Journal of Multivariate Analysis 101 (2010) 2184–2199

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Exploring uses of persistent homology for statistical analysis of landmark-based shape data

Jennifer Gamble*, Giseon Heo

Statistical Consulting Centre, Department of Dentistry, University of Alberta, Edmonton, Canada, T6G 2N8

ARTICLE INFO

Article history:
Received 16 September 2009
Available online 4 June 2010

AMS subject classifications:
62-07
62P10
55U15
55U10
62H25

Keywords:
Persistent homology
Wasserstein distance
Landmark-based data
Shape analysis
Persistence diagrams
Multidimensional scaling

ABSTRACT

A method for the use of persistent homology in the statistical analysis of landmark-based shape data is given. Three-dimensional landmark configurations are used as input for separate filtrations, persistent homology is performed, and persistence diagrams are obtained. Groups of configurations are compared using distances between persistence diagrams combined with dimensionality reduction methods. A three-dimensional landmark-based data set is used from a longitudinal orthodontic study, and the persistent homology method is able to distinguish clinically relevant treatment effects. Comparisons are made with the traditional landmark-based statistical shape analysis methods of Dryden and Mardia, and Euclidean Distance Matrix Analysis.

1. Introduction

To describe the dimensions of objects or images, data sets may be encoded as a set of labeled points in two or three dimensions. In morphometrics literature, such points are called *landmarks*. These landmarks serve as reference points for a partial geometric description of an object. If each image or object is represented as k landmarks in $\mathbb{R}^d$, this can be written as a $k \times d$ configuration matrix. Two landmark configurations are said to have the *same shape* if a rigid motion (translation or rotation) and rescaling of one will coincide with the other. If size information is also of interest, then it may be retained by omitting the rescaling.

When multiple objects or subjects are under consideration, the landmarks are assigned to each in a corresponding way. This is often done by an expert in the field, and the landmarks chosen are biologically or clinically relevant and appropriate to the objectives of the study. For example, the data set that will be used later in this paper is from an orthodontic study involving an upper jaw expansion treatment, and the landmarks used are well-defined upper jaw landmarks in the orthodontic literature, with their placement by an experienced orthodontist checked for reliability. The field of statistical shape analysis was developed to allow statements to be made about whether there are statistically significant differences in the mean shape between groups, based on the observation of some samples. If differences are observed, it is also desirable to localize and describe them. Two methods of traditional statistical shape analysis that will be applied in this paper are Euclidean Distance Matrix Analysis (EDMA) [17], and the methods discussed in the book *Statistical Shape Analysis* by Dryden

* Corresponding author.
E-mail addresses: jpgamble@ualberta.ca (J. Gamble), gheo@ualberta.ca (G. Heo).

0047-259X/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2010.04.016

and Mardia [9]. EDMA represents each landmark configuration as the set of all pairwise inter-landmark distances, which can be written as a $k(k-1)/2$-vector (where k is the number of landmarks). Dryden and Mardia's methods represent each landmark configuration as a single point in a high-dimensional shape space (or more specifically, a point in a tangent space approximation to the shape space). These methods will be described in more detail in Section 4. A thorough review, including other methods in statistical shape analysis for landmark data, is not presented here, but may be found in [15].

Persistent homology (see [11,22]) is a recently developed method of determining topological information from sets of points, shapes or functions. We will first give an outline of the basics behind persistent homology, and then propose a method for its use in the analysis of landmark-based shape data. Methods from persistent homology have been used in some practical applications (e.g. [13,19,7]), and have just recently begun to be applied to the analysis of medical image data [5]. The methods used in the present paper are based on established theory of metrics on the space of persistence diagrams (to be discussed in Section 3), applied for the first time with dimensionality reduction and statistical analysis. In [5], Chung et al. also used a metric to compare persistence diagrams, and similarities to the current paper will be discussed at the end of Section 2. Our method will be applied to an example data set from orthodontics in Section 5.1, and the results compared with those from the other shape analysis methods in Section 5.2.

2. Persistent homology (an outline)

Full development of the theory and methods behind persistent homology is not given in this brief outline, but is available in [11,22] or [21]. More recent surveys [10,12,4] include applications and future directions for the subject.

Algebraic topology is a large area of mathematics, which involves the classification of topological spaces using algebraic invariants. Here, when we speak of topological features of interest, this includes things like the number of connected components, and the number of holes. If working in greater than two dimensions, this extends to enclosed voids (such as the empty space inside of a sphere) in three dimensions, and higher-dimensional analogues. One method of classifying a space is by counting the number of topological features of each dimension. These counts are referred to as *Betti numbers*, with the nth Betti number, β_n, recording the number of n-dimensional topological features. The number of connected components are encoded then as β_0, the number of loops (or holes) as β_1, and the number of enclosed voids as β_2. An often cited resource as an introduction to this field is Hatcher's *Algebraic Topology* [14].

Persistent homology is a computational method developed to determine *topologically significant* features from a set of data points (often thought to represent a sample from some underlying space, such as a Riemannian manifold). The next two paragraphs give the notions of a simplicial complex, and an abstract simplicial complex, which are required for our later discussions.

A k-dimensional simplex (or a k-simplex) is the convex hull of $k+1$ affinely independent points $\{v_0, v_1, \ldots, v_k\}$, which are called the *vertices* of the simplex. For k between zero and three, a k-simplex is a point, line segment, triangle, or tetrahedra (respectively), as in Fig. 1 (left). If σ is a simplex, and a subset of its vertices are the vertices of another simplex τ, then τ is called a *face* of σ. A *simplicial complex* K is a finite set of simplices such that (a) if a simplex is in K, then all its faces are also simplices in K and (b) if two simplices are in K, then their intersection is either a face of both simplices, or is empty. Combinations of simplices which do not form simplicial complexes are shown in Fig. 1 (middle).

A combinatorial definition of a simplicial complex is also possible, where an *abstract simplicial complex* is a set together with a collection $\mathcal{S}$ of its subsets, where an element v of the set is considered a *vertex* if $\{v\}$ is in $\mathcal{S}$. Additionally, there is the property that for every $\sigma \in \mathcal{S}$, all subsets of σ are also in $\mathcal{S}$. For example, if the set is $\{a, b, c, d\}$ then one possible collection of subsets that forms an abstract simplicial complex is

$$\mathcal{S} = \{\{\emptyset\}, \{a\}, \{b\}, \{c\}, \{d\}, \{ab\}, \{bc\}, \{ca\}, \{cd\}, \{abc\}\}.$$

Note that every element in $\mathcal{S}$ also has all of its subsets in $\mathcal{S}$ (and the empty set is included since it is a subset of all sets). Every simplicial complex can be represented as an abstract simplicial complex such as this, and this combinatorial representation is useful for computational purposes (e.g. during the persistent homology algorithm). Conversely, every abstract simplicial complex has a *geometric realization* as a simplicial complex. Two geometric realizations of $\mathcal{S}$ are given in Fig. 1 (right). For a given abstract simplicial complex, a geometric realization does not necessarily exist in all dimensions, however. For example, a tetrahedron has no geometric realization in two dimensions. Two abstract simplicial complexes are said to be *isomorphic* if there is a bijection between their vertices which maps the sets of one onto the sets of the other.

One way to approximate the topology of a space based on a sample of points from it, is to build a simplicial complex by using the sample points as vertices, and connecting vertices by an edge whenever they are within some distance $\varepsilon > 0$ of each other. When three points are all pairwise connected, then the corresponding triangle is also part of the simplicial complex. Higher-dimensional simplices are part of the complex if all of their vertices are pairwise within ε. This definition forms a *Rips complex* (also known as a *Vietoris–Rips complex*). A true Rips complex can contain simplices of any dimension (up to $n-1$ for n sample points), but the complex may also be restricted to only include simplices up to a small dimension (say two or three). For the remainder of this paper, we will work with complexes restricted to two or three dimensions. Methods for determining the topology (in particular, the number of connected components, loops, voids, etc.) of simplicial complexes are well developed in the field of simplicial homology. Depending on which ε is chosen, the complex and its associated topology will change.

