorithm computes each function $z \mapsto d(z, \ell)$ separately, so these can be calculated for $\ell_1, \ell_2, \dots, \ell_n$ sequentially as needed. Note that the parallel Dijkstra algorithm described earlier cannot be applied in this context, since it requires prior knowledge of all the landmark points. In any case, both metrics are the same in local neighbourhoods so there seems to be little reason to prefer one over the other; at least when λ and the neighbourhood size k are of the same order of magnitude.

Other than simplicity, the main advantage of random choice is that it is statistically representative. The landmark points will typically be located in high-density regions of the data, rather than in low-density background noise. This is especially true when the number of landmarks n is small. The advantage of sequential maxmin is that the landmarks are guaranteed to be well-separated. On the other hand, the rigid constraints frequently lead to the selection of outliers as landmarks. Finally, theoretical analysis is more difficult when the landmarks are chosen by a deterministic inductive procedure, as opposed to randomly. However, experience suggests that sequential maxmin is often worth the extra effort.

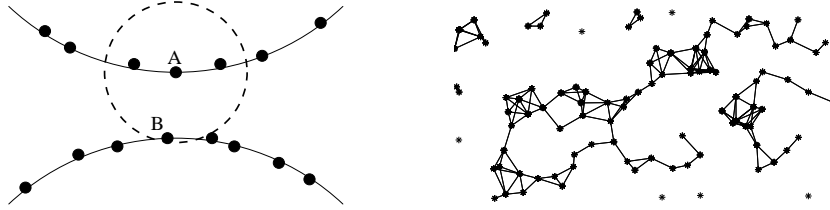


Fig. 2. Choosing the wrong neighbourhood size: too large (left), too small (right)

3.6 *Setting the neighbourhood size*

There are two conflicting requirements when choosing the neighbourhood size parameter, k or ϵ .

- (1) It must be small enough to avoid ‘false neighbours’ coming from quite separate regions of the space X .
- (2) It must be large enough so that the neighbourhood graph has enough edges to represent a viable 1-skeleton for X .

These twin dangers are illustrated in Figure 2. The panel on the left shows part of a one-dimensional data set with two closely approaching branches. The neighbourhood size chosen for point A is just large enough to include point B belonging to the other branch, so there will be a misleading extra edge in the graph. The right-hand panel shows the neighbourhood graph on a set of points sampled randomly in a rectangle. Because the neighbourhood size is too small, the graph is too sparse to represent the two-dimensional structure of the data correctly.

A generic d -dimensional lattice has degree $d(d+1)$ at each vertex, so for a d -dimensional data set $k \gg d(d+1)$ is a reasonable estimate for “large enough”. On the other hand, if N points are sampled randomly from a fixed space X , then the upper bound for “small enough” takes the form $k < O(N)$. When the sample size N is large enough both requirements can be met simultaneously. For further discussion we refer the reader to the literature of non-linear dimensionality reduction, where this is a topic of central importance [1,17]. We note that k nearest neighbours is generally preferable to ϵ , being a scale-invariant parameter.

3.7 *Persistent homology*

There is a natural way of extending witness-based constructions to give nested families of simplicial complexes, suitable for persistent homology.

The paradigm is Edelsbrunner’s family [15] of α -shape complexes $A(Z, R)$ de-

defined for $R \in [0, \infty]$. The smallest complex $A(Z, 0)$ is a discrete collection of points, and the largest complex $A(Z, \infty)$ is a complete simplex on the vertex set Z . The appearance, survival and disappearance of homology classes, as R varies through intermediate values, provides detailed topological information that is statistically more robust than the Betti numbers of the complex for any single value of R . Thus we seek similar nested families based on the Martinetz–Schulten and combinatorial Delaunay complexes.