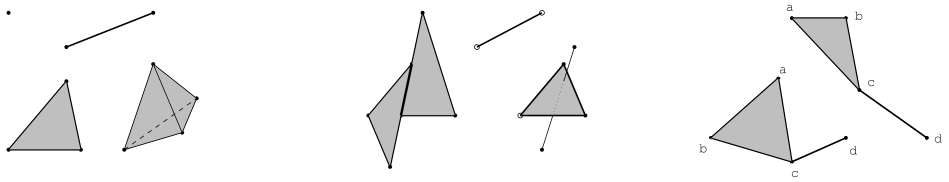

Fig. 1. Left: simplices of dimension zero (point), one (line), two (triangle) and three (tetrahedron). Middle: configurations of simplices which are not simplicial complexes because they either intersect over a region which is not a face of both simplices, or they contain simplices which have faces that are not part of the complex. Right: two simplicial complexes that are isomorphic to the abstract simplicial complex $\mathcal{S} = \{\{\emptyset\}, \{a\}, \{b\}, \{c\}, \{d\}, \{ab\}, \{bc\}, \{ca\}, \{cd\}, \{abc\}\}$.

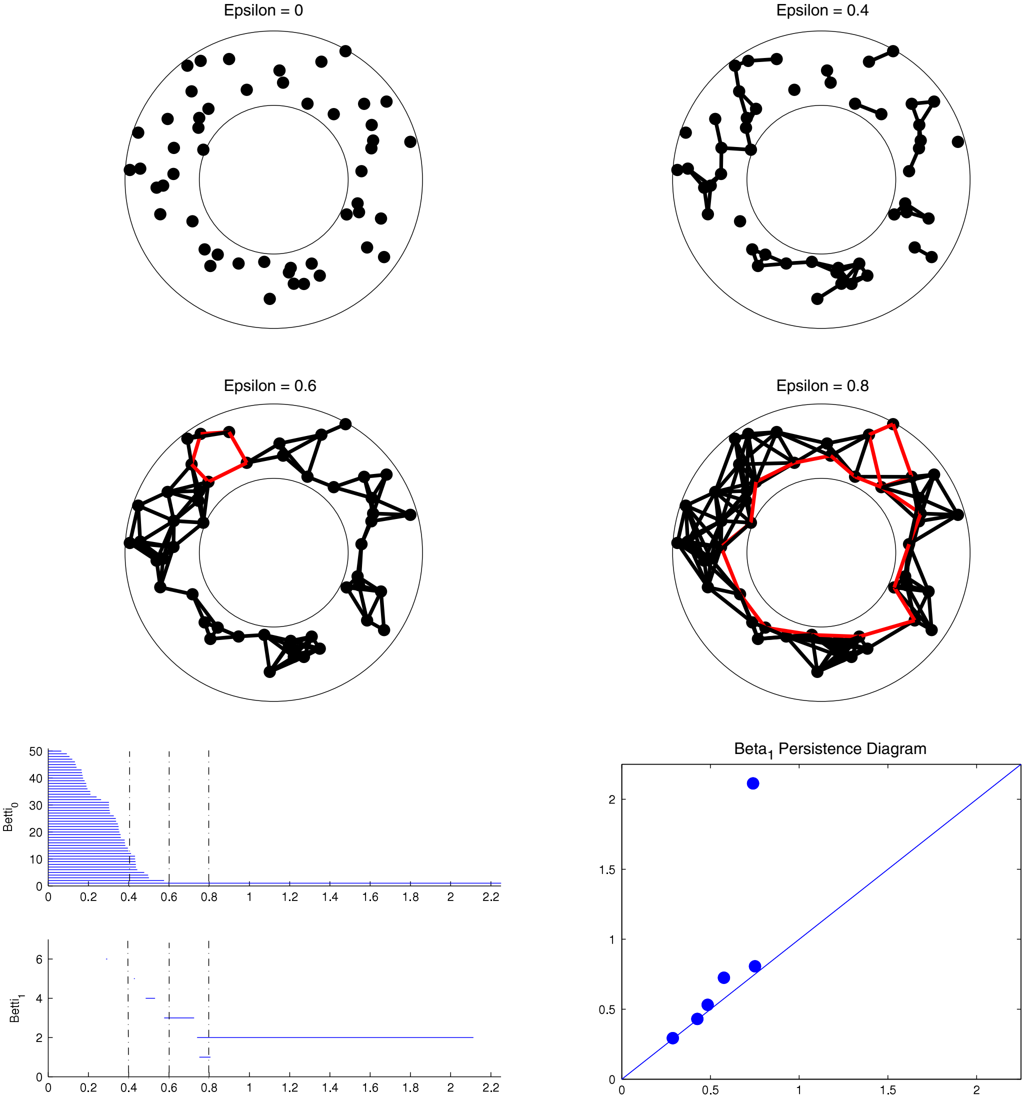

Fig. 2. Top left: a sample of 50 points from an annulus. Top right: points within $\varepsilon = 0.4$ of each other are connected. Middle row: points within $\varepsilon = 0.6$ (left) and $\varepsilon = 0.8$ (right) of each other are connected. One path around each loop is highlighted in red. Bottom left: β_0 and β_1 barcodes obtained by performing persistent homology on the sample. Bottom right: the corresponding β_1 persistence diagram. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

As a simple example, the points could be sampled from the surface of an annulus in two dimensions with inner radius 1 and outer radius 2 (see Fig. 2, top left). In this example, the topology is calculated using the Rips complex restricted to two dimensions. In other words, edges and triangles are part of the complex when their landmarks are all pairwise within ε. The first two rows of Fig. 2 illustrate the 1-skeleton of the complex (i.e. triangles are not shown) for various values of ε. When

$\varepsilon = 0.4$ there are a number of connections, but still many separate connected components. When $\varepsilon = 0.6$, there is a single connected component, and one loop, highlighted in red (although it does not correspond to the 'main' loop of the annulus). With $\varepsilon = 0.8$ the main loop is now closed, but an extra loop due to sampling still remains. In this case, there is only one significant topological feature, but in cases where 'true' topological features are present at multiple scales, then perhaps no one ε value will reveal them all. Alternatively, features may be visible at some scales which are only artifacts of noise, and do not reflect true topological features (for example, the additional loop at $\varepsilon = 0.6$).

In general, when $\varepsilon = 0$, the complex ($\mathcal{K}_0$) will consist only of all the sample points (with no edges connecting them), and for some large ε all edges and faces (and tetrahedra, if allowing three-dimensional simplices) will be present. This largest simplicial complex, call it $\mathcal{K}$, is called the *flag complex* (restricted to two or three dimensions). If the edges and faces of $\mathcal{K}$ are ordered by their entry time (i.e. the set of simplices δ^i are those that join at the ith discrete ε value), and $\mathcal{K}_i = \mathcal{K}_{i-1} \cup \delta^i$, then the nested sequence $\mathcal{K}_0 \subseteq \mathcal{K}_1 \subseteq \mathcal{K}_2 \subseteq \cdots \subseteq \mathcal{K}$ is a *filtration* of $\mathcal{K}$.

A sensible way to distinguish true features from noise, is to consider the features that persist through multiple scales of resolution to be *significant*. Persistent homology takes a filtered simplicial complex, and calculates how topological features persist through the filtration. The results are then visually represented in a *barcode*, with one barcode for each dimension (e.g. Fig. 2, bottom left). For each, the horizontal axis represents ε, and each horizontal line represents a topological feature of that dimension. For example, in a β_1 barcode, a horizontal line begins at the ε value when a loop is formed (through pairwise connections), and ends when the loop 'dies' (i.e. is filled in by triangles joining the simplicial complex). For the β_0 barcode, at $\varepsilon = 0$ each data point is a separate component (with no connections between them), so a horizontal line begins for every point. As ε grows and connections are made, some components join together, and the horizontal lines end, with just one *persisting* indefinitely. In Fig. 2 (bottom left) the β_0 and β_1 barcodes are shown, with vertical dashed lines indicating $\varepsilon = 0.4, 0.6, 0.8$. The number of horizontal lines intersecting each of these dashed lines indicate the number of topological features present (e.g. two loops at $\varepsilon = 0.8$). The long horizontal bar in the β_1 barcode corresponds to the hole in the annulus.