In this subsection we construct a family of simplicial complexes $W(D; R, \nu) = W_1(D; R, \nu)$, where $R \in [0, \infty]$ and ν is a non-negative integer, extending the definition of the witness complex $W(D) = W_1(D)$. The family is doubly nested in R and ν , so that if $R \leq R'$ and $\nu \leq \nu'$ then $W(D; R, \nu) \subseteq W(D; R', \nu')$. For a fixed value of ν one gets a nested family of complexes indexed by a single parameter R , analogous to the α -shape family. The special cases $\nu = 0, 1, 2$ are of particular relevance. Here is the definition of $W(D; R, \nu)$.

- Vertex set: $\{1, 2, \dots, n\}$
- If $\nu = 0$, then for $i = 1, 2, \dots, N$ set $m_i = 0$.
- If $\nu > 0$, then for $i = 1, 2, \dots, N$ set m_i equal to the ν -th smallest entry of the i -th column of D .
- Definition: the edge $\sigma = [ab]$ belongs to $W(D; R, \nu)$ iff there exists a witness $i \in \{1, 2, \dots, N\}$ such that $\max(D_{ai}, D_{bi}) \leq R + m_i$.
- Definition: the p -simplex $\sigma = [a_0 a_1 \dots a_p]$ belongs to $W(D; R, \nu)$ iff all its edges belong to $W(D; R, \nu)$; equivalently iff there exists a witness $i \in \{1, 2, \dots, N\}$ such that $\max(D_{a_0 i}, D_{a_1 i}, \dots, D_{a_p i}) \leq R + m_i$.

Note that $W(D; 0, 0)$ and $W(D; 0, 1)$ are both equal to the discrete complex on n vertices, and $W(D; 0, 2) = W(D)$. Also $W(D; \infty, \nu)$ is equal to the complete simplex on n vertices, for all values of ν . Thus the 1-parameter families for $\nu = 0, 1$ are closely analogous to $A(Z, R)$, since they exhibit the same behaviour at $R = 0$ and $R = \infty$. On the other hand the $\nu = 2$ family at $R = 0$ is already equal to the good approximation $W(D)$, so it may be preferable for that reason.

The persistent homology groups over an interval $R \in [0, r]$ can be computed using algorithms in [11]. The preprocessing task is to generate a list of simplices (up to dimension $p + 1$ for p -dimensional homology). For each simplex σ , one needs to identify its faces and determine its *time of appearance*, which is the smallest value $R = R_\sigma$ for which $\sigma \in W(D, R)$. By definition, $R_\sigma = \min\{R_\tau : \tau \text{ is an edge of } \sigma\}$ and so we break up the task as follows:

- (1) Compute the $n \times n$ matrix E with off-diagonal entries $E_{ij} = R_{[ij]}$, which records the time of appearance of each edge.
- (2) Generate a list of simplices which appear by time r .
- (3) Compute the appearance time of each simplex as the maximum of the

appearance times of its edges.

Step 1 can be expressed algebraically as a kind of ‘minmax’ matrix product: $E = D \odot D^*$. Here \odot represents the operation

$$[A \odot B]_{ij} = \min_k \max(A_{ik}, B_{kj})$$

which presents no difficulty in implementation. For Step 2, a list of edges which appear by time r is used to generate higher-dimensional cells by Rips expansion, as described in Step 3 of Section 3.4. The appearance time of each new cell can be computed when it is generated, for instance as the maximum of the already computed appearance times of (b_0, b_1, \dots, b_i) , $(b_0, b_1, \dots, b_{i-1}, a_1)$ and (a_0, a_1) .

Remark. It is also possible to define parametrised families of complexes based on the strict witness complex $W_\infty(D)$. However, the natural definitions seem to lead to rather cumbersome computations, so we omit discussing them here.

3.8 Mathematical motivation

Simplicial complex approximations are well understood if they can be interpreted as the nerve of a covering of a space [13]. As discussed earlier, the Čech complex and the related α -shape complex are nerves of different coverings of a union of balls $Z_R = \bigcup B(z, R/2)$. Since the covering sets are convex, the Čech theorem implies that the resulting complexes recover the homotopy type of Z_R . Thus the question of successful recovery reduces to understanding when $Z_R \simeq X$.

We now give a covering-based interpretation of the family of simplicial complexes $W(D; R, 1)$ which partially motivates the construction in the preceding section. The starting point is a metric space X , together with a collection L of n landmark points and a much larger collection Z of data points. The metric on X defines a continuous function $\varphi : X \rightarrow \mathbb{R}^n$ by the formula:

$$\varphi(x) = [d(x, \ell_1), d(x, \ell_2), \dots, d(x, \ell_n)]$$

If there are enough landmark points then this will generally be an embedding of X inside \mathbb{R}^n ; in which case, knowledge of this function is therefore enough to recover the topology of X .

However we now choose to view \mathbb{R}^n through a crude topological lens, by ex-

pressing it as a union of spaces \mathbb{R}_i^n defined:

$$\mathbb{R}_i^n = \{x \in \mathbb{R}^n : x_i = \min(x_1, x_2, \dots, x_n)\}$$

This can be pulled back to a covering of X by sets $\Phi_i = \varphi^{-1}(\mathbb{R}_i^n)$. If the landmark points are sufficiently finely distributed, it can be shown under reasonable assumptions that the nerve of the covering $\{\Phi_i\}$ recovers the homotopy type of X .

In practice we have a discrete sample Z instead of the continuous space X . We can still define a covering of Z by sets Φ_i , but with probability 1 the pairwise intersections $\Phi_i \cap \Phi_j$ will all be empty, so the nerve complex will be a discrete set of vertices with no edges or higher cells. In order to remedy this problem, we thicken the sets \mathbb{R}_i^n so that the overlaps will have positive measure. For example we can define $\mathbb{R}_i^n(R) = \{x \in \mathbb{R}^n : \text{dist}(x, \mathbb{R}_i^n) \leq R/2\}$. The new parameter R naturally produces nested complexes and hence gives rise to persistent homology groups. If distance function on \mathbb{R}^n used in the definition is the ℓ_∞ -norm, $\|x\|_\infty = \max_i(|x_i|)$, then the nerve of the corresponding covering of Z is precisely equal to $W(D; R, 1)$.

4 Examples

4.1 Points on a sphere $S^2 \subset \mathbb{R}^3$

We applied several simplicial complex approximations to the task of recovering the correct Betti numbers of the sphere $S^2 \subset \mathbb{R}^3$. In each trial, 500 points were generated uniformly randomly on the unit sphere, and 12 landmark points were chosen randomly or by sequential maxmin. Seven constructions were applied to each of these data sets: the Rips complex (on the landmark points alone), the Martinetz–Schulten complex with $\nu = 0, 1, 2$, and the CDT complex (using neighbourhood size $k = 12$) with $\nu = 0, 1, 2$. The calculation was organised so as to determine the Betti numbers b^0 , b^1 and b^2 , for all possible values of the moveable parameter R . However, no persistent homology groups were computed.

A selection of statistics from these experiments is presented in Figures 3 and 4. We ran 100 trials for each method of generating the landmarks. For each trial and each construction, we determined four constants R_0 , R_1 , K_0 and K_1 . These are determined as follows. R_0 and R_1 are chosen so that $(b^0, b^1, b^2) = (1, 0, 1)$ for $R \in [R_0, R_1)$, agreeing with the 2-sphere; but $(b^0, b^1, b^2) \neq (1, 0, 1)$ for $R \geq R_1$ and for $R = R_0 - \epsilon$. In other words, $[R_0, R_1)$ is the rightmost contiguous interval over which the homology of S^2 is correctly recovered. At $R = K_0$ the

| 12 LANDMARK POINTS CHOSEN RANDOMLY | | | | | | | |
|---|-------|--------------------|-----------|-----------|------------------------|-----------|-----------|
| | Rips | Martinetz-Schulten | | | Combinatorial Delaunay | | |
| | | $\nu = 0$ | $\nu = 1$ | $\nu = 2$ | $\nu = 0$ | $\nu = 1$ | $\nu = 2$ |
| % success | 54 | 51 | 99 | 99 | 53 | 100 | 97 |
| <i>when a successful reconstruction exists for some R:</i> | | | | | | | |
| median relative dominance | 0.038 | 0.059 | 0.620 | 0.808 | 0.062 | 0.600 | 0.798 |
| median absolute dominance | 0.034 | 0.047 | 0.347 | 0.163 | 0.046 | 0.318 | 0.152 |
| median number of cells | 208 | 199 | 86 | 94 | 208 | 92 | 92 |