An alternative way to visually represent persistent homology results is in a *persistence diagram*. Each horizontal line in the barcode representation is uniquely determined by its start and end points (call them a and b), and can be represented by the point (a, b) in the Euclidean plane. The degree-n persistence diagram consists of the points representing the birth and death times for each topological feature of degree n, along with all points along the diagonal (for technical purposes). The β_1 persistence diagram for the annulus example is given in Fig. 2 at the bottom right. Each loop formed (at any point during the filtration) is represented by an off-diagonal point, but those that are a product of noise lie very close to the diagonal. The two that lie extremely close to the diagonal in fact, are only visible as faint points in the barcode (instead of lines), since they disappear so quickly. The significant feature (the main hole) is represented by the point at (0.74, 2.11), since the main loop is formed when $\varepsilon = 0.74$, and is 'filled in' by triangles at $\varepsilon = 2.11$.

The persistent homology algorithm [11,22], takes as input a filtered simplicial complex and creates pairings between the simplices that mark the birth and death times for topological features. As such, it has a wide range of applications, since there are many different ways a filtered simplicial complex could be obtained. A typical way is as above, forming a Rips complex using the Euclidean distance on a set of data points, but it is also possible to use alternative distance measures, to define the filtration differently (e.g. based on a Čech or witness complex), or more generally, to obtain a filtered abstract simplicial complex based on the critical values of a Morse function (see [21] for details).

The majority of the literature on persistent homology does not incorporate any statistical methodologies, but instead focuses on mathematical and computational theory, or uses persistent homology as a data analysis method on its own. Some notable exceptions are some recent papers by Bubenik and Kim [3,5,2]. Their work [3], and their work with Carlsson and Luo [2] are the first to take both parametric and nonparametric statistical approaches to the subject. In their work with Chung [5], they analyze cortical thickness in autistic patients and controls by first applying kernel smoothing to the cortical thickness data, followed by persistent homology to obtain a persistence diagram for each subject. They discuss using distances between persistence diagrams to compare subjects, and use a plot of pairing concentrations to distinguish between groups. In the current paper, we also consider distances between persistence diagrams, but instead calculate them and apply dimensionality reduction methods followed by statistical analysis to determine significant differences between groups.

In the following section, we present our method for the use of persistent homology to analyze landmark-based shape data.

3. Persistent homology on landmark configuration shape data

The outline of our proposed method is as follows:

1. Using the same fixed simplicial complex, perform persistent homology on each of the three-dimensional landmark configurations individually. The choice of filtration, and distance measure between nodes will be discussed below. The β_0, β_1, and β_2 barcodes or persistence diagrams are obtained for each.
2. Distances between the barcodes or persistence diagrams can be obtained using the Wasserstein distance as described in [6], possibly restricted to each dimension. If there are n landmark configurations under consideration, then an $n \times n$ distance matrix is obtained for each of β_0, β_1, and β_2, as well as the full Wasserstein distance.

3. Statistics such as hierarchical clustering can be performed directly on the distance matrices. To perform other multivariate statistics, dimensionality reduction methods can be used to embed the points in a lower number dimensions. These embeddings can be obtained using multidimensional scaling (MDS), or ISOMAP [20] if necessary.

3.1. Performing the filtration

Since each configuration is represented by k homologous landmarks, these can be used as vertices of a simplicial complex, which is fixed across all subjects. The simplest way to do this would be to use the flag complex obtained from the k landmarks, restricted to a low number of dimensions, say three (i.e. every pair of landmarks will have an edge between them, and all the associated triangles and tetrahedra will also be part of the complex). This flag complex will not generally have a geometric realization in a low number of dimensions, however. If the 'true nature' of the data is known to be inherently two or three dimensional (our orthodontic data is inherently three-dimensional, since the landmarks come from human subjects), to preserve this nature it is perhaps preferable to choose from this flag complex a subcomplex that has a geometric realization in the appropriate number of dimensions. The Delaunay triangulation of the k landmarks can be used for this purpose. This way, when persistent homology is performed on each of the 240 landmark configurations the same abstract simplicial complex can be used (but with different filtrations). See the end of Section 5.1 for discussion of how choosing the Delaunay triangulation differs from the three-dimensional flag complex in our example.

If the Delaunay triangulation was calculated separately for each of the 240 landmark configurations, it is likely that they would all be equivalent (i.e. isomorphic) as abstract simplicial complexes. There is a small chance however, that due to differences in the distances between landmarks for each subject, the Delaunay triangulations could be slightly different. To avoid this, the Delaunay triangulation obtained from Dryden and Mardia's *mean shape* [9] of the entire group can be chosen as a representative simplicial complex.

There are a number of different choices of filtration available. The most obvious one is to perform a Rips filtration using the Euclidean distance between landmarks/vertices. In other words, all nodes enter at time $t = 0$, and the edge between landmarks i and j enters at time $t = ild(i, j)$, where $ild(i, j)$ is the Euclidean inter-landmark distance between landmarks i and j. A triangle joins the filtration when all three of its edges have joined, and the three-dimensional tetrahedron joins when all of its faces have. The point in the filtration t at which a simplex joins is called its *entry time*. This type of filtration would likely be most appropriate when the absolute structure of a landmark configuration is of interest: landmarks that are close together will join with each other more quickly, and loops and voids will form around areas without landmarks.

When the goal of analysis is to compare multiple landmark configurations, a filtration which may be more appropriate is one which considers inter-landmark distances within an individual *relative to those same inter-landmark distances in other individuals*. Define a Rips filtration where all the nodes join at $t = 0$, but the edge between landmarks i and j joins at time

$$t = \frac{ild(i, j) \text{ in that subject}}{\text{average of } ild(i, j) \text{ over all subjects}}, \tag{1}$$

and higher-dimensional simplices join once all of their faces have. When this filtration is performed on an individual landmark configuration, the edges that join the filtration first are those that are smallest in that subject, relative to the entire group. For example, there may be a subject where every inter-landmark distance in that subject is smaller than the group averages, but the edges with the smallest ratio as defined in Eq. (1) will join the filtration first.

Persistent homology will be performed using the adapted Rips filtration defined in Eq. (1), restricted to the abstract simplicial complex obtained from the Delaunay triangulation of Dryden and Mardia's mean shape (as opposed to the entire flag complex of the vertices). It should be noted that although this particular filtration was chosen (for reasons given in the previous paragraph), choices of other filtrations are possible. For example, a filtration using raw inter-landmark distances, a Čech filtration, or one based on the critical values of a Morse function, could also easily have been used. Also, the entire flag complex, or another subcomplex (besides the Delaunay triangulation) could have been used as the fixed abstract simplicial complex. Of course, the results may depend on these choices, and this is an area for future research (discussed further in Section 6).

After choosing a filtration and performing persistent homology on each landmark configuration, β_0, β_1, and β_2 barcodes and persistence diagrams are obtained for each. It is the Wasserstein distance between persistence diagrams that will be used as a distance measure (discussed in the next section), but barcodes may still be useful for visualization and interpretation of the filtrations.

3.2. Distance between persistence diagrams

It is desirable to obtain some type of distance measure between landmark configurations, in order to judge similarity or dissimilarity between multiple configurations. The Wasserstein distance between persistence diagrams can be used for this purpose. A full description of this distance, along with stability properties can be found in [6].