Fig. 3. Reconstructing the sphere $S^2 \subset \mathbb{R}^3$: landmarks chosen randomly.

| 12 LANDMARK POINTS CHOSEN BY SEQUENTIAL MAXMIN | | | | | | | |
|---|-------|--------------------|-----------|-----------|------------------------|-----------|-----------|
| | Rips | Martinetz-Schulten | | | Combinatorial Delaunay | | |
| | | $\nu = 0$ | $\nu = 1$ | $\nu = 2$ | $\nu = 0$ | $\nu = 1$ | $\nu = 2$ |
| % success | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| <i>when a successful reconstruction exists for some R:</i> | | | | | | | |
| median relative dominance | 0.184 | 0.215 | 0.752 | 0.924 | 0.216 | 0.744 | 0.922 |
| median absolute dominance | 0.161 | 0.162 | 0.519 | 0.252 | 0.153 | 0.466 | 0.209 |
| median number of cells | 74 | 78 | 66 | 79 | 82 | 66 | 80 |

Fig. 4. Reconstructing the sphere $S^2 \subset \mathbb{R}^3$: landmarks chosen by sequential maxmin.

Betti profile changes permanently to $(1, 0, 0)$, indicating that the data have coalesced into a single contractible blob. Finally, $R = K_1$ marks the time when the complex becomes the complete simplex on 12 vertices; all possible cells have been included.

In the tables, “% success” indicates the number of trials (out of 100) where the homology of S^2 is correctly recovered for some interval of values of R , no matter how small. For each successful trial, *relative dominance* and *absolute dominance* are defined to be $(R_1 - R_0)/K_0$ and $(R_1 - R_0)/K_1$ respectively. Relative dominance compares the lengths of the successful interval $[R_0, R_1)$ and the interval $[0, K_0]$ of homological activity. Absolute dominance compares the successful interval interval $[0, K_1]$ of *cellular* activity. If either of these quantities is large, this indicates that the Betti profile $(1, 0, 1)$ can be taken seriously *a priori*, and not just because we know the correct answer.

The last three rows of the tables give median values of these statistics, and of

the total number of cells (up to dimension 3) at $R = R_0$. The median is taken over successful trials only. For unsuccessful trials R_0 and R_1 are not defined; although both dominances may be taken to be 0 in those cases.

We make several observations.

- (1) In these tests, there is little to choose between MS and CDT. This is to be expected since there is a monotonic relationship between Euclidean distance and geodesic distance on a sphere.
- (2) The $\nu = 0$ cases of MS and CDT behave very similarly to Rips. Again, there is a monotonic relationship between the Euclidean length of an edge, and the value of R for which a witness to that edge is likely to exist; since 500 points cover the sphere finely relative to the 12 landmark points.
- (3) The $\nu = 1, 2$ cases of MS and CDT give results which are consistently, strikingly better than Rips and $\nu = 0$, particularly when the landmarks are chosen randomly. Fewer cells are needed and the successful intervals have greater dominances.
- (4) In every case with sequential maxmin, there is a range of R for which the correct Betti numbers are recovered. This is surely a result of the strong constraints on landmark selection. In terms of the dominance, there is still a clear advantage in using the witness-based techniques with $\nu = 1, 2$.
- (5) The $\nu = 2$ cases have extremely high relative dominances, but much lower absolute dominances. Which should be taken more seriously? The underlying cause of the difference is that K_0/K_1 is small; in other words homological activity dies down long before cellular activity, as R increases. It is tempting to think of this as a good thing, but a deeper understanding is called for.

The overall message is reasonably clear; which is that witness complexes give topological approximations which are more reliable, use fewer cells, and are statistically more defensible than Rips complexes. At the same time there are significant benefits to choosing landmark points by sequential maxmin rather than randomly.

4.2 *Natural image statistics*

We applied the same simplicial approximation techniques to a point cloud data set derived from natural image data, provided by David Mumford. The data set can be viewed as a probability distribution over the sphere $S^7 \subset \mathbb{R}^8$, which is known to be highly concentrated over a particular region topologically equivalent to an annulus $S^1 \times [0, 1]$. A natural question is whether this annulus can be detected using topological approximation techniques. It turns out to

be possible, using b^1 persistent homology.

The data set in question consists of about 4.2×10^6 normalised high-contrast 3×3 optical image patches, extracted from van Hateren’s still image collection [18] and described in Lee, Pederson and Mumford [19]. The normalisation consists of subtracting the mean and rescaling to unit norm, which reduces each 9-dimensional data vector to a point in the unit sphere $S^7 \subset \mathbb{R}^8$ after a change of coordinates. To bring the data down to a manageable size, we randomly sampled 5×10^4 points from this collection and regarded that subset as our primary source.

Edge features in a natural image can be idealised as perfectly straight boundaries between two homogeneous regions of different brightness levels. The family of idealised edges has the topology of an annulus, being parametrised by angle and distance from the image centre; and this translates to a topological annulus in the space S^7 of normalised patches. Since edges are a prominent and common feature of natural images, one expects to find a high concentration of data points on or near this annulus. See [19] for a detailed discussion.

A direct application of simplicial complex approximation to the 5×10^4 data points is destined to fail, since there are points distributed all over the sphere and not just in the high-density regions. To extract a high-density sample, we thresholded on a simple density function

$$\rho_k(x) = |x - x_k|, \quad \text{where } x_k \text{ is the } k\text{-th nearest neighbour of } x,$$

using $k = 125$. Our subsample was the 25% cut of points with smallest ρ_{125} , a set of size $N = 12500$.

We then ran each of the seven reconstruction algorithms investigated in Section 4.1, randomly choosing $n = 50$ landmark points. In each case we computed the full persistent homology for the Betti number b^1 , in the hope of discovering a 1-dimensional homology cycle of long persistence; this would be the cycle carried by the annulus in its angle variable. A typical set of results is shown in Figures 5, 6 and 7, in the form of persistence interval graphs [12]. Each interval represents the appearance of a homology cycle, its lifetime, and ultimate disappearance.

The results are consistent with the general observations of Section 4.1. The Rips complex and the $\nu = 0$ cases of Martinetz–Schulten and CDT all give similar results, with a small number of ‘noisy’ homology cycles with short lifespan, together with one cycle which clearly survives over an extended interval. The results for $\nu = 1$ and $\nu = 2$ are even clearer, in the sense that the persistent cycle appears right at the beginning of the process. We note the difference between $\nu = 1$, where we observe a large number of noise cycles with extremely short persistence, and $\nu = 2$, which is shockingly clean.