To calculate the Wasserstein distance between persistence diagrams a bipartite matching algorithm [18] is performed. To match two persistence diagrams, each point in one must be matched to either a point in the other, or a point along the diagonal. The algorithm obtains the matching which minimizes the square root of the sum of squared L_∞-distances between pairs. In the notation of [6], if a and b are two landmark configurations with corresponding persistence diagrams (of degree l) $\mathrm{Dgm}_l(a)$ and $\mathrm{Dgm}_l(b)$, and $\gamma_l(x) : \mathrm{Dgm}_l(a) \to \mathrm{Dgm}_l(b)$ is a matching between the β_l persistence diagrams, then the degree-p Wasserstein distance between the persistence diagrams of landmark configurations a and b is

$$W_p(a, b) = \left[\sum_l \inf_{\gamma_l} \sum_{x \in \mathrm{Dgm}_l(a)} \|x - \gamma_l(x)\|_\infty^p \right]^{1/p}.$$

For our analysis $p = 2$ is used. Note that in the definition, the sum of squared distances for the β_0, β_1 and β_2 matchings are each calculated, and then the three are summed and the square root taken to obtain the Wasserstein distance. It may also be informative however, to keep the three separate, and analyze each of the dimensions individually. The stability of W_p (as discussed in [6]), implies the stability of the squared distances $\sum_x \|x-\gamma_l(x)\|_\infty^p$ for each of $l = 0, 1, 2$ (since each is strictly less than W_p). We will simply refer to these sum of squared distances for each dimension as the β_0, β_1 and β_2 distances respectively. If n configurations are under comparison, then the Wasserstein distances between persistence diagrams may be represented as a $n \times n$ matrix of distances. Additionally a $n \times n$ matrix of distances may be retained for each of the β_0, β_1 and β_2 matchings.

3.3. Statistical analysis

The matrix of distances (either the full Wasserstein distance, or each of the three dimensions individually) may be analyzed directly, using methods such as hierarchical clustering. To perform usual multivariate statistics, dimensionality reduction can be performed to embed the points in a low number of dimensions. Multidimensional scaling (MDS) (see [1], or the appropriate chapter of most textbooks on multivariate statistical analysis) can be used for this purpose, with the appropriate number of dimensions chosen using a scree plot. A non-linear dimensionality reduction method such as ISOMAP [20] could also be used, but is only necessary if it gives different results from the simpler linear method.

After the embedding, any usual statistical methods for Euclidean data may be applied.

4. Traditional shape analysis methods for landmark-based data

Many methods have been proposed for the analysis of landmark-based shape data, each with their own advantages and drawbacks. Here we discuss two of them, for comparison with the persistent homology-based method.

4.1. Euclidean Distance Matrix Analysis (EDMA)

This method of shape analysis was developed in the early 1990s, and is presented fully in [17]. It represents each configuration of k landmarks as the set of all pairwise inter-landmark distances. These can be held in a symmetric $k \times k$ *form matrix*, with zeros along the diagonal, but to avoid repetition is often combined into a vector of length $k(k-1)/2$. This representation is invariant to rotations and translations, and the term 'form matrix' is often used interchangeably to represent the matrix itself, or its equivalent vector.

To compare two configurations, the ratio of their corresponding form matrices is calculated entry-wise to create a *form difference matrix*, or FDM. For this, one configuration must be chosen as the 'numerator' and one as the 'denominator'. If the entry of the FDM corresponding to landmarks i and j is larger than 1, then $ild(i, j)$ is greater in the numerator configuration than the denominator configuration (and vice versa if the corresponding FDM entry is less than 1). The ratio of the largest to the smallest entry in the FDM can be used as a statistic for degree of shape difference, and is called the T statistic. If the configurations are very similar, then all the entries in the FDM will be close to 1 (or some other constant if there is size but no shape difference), and T will be close to 1. Larger values of T indicate a greater degree of dissimilarity. Note that T is invariant to which of the configurations is used as the numerator, and which as the denominator.

When multiple groups of configurations are under comparison, an estimated mean form matrix can be calculated for each. While this does not uniquely determine a landmark configuration (since it is invariant to rigid motions), an instance may be chosen as an *icon* (perhaps centred at the origin) for graphical purposes. The algorithm for obtaining an estimated mean form matrix, along with an instance of the mean shape, is given in [17]. The ratio of estimated mean form matrices for two groups can be used as an estimated FDM, with an estimated T statistic calculated to measure dissimilarity. Without knowing the variability of the underlying populations, it is difficult (if not impossible!) to determine how large a T statistic must be to signify a 'significant' difference in mean form matrices. To estimate what type of variability in T would be expected if the samples under comparison were, in fact, drawn from the same population, a bootstrap procedure can be used (with one of the samples chosen as baseline).

If two groups are seen to be significantly different from each other, the estimated FDM entries can be analyzed directly to help localize where those differences occur. Another bootstrap procedure can be performed to obtain confidence intervals on the individual FDM entries, and determine which ILDs are significantly larger or smaller in one group than the other.

EDMA will be applied to the example orthodontic data set in Section 5.2.1.

4.2. Dryden and Mardia's methods

The mathematical details for all the methods outlined in this subsection are available in Dryden and Mardia's book *Statistical Shape Analysis* [9].

The first step in this shape analysis method is to factor out the rigid motions of translation and rotation (and scaling, if desired) in a standardized way. The method used to do this is called Procrustes superimposition. The method involves 'fitting' the landmark configurations over top of each other in such a way that the sum of squared distances between corresponding landmarks is minimized. If only two landmark configurations are under consideration, then there exists an explicit formula to fit one onto the other (although the fitting is not invariant to the choice of which of the two configurations is 'fixed'). If there are more than two configurations to be fit, however, the minimization must be performed through an iterative algorithm.

Although the details are beyond the scope of this paper, Dryden and Mardia consider each landmark configuration (mod rotation and translation) as a single point in a higher-dimensional *shape space*. This point represents the entire equivalence class of a given landmark configuration, under the group action of rigid motions. To allow for distance measures between points in the shape space, and for regular statistical methods to be applied, a tangent space approximation to the shape space, at the mean shape is considered. Distances between configurations may also be calculated directly in the shape space (with three choices of distance measure available), however this may only be done in a pairwise fashion. Tangent space approximations allow Euclidean coordinates to be assigned to each point, which can be used as input for statistical analysis.

To compare two samples, Dryden and Mardia propose the use of an adapted Hotelling's T^2 test, which uses the Moore–Penrose generalized inverse of the estimated covariance matrix. Usual multivariate statistical analyses are still possible in the tangent space, but become problematic if the number of subjects is small compared to the dimension of the tangent space. This is often the case, especially in medical and biological applications, where the number of subjects is often relatively small. For k landmarks in d dimensions, the dimension of the tangent space is $(k-1)d - \frac{d(d-1)}{2}$, which can become large. One way to address this is to apply principal component analysis to the tangent space coordinates, and use only the few most useful dimensions for further analysis.

These methods will be applied to the example orthodontic data set in Section 5.2.2.

5. Application to orthodontic data set

In this section a landmark-based data set will be analyzed. An ongoing orthodontic study was undertaken at the University of Alberta's Division of Orthodontics (principal investigators Manuel Lagravere and Paul Major), to determine the effect of two types of maxillary (i.e. upper jaw) expansion treatment. Sixty subjects were randomly allocated to three different treatment groups (Control: $n = 19$, Bone-anchored: $n = 21$, or Tooth-anchored: $n = 20$) were measured at four different time points (Time 1: baseline, Time 2: mid-treatment, Time 3: appliance removal, Time 4: follow-up). The control group was not measured at time point 2, so their time 1 measurements were used for both points in the analysis. Each subject at each time point is represented by a set of 22 three-dimensional landmarks that were chosen to be of primary interest to the study, and whose placement was checked for reliability. These landmarks are labeled 1 through 22, and are displayed in Fig. 3 (left) and 4.

The primary objective of this study is to identify any significant differences between the three groups over time. Specifically, whether there is increased maxillary expansion in the treatment groups, as compared to the control group, and whether there are any differences between the two treatment groups.

5.1. Persistent homology on each landmark configuration

The persistent homology method discussed in Section 3 was applied to this data set. The computations were performed in MATLAB, with code adapted from the program Plex [8]. The filtration defined in Eq. (1) was used, where the average in the denominator of t is taken over all subjects at all time points ($n = 240$). All 240 were also used to obtain the mean shape whose Delaunay triangulation was used. The Delaunay triangulation of the mean shape is shown in Fig. 3 on the right. The β_0, β_1 and β_2 barcodes and persistence diagrams were obtained for each landmark configuration. The β_0, β_1 and β_2 barcodes as well as the β_1 and β_2 persistence diagrams for one of the subjects at baseline are shown in Fig. 5.