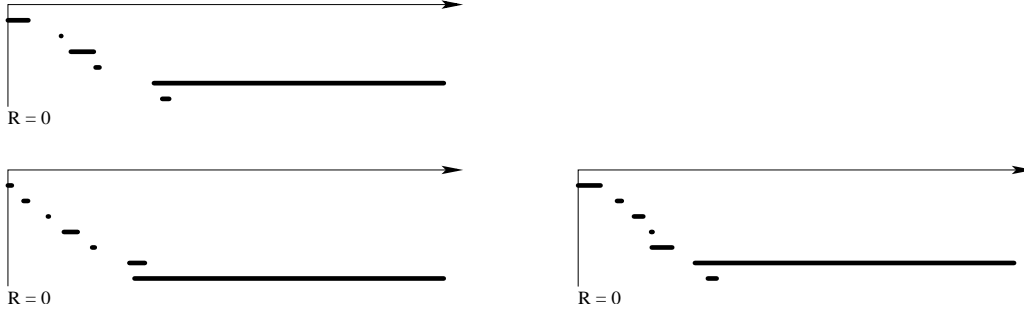


Fig. 5. Persistence intervals in b^1 : Rips (top left); Martinetz–Schulten, $\nu = 0$ (bottom left); combinatorial Delaunay, $\nu = 0$ (bottom right).

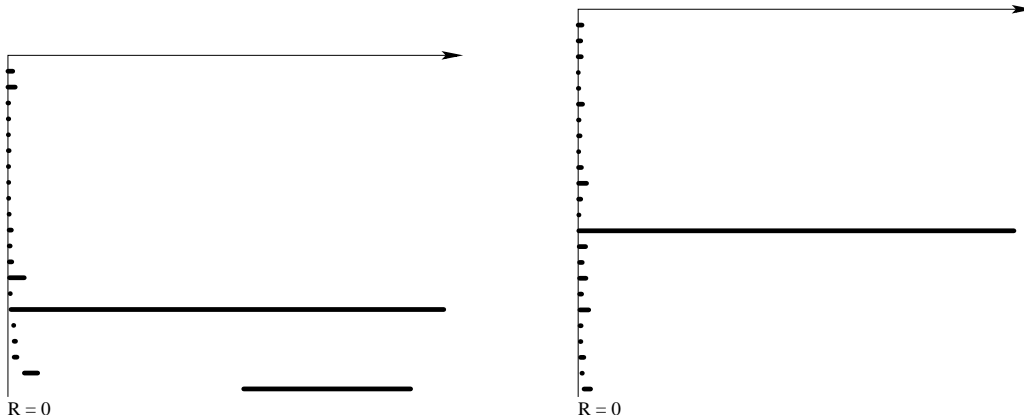


Fig. 6. Persistence intervals in b^1 : Martinetz–Schulten, $\nu = 1$ (left); combinatorial Delaunay, $\nu = 1$ (right).



Fig. 7. Persistence intervals in b^1 : Martinetz–Schulten, $\nu = 2$ (left); combinatorial Delaunay, $\nu = 2$ (right).

The Martinetz–Schulten algorithm gives slightly different results in this trial. There is a second 1-cycle, of medium persistence, in the $\nu = 1$ case; for $\nu = 2$ this has grown into a second cycle of essentially equal importance to the principal one. On repeated trials with 50 randomly-chosen landmarks this discrepancy seems to occur about half the time, with the expected single interval occurring in the remaining half of cases. This leads to the question of what this phenomenon may be saying about the geometry of the data. This is not a question we know how to answer at this time.

Repeating the experiment with $n = 10$ randomly-chosen landmarks, it turns out that all seven trials give extremely clean results with one long interval and virtually no noise. On the other hand it can be argued that the results with $n = 50$ are more meaningful, since they allow the potential detection of more complicated topological behaviour; whereas with only 10 points the possibilities are limited.

Concluding Remarks

Modern statistical analysis increasingly calls for the use of nonlinear techniques, capable of resolving the underlying structure of a data set. The modern theory of nonlinear dimensionality reduction (NLDR) gives several examples of such techniques. The emphasis is on identifying the essential parameters of the data as coordinate functions of a low-dimensional embedding. This presupposes that such an embedding exists; which in turn restricts the use of such techniques to data manifolds with comparatively simple topology. On the other hand, it is clear that many naturally occurring data sets exhibit non-trivial topology. One distant hope is to use topological information to reduce complicated data manifolds to simpler pieces which might then be amenable to NLDR techniques. The reliable estimation of topological invariants will be an essential part of such a scheme.