Performing clustering methods (single linkage hierarchical clustering) directly on the matrix on distances yields one large cluster, with the other clusters containing only one or two elements, corresponding to the farthest outliers. This is true for all four distance matrices (corresponding to β_0, β_1, β_2 and full Wasserstein). This indicates that the configurations are not split into clearly defined groups of different shapes. To perform usual statistical analysis, dimensionality reduction is performed. It was found that the results using ISOMAP were similar to those found using MDS, so all analysis is performed using the MDS embedded coordinates.

The maxilla, when viewed from overhead, forms a 'U' shape, which is expected to widen as a result of maxillary expansion treatment. Since this 'U' is essentially a one-dimensional structure, it intuitively seems that one-dimensional features (i.e. loops) might be best at highlighting differences in jaw width. Using this intuition, we will analyze the MDS embedded

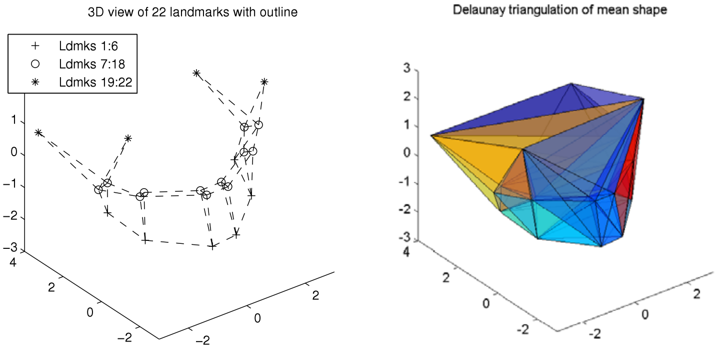
3D view of 22 landmarks with outline
Ldmks 1:6
Ldmks 7:18
Ldmks 19:22
1
0
−1
−2
−3
4
2
0
−2
−2
0
2
Delaunay triangulation of mean shape
3
2
1
0
-1
-2
-3
4
2
0
-2
-2
0
2

Fig. 3. Three-dimensional plot of the 22 landmarks (left) and their Delaunay triangulation (right).

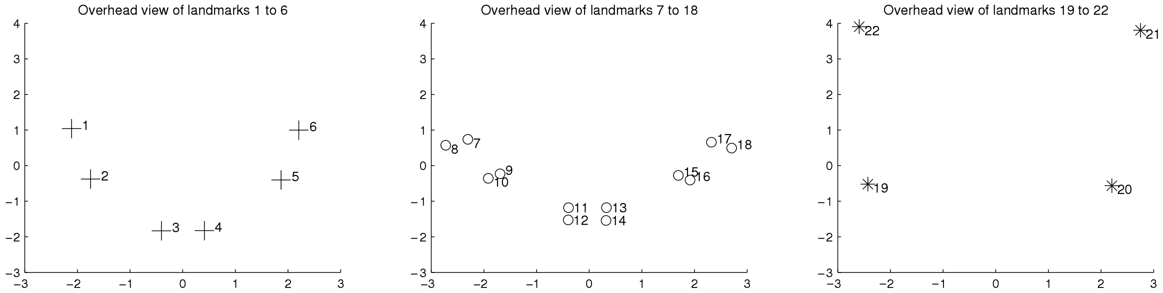

Fig. 4. Overhead views of landmarks 1 to 6 (left), 7 to 18 (middle) and 19 to 22 (right) for labeling purposes.

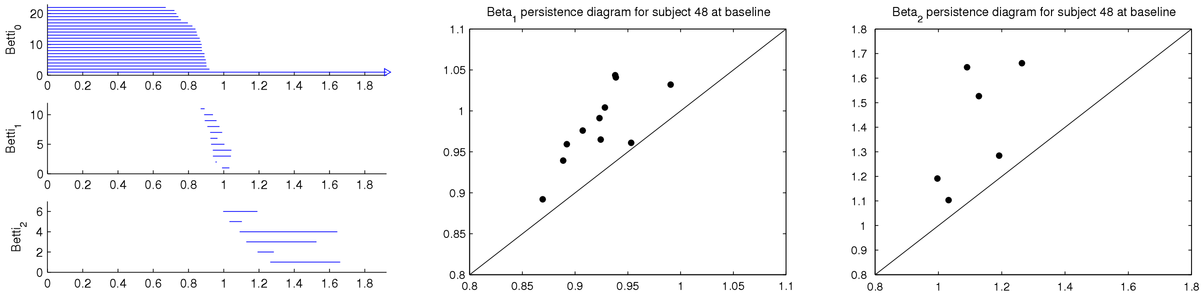

Fig. 5. Left: β_0, β_1 and β_2 barcodes for one of the subjects at baseline (chosen randomly for illustrative purposes). Middle and right: the corresponding β_1 and β_2 persistence diagrams for that subject.

coordinates obtained from the β_1 distances first (before the MDS coordinates obtained from the β_0, β_2, or full Wasserstein distances).

If Y is the $n \times p$ matrix of embedded coordinates (for some dimension $p < n$), then the eigenvalues of Y^TY can be analyzed to determine an appropriate number of dimensions to use, and to check if the original distance matrix was Euclidean. If the largest d eigenvalues are significantly larger than the remaining ones, then d is an appropriate number of dimensions for embedding. A line plot of the eigenvalues (called a scree plot) can help to determine a cut off. The scree plot for the β_1 MDS coordinates is shown on the left in Fig. 6, and the pronounced 'elbow' at $d = 3$ indicates that the first two embedded dimensions account for a large proportion of the variability between subjects. Any negative eigenvalues indicate non-Euclidean structure in the original distance matrix, but this is only a problem if the largest negative eigenvalue (in absolute value) is large compared to the largest positive eigenvalue. One rule of thumb is to only include dimensions whose positive eigenvalues are larger than the largest (in absolute value) negative eigenvalue. In this case the negative eigenvalue with the largest absolute value is -0.2848. There are seven positive eigenvalues with absolute value greater than this, so only the first seven would be appropriate for use (although the scree plot has already indicated that only two are necessary).

Looking only at the scatterplot of the first two MDS dimensions (see Fig. 6, right) with all 240 points labeled by treatment group, the differences between the groups are not obvious, although it appears that the control group's values in both coordinates are generally smaller than the coordinates for the two treatment groups. Repeated measures MANOVA is performed to determine if any statistically significant differences exist. The factors of interest are *time* and *group*, as well as

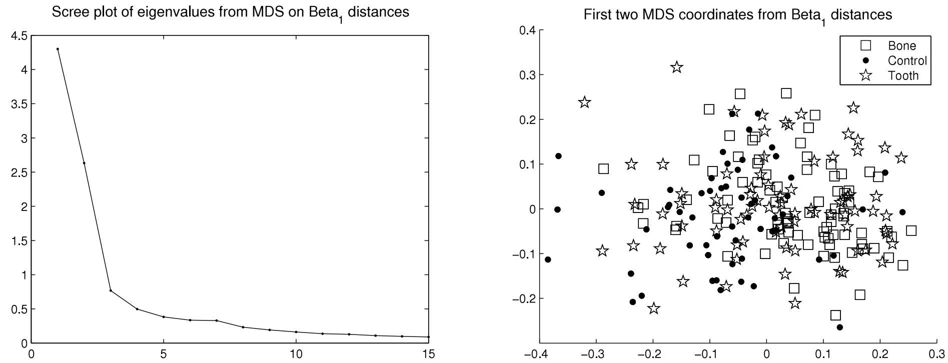

Fig. 6. Scree plot for the ordered eigenvalues from MDS on the β_1 distances (left), and the first two embedded MDS coordinates, labeled by treatment group (right).

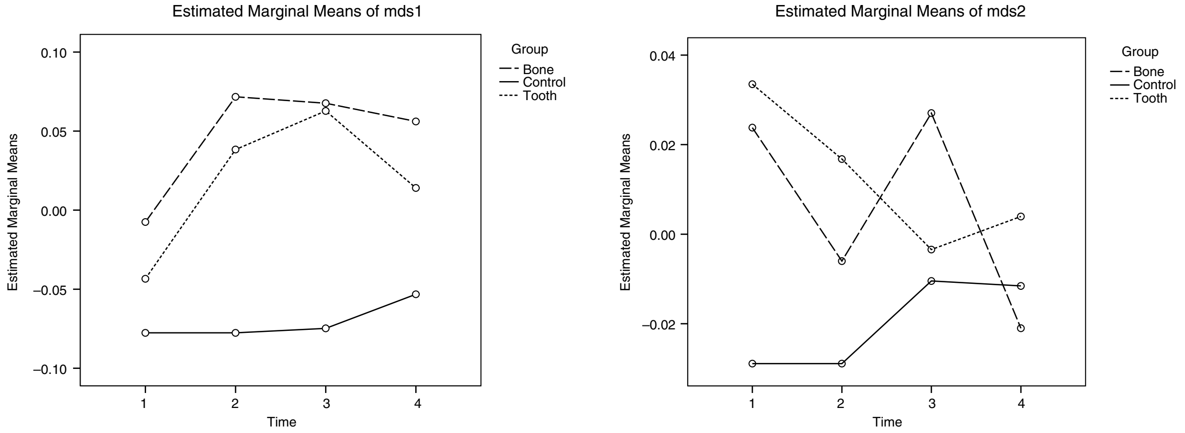

Fig. 7. Profile plots for the first (left) and second (right) embedded MDS dimensions from the β_1 distances.

a *time*group* interaction. The *time*group* interaction will be significant if the way that the landmark configurations change through time (as measured by the first two embedded MDS coordinates) is different depending on the treatment group.

It should be noted that the assumptions for MANOVA are satisfied, with Box's M, Mauchley's test for sphericity, as well as univariate Levene's tests at each time point, all non-significant. The overall repeated measures MANOVA multivariate test (Wilks' Lambda) shows strong significance for *time*, *group*, and *time*group* (with *p*-values of ∼0.000, 0.030, and 0.011 respectively). Upon further analysis, the univariate tests on each embedded coordinate show that only the first MDS dimension is significant for each of *time*, *group* and *time*group* (*p*-values ∼0.000, 0.007, and ∼0.000), while within- and between-subject tests show the second MDS dimension is not significant for either of the factors or their interaction. Examining the profile plot of the first coordinate (Fig. 7 left), the control group remains largely unchanged through time, while the two treatment groups increase significantly from time point 1 to time point 2, and then stay comparatively high for time points three and four (although the tooth-anchored group shows some slight regression at the fourth time point). The profile plot of the second embedded coordinate (Fig. 7 right) shows slight variation over time and between groups that appears somewhat random, and is not statistically significant.

The other shape analysis methods discussed in this paper perform tests to compare the groups *within each time point*, so to facilitate comparison, ANOVA is also performed on the first MDS coordinate at each of the four time points (the second MDS coordinate is not used, since it was already determined to be non-significant. If it was however, MANOVAs could be performed). The results are displayed in Table 1.

To help interpret what these differences mean clinically, it would be useful to know if the first embedded MDS coordinate corresponds to any features in the subjects. To check for this, a linear correlation is performed between the first embedded coordinate and each of the inter-landmark distances using all the landmark configurations. A correlation is also performed between the first embedded coordinate and the centroid size of each configuration (which is the square root of the sum of the squared distances from the landmarks to their centroid).

It is seen that a number of inter-landmark distances (ILDs) correlate strongly with the first embedded MDS coordinate. Fig. 8 (left) shows all the inter-landmark distances, colour coded by the strength of their correlation with the first MDS coordinate. The distances with the strongest correlation (coloured in red and orange) are those that cross the mouth (either diagonally or horizontally). In other words, those that correspond to maxillary width. The distances that correlate the least

Table 1
Results of ANOVAs on the first β_1-based MDS coordinate at each of the four time points. Overall p-value is from Wilks' Lambda, and post hoc differences between groups are found using Bonferroni (p-value given for each pairwise comparison). P-values significant at $\alpha = 0.05$ are in bold.

Time point	ANOVA	Post hoc (Bonferroni)		
	p-value	B vs. C	C vs. T	T vs. B
1	0.266	0.317	1.000	1.000
2	**0.001**	**0.001**	**0.011**	1.000
3	**0.001**	**0.002**	**0.003**	1.000
4	**0.017**	**0.014**	0.239	0.770

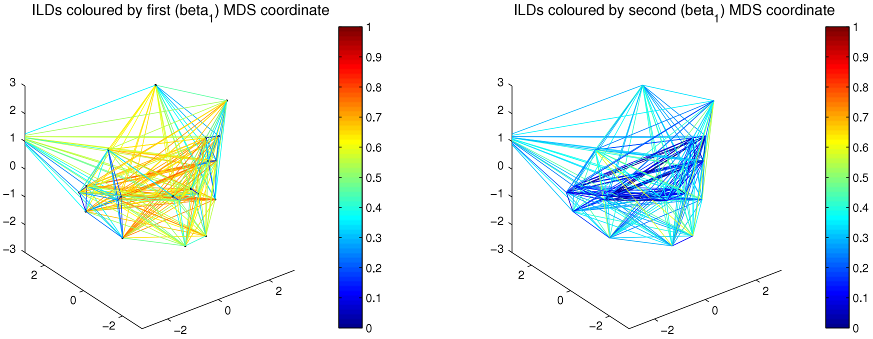

Fig. 8. Inter-landmark distances colour coded by their correlation with the first (left) and second (right) MDS coordinates derived from the β_1 distances.

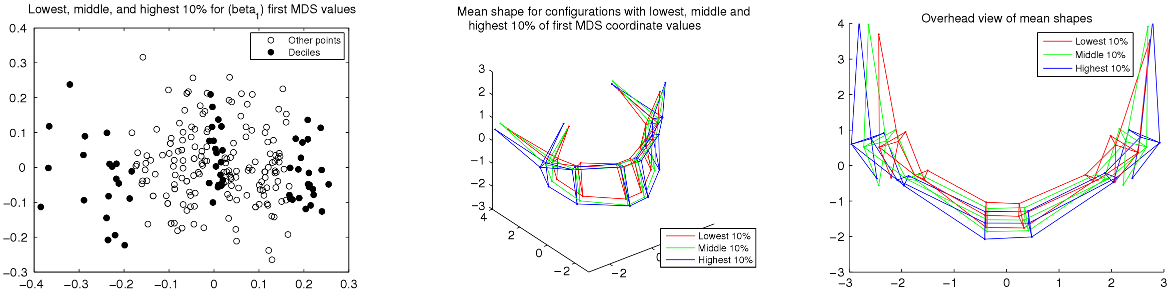

Fig. 9. The mean shapes of the landmark configurations which correspond to the lowest, middle, and highest first MDS values (based on the β_1 distances).

with the first MDS coordinate are those that are very near each other (e.g. landmarks within the same tooth). The centroid size also has a strong correlation with the first coordinate ($r = 0.894$). This interpretation corresponds well to the treatment of maxillary expansion. As the treatment progresses subjects in the treatment groups show increases in their first MDS dimension (Fig. 7) (which is interpreted as an increase in the inter-landmark distances corresponding to maxillary width), while patients in the control group see no discernible change. The second MDS dimensions is not correlated with maxillary width, but with the maxillary bone landmarks (landmarks 19 and 20), although the correlations are quite a bit weaker (Fig. 8 right).

To help with visualization, the mean shape was calculated for the lowest, middle and highest 10% of the first MDS values (Fig. 9 left). They are plotted together in Fig. 9 in three dimensions (middle), and in an overhead view (right). It can be seen clearly that the jaw expands as the first MDS dimension increases. Returning to the results of the repeated measures ANOVA on the first MDS dimension, this means that the control group does not see any jaw expansion, while the two treatment groups both do, with a slight regression in width for the tooth-anchored group at time point four.

This entire analysis can be performed again, using the β_0, β_2, or full Wasserstein distance matrices as input for MDS. These were each performed, but their results do not paint as clear a picture as the analysis based on the β_1 distances did.

Using β_0 distances, the scree plot indicated that only one embedded MDS dimension was useful for analysis (Fig. 10, left). While it was seen to have moderate correlations with a number of inter-landmark distances associated with maxillary width, it also showed correlations with other inter-landmark distances, such as those involving the maxillary bone landmarks (Fig. 10, middle). Repeated measures ANOVA on this dimension showed significant *time* effects, but no significant *group* differences or *time*group* interaction. The profile plot of the first MDS dimension values for the three groups across time is given in Fig. 10, right.