In other areas of research, there is a growing body of algorithms which exploit topological information carried in point-cloud data. It is becoming clear that rapid topological profiling is an essential tool in these developments. As we have tried to suggest in the examples of Section 4, it is not simply enough to pick a single algorithm and hope that it will work most or some of the time. Instead, one should have a battery of techniques available, and a detailed understanding of their behaviour under different circumstances and varying parameter settings. If a particular technique appears to give an anomalous answer, what possible geometric configurations can explain it?

At present this appears to be a daunting task, given the wealth of possible examples and techniques. Part of the challenge is to develop explanatory and predictive theoretical tools for analysing simplicial complex approximations, including those with *ad hoc* definitions that may be theory-unfriendly. There is also call for a statistics of Betti numbers and of persistence interval graphs, so that quantitative statements can be made with known levels of confidence. We hope that this paper represents a small step towards fulfilling this ambitious programme.

Acknowledgements

We gratefully acknowledge the support of the NSF, through grant DMS-0101364. This work was carried out at the Department of Mathematics, Stanford University.

We wish to thank several individuals: Afra Zomorodian, for numerous discussions and for use of his code for persistent homology calculations; David

Mumford, for making available his database of 3×3 patches; and Debashis Paul, for his considerable assistance with preliminary investigations of the Mumford data. The second author also wishes to thank Josh Tenenbaum, for generously sharing his insights into nonlinear statistics and for emphasizing the importance of landmark-based techniques; and Carrie Grimes, for several helpful comments on the experimental studies in this paper.

References

- [1] J. B. Tenenbaum, V. de Silva, J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (2000) 2319–2323.
- [2] A. B. Lee, K. S. Pedersen, D. Mumford, The nonlinear statistics of high-contrast patches in natural images, Tech. Rep. APPTS #01-3, Division of Applied Mathematics Brown University (December 2001).
URL <http://www.dam.brown.edu/ptg/publications.html>
- [3] W. Massey, *A Basic Course in Algebraic Topology*, Springer-Verlag, New York, 1991.
- [4] E. Carlsson, G. Carlsson, V. de Silva, An algebraic topological method for feature identification, [submitted].
- [5] G. Carlsson, A. Collins, L. Guibas, A. Zomorodian, Persistent homology and shape description: I, [in preparation].
- [6] T. Martinez, K. Schulten, Topology representing networks, *Neural Networks* 7 (3) (1994) 507–522.
- [7] W. S. Torgerson, *Theory and Methods of Scaling*, Wiley, New York, 1958.
- [8] T. F. Cox, M. A. A. Cox, *Multidimensional Scaling*, Chapman & Hall, London, 1994.
- [9] M. Erwig, The graph Voronoi diagram with applications, *Networks* 36 (3) (2000) 156–163.
- [10] N. Amenta, S. Choi, T. K. Dey, N. Leekha, A simple algorithm for homeomorphic surface reconstruction, *International Journal of Computational Geometry and Applications* 12 (1-2) (2002) 125–141.
- [11] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, in: *IEEE Symposium on Foundations of Computer Science*, 2000, pp. 454–463.
- [12] A. Zomorodian, G. Carlsson, Computing topological persistence, [preprint] (2003).
- [13] E. H. Spanier, *Algebraic Topology*, McGraw-Hill Book Co., 1966.

- [14] V. de Silva, Simplicial complexes in manifold learning, [in preparation] (2003).
- [15] H. Edelsbrunner, The union of balls and its dual shape, *Discrete & Computational Geometry* 13 (3-4) (1995) 415–440.
- [16] V. de Silva, A weak definition of Delaunay triangulation, [preprint] (2003).
URL <http://math.stanford.edu/comptop/preprints/>
- [17] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (2000) 2323–2326.
- [18] J. H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex, *Proceedings of the Royal Society of London B* 265 (1998) 359–366.
- [19] A. B. Lee, K. S. Pedersen, D. Mumford, The nonlinear statistics of high-contrast patches in natural images, *International Journal of Computer Vision* 54 (1-3) (2003) 83–103.