When MDS is performed on the β_2 distances, the first embedded coordinate only shows strong correlation (absolute value above 0.5) with two inter-landmark distances: between landmarks 7 and 8, and landmarks 17 and 18. These are pairs within the two teeth at the back of the mouth (where each pair is from one tooth).

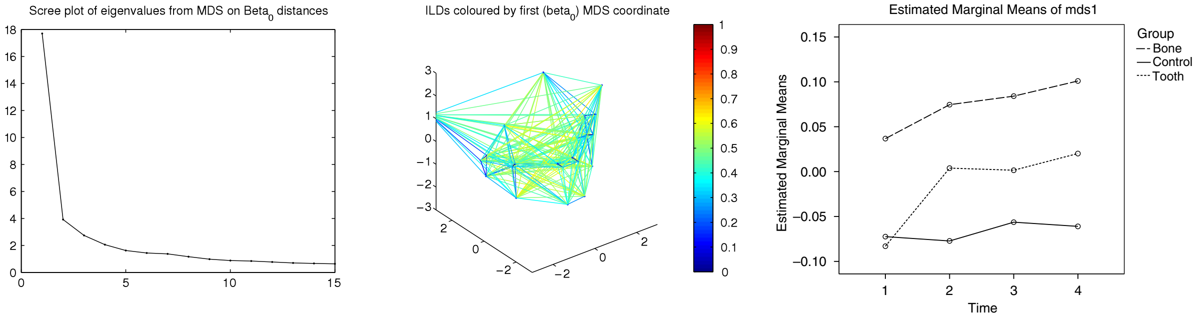

Fig. 10. Output from MDS on β_0 distances. Scree plot (left), inter-landmark distances colour coded by correlation with the first MDS coordinates (middle), and profile plot from repeated measures ANOVA on the first MDS coordinates (right).

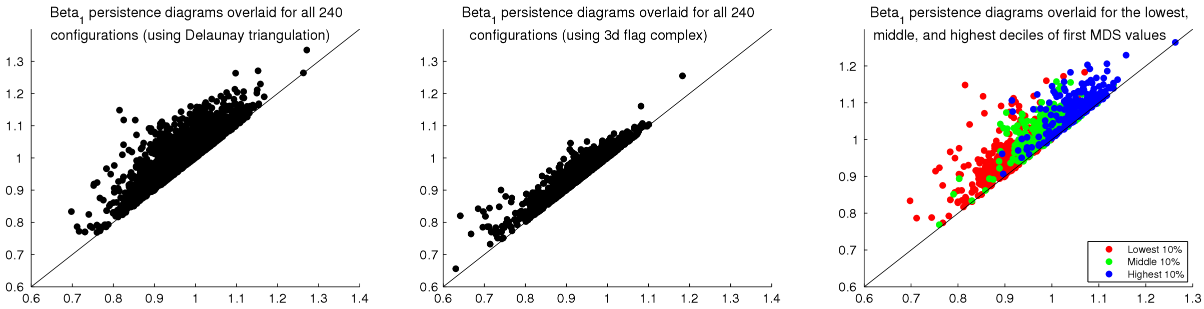

Fig. 11. 240 overlaid persistence diagrams calculated using the Delaunay triangulation (left), and the 240 persistence diagrams calculated using the three-dimensional flag complex (middle). Those calculated using the three-dimensional flag complex generally lie closer to the diagonal line. On the right are the overlaid persistence diagrams from the configurations with the lowest, middle, and highest deciles of first MDS values (using the Delaunay triangulation).

When the entire Wasserstein distances are used for MDS, the results are similar to those from the β_0 analysis. This is since the β_0 distances are generally larger than the β_2 and β_1 distances (with the β_1 distances the smallest), so they make a larger contribution to the Wasserstein distance. Performing repeated measures MANOVA on the first four embedded MDS dimensions, they are only significant for *time*, but not for *group* or *time***group*.

Recall from Section 3.1, that persistent homology was performed using the abstract simplicial complex formed by the Delaunay triangulation of the mean shape. This was chosen instead of the three-dimensional flag complex of the k landmarks, in the hope of retaining the inherently three-dimensional structure of the data. To validate this decision, the analysis above was re-run using the three-dimensional flag complex instead of the Delaunay triangulation. The results were similar, although not as pronounced: distances between β_1 persistence diagrams were the ones that displayed differences between the groups, but the statistical significance was moderate. The first MDS coordinate (based on the β_1 persistence diagrams) was correlated again with the ILDs corresponding to maxillary width, but the correlations were weaker (closer to 0.5) than those seen when the Delaunay triangulation was used. We speculate that when the flag complex is used, and any edge or triangle is allowed to enter during the filtration (instead of just those in the Delaunay triangulation), the effect is that loops and voids tend to close more quickly, and this makes any effects seen in the persistence diagrams less pronounced. To illustrate this, the 240 persistence diagrams calculated using the Delaunay triangulation are overlaid in Fig. 11 (left), and the 240 persistence diagrams calculated using the three-dimensional flag complex are overlaid in Fig. 11 (middle). Generally, the points in those calculated using the flag complex lie much closer to the diagonal line, which indicates that loops do not persist as long during the filtration.

The barcodes and filtrations of individual landmark configurations can be subject to further analysis. This will help to make more precise our original intuition that information about the persistent loops would give the 'best' information about maxillary width. Part of the explanation is likely that (based on the filtration we used) loops generally form earlier in the filtration in configurations that are smaller than average, and form later in the larger configurations. Thus, the distances between the β_1 persistence diagrams would be larger between configurations of different sizes. The lowest, middle and highest deciles of first MDS values (originally seen in Fig. 9) are used in the overlay plot of persistence diagrams shown in Fig. 11 (right). It shows that the configurations with larger first MDS values do tend to have loops that form later in the filtration. The β_1 persistence diagrams give more than just size information however. If, instead of the first MDS coordinate, centroid size is considered for analysis, it is not as easy to localize where the differences between groups are occurring. A repeated measures ANOVA on centroid size does show significant differences in *time*, *group*, and *time***group* (p-values $\sim$0.000, 0.007, and 0.001, respectively), so the size change between the groups is different over time (profile plot in Fig. 12, left). However the correlations between the inter-landmark distances and centroid size are large for many ILDs (Fig. 12, right),

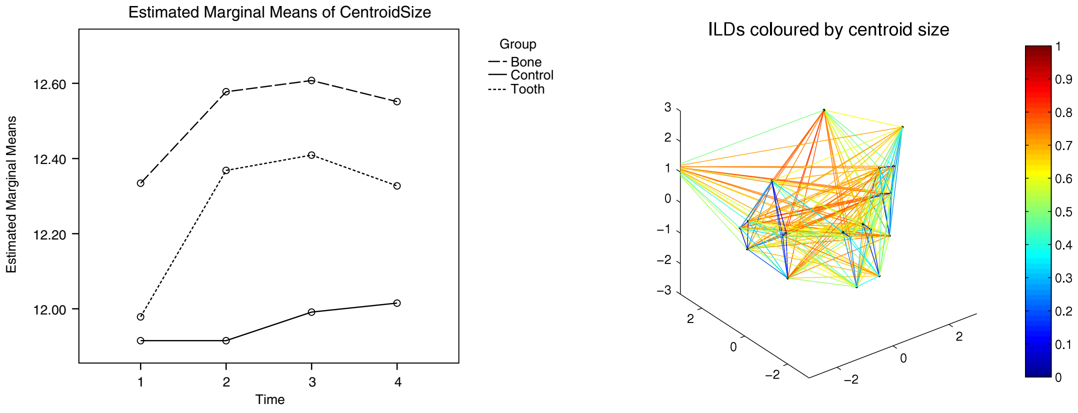

Fig. 12. Profile plot for repeated measures ANOVA in the centroid size (left), and the inter-landmark distances colour coded by their correlation with centroid size (right).

Table 2
T statistics and confidence intervals for pairwise comparison of the three groups at each of the four time points. *T* statistics which lay outside at least one of the two confidence intervals are in bold.

A vs. B	*T*	95% confidence intervals	
		B as baseline	A as baseline
B1 vs. C1	1.3605	1.115–1.501	1.123–1.562
B1 vs. T1	1.2884	1.130–1.508	1.126–1.451
C1 vs. T1	1.3241	1.132–1.577	1.110 1.500
B2 vs. C2	**1.5780**	1.107–1.505	1.133–1.520
B2 vs. T2	1.2209	1.139–1.599	1.125–1.458
C2 vs. T2	**1.5729**	1.156–1.656	1.124–1.532
B3 vs. C3	**1.6704**	1.121–1.495	1.135–1.492
B3 vs. T3	1.4268	1.132–1 612	1.133–1 543
C3 vs. T3	**1.6536**	1.157–1.684	1.122–1.481
B4 vs. C4	**1.6253**	1.109–1.662	1.133–1.580
B4 vs. T4	1.3229	1.145–1.551	1.123–1.720
C4 vs. T4	**1.6654**	1.142–1.605	1.119–1.515

including the long ILDs involving landmarks 21 and 22 (not associated with maxillary width), so little information is given about *shape* differences between the groups.

5.2. Comparing with results from traditional landmark-based shape analysis

In this section we will compare the results and conclusions reached using persistent homology on the landmark configurations, with results from EDMA and the methods of Dryden and Mardia.

5.2.1. EDMA on the orthodontic data set

An outline of the EDMA method was given in Section 4.1.

To analyze the 22 landmark orthodontic data set using EDMA, first the groups were compared to each other within each of the time points. The *T* statistic and confidence interval comparing the bone-anchored and control groups at time one was calculated using the bone-anchored group as baseline, and then again using the control group as baseline, using 200 repetitions for the bootstrap procedure. This was done for the other two comparisons (bone vs. tooth, and tooth vs. control) at time point one, and again for all three comparisons at the other three time points. The *T* statistics and confidence intervals are summarized in Table 2, with significant *T* statistics in bold. The abbreviations used for the groups at each time point are: B1 for bone-anchored group at time point one, C1 for control, T1 for tooth-anchored, etc. It can be seen that at time point one, none of the groups of landmark configurations show significant differences in shape from one another, and also that the bone- and tooth-anchored groups are not significantly different from each other at any of the time points. The bone- and tooth-anchored groups are significantly different from the control group at each of time points two, three and four.

For each of the significant pairwise comparisons, the bootstrap procedure on the estimated FDM (described in Section 4.1) returns confidence intervals for each of the inter-landmark distances. Those ILDs with confidence intervals entirely above or below 1 show a significant difference in length between the two groups. Plots of the ILDs comparing Bone to Control (left) and Tooth to Control (right) are shown in Fig. 13. Blue ILDs show no differences between the groups, while those shown in pink are larger in the treatment group than the control group (their 95% confidence intervals do not contain 1). It appears that the tooth-anchored treatment group is different from control in the ILDs corresponding to maxillary width, whereas

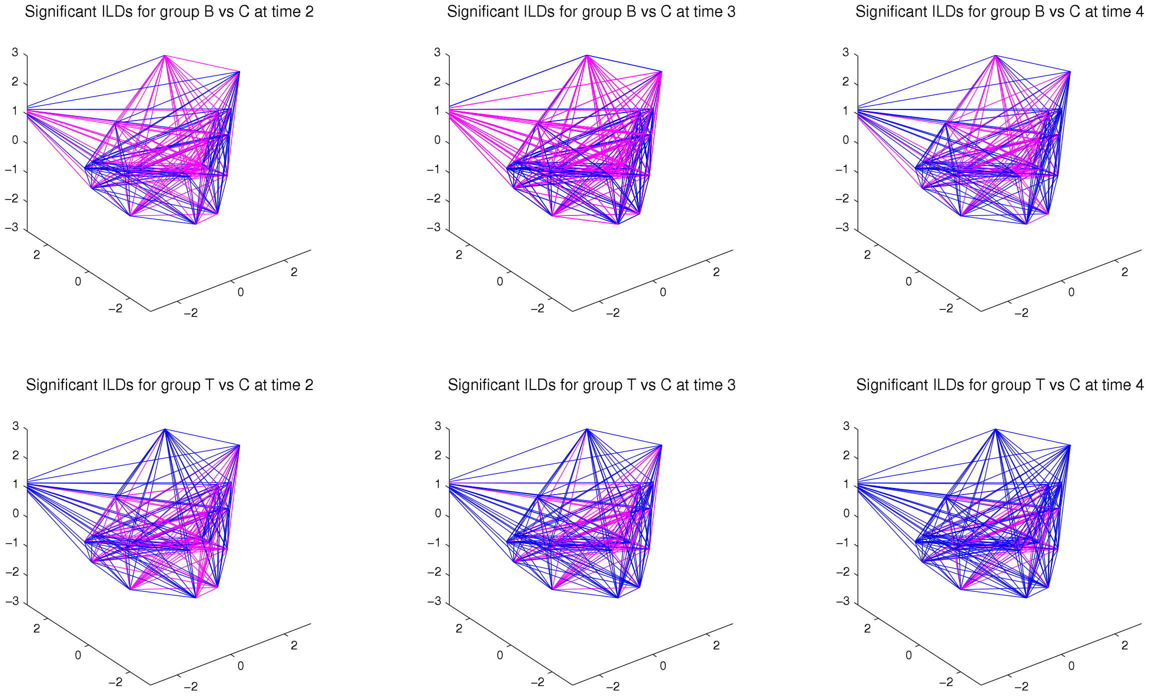

Fig. 13. EDMA results of bootstrap on estimated FDMs for Bone vs. Control (top row) and Tooth vs. Control (bottom row) at time points two, three and four. Inter-landmark distances shown in blue are not statistically different between the groups, while those shown in pink are (at the 0.05 significance level) larger in the treatment group than the control.

the bone-anchored group is different from control in the ILDs corresponding to maxillary width *as well as* the ILDs relating to maxillary bone landmarks (ldmks 19 to 22). This could indicate that the bone-anchored group saw some additional forward movement of the teeth relative to the maxillary bones.

5.2.2. Dryden and Mardia's methods applied

An outline of Dryden and Mardia's methods was given in Section 4.2.

An overhead view of the orthodontic data set landmarks for all 240 configurations are given in Fig. 14 before (left) and after (right) Procrustes superimposition. As was done with EDMA, pairwise comparisons between each of the groups can be made at each of the time points. The results show no obvious differences between the groups. The adapted Hotelling's T^2 test in the 60-dimensional tangent space is performed on the groups two-at-a-time, within time points. The F-statistics and p-values are given in Table 3, and none of the groups show any statistically significant shape differences at any of the time points. Despite this, a plot of the mean shapes for the three groups at time point four does seem to indicate that the control group mean shape has a narrower maxilla (see Fig. 15, left). Perhaps this is because the Hotelling T^2 tests for significant differences in shape as related to *overall* shape variability, whereas the first persistent homology-based MDS coordinate restricts the analysis to the 'dimension' along which the subjects show the greatest variability. More similar to this approach, is to perform principal component (PC) analysis on the tangent space coordinates, and analyze the first one (or more) PC scores though repeated measures (M)ANOVA.

A scree plot (Fig. 15, right) can be used again to choose the number of principal components for analysis. In this case, no clear 'elbow' is seen, so the choice is somewhat arbitrary. Correlations between the first PC score and the inter-landmark distances shows that it is most strongly related to landmarks 21 and 22 (the rear maxillary bone landmarks), while the second PC score is most strongly associated with the other two maxillary bone landmarks (landmarks 19 and 20). Colour-coded plots for these are given in Fig. 16. Analyzing the first two PC scores, repeated measures MANOVA shows no significant differences between the groups. *Time* is significant, with p-value ∼0.000, but both *group* and *time*group* are not significant (p-values 0.80 and 0.17 respectively). Additionally, separate MANOVAs using the first two PCs at each of the four time points show no significant differences between the groups.

The principal component analysis applied above is based on a linear approximation to the shape space. A recently proposed method by Huckemann et al. [16] called *geodesic principal components* is more reflective of manifold curvature, and could possibly produce more accurate results.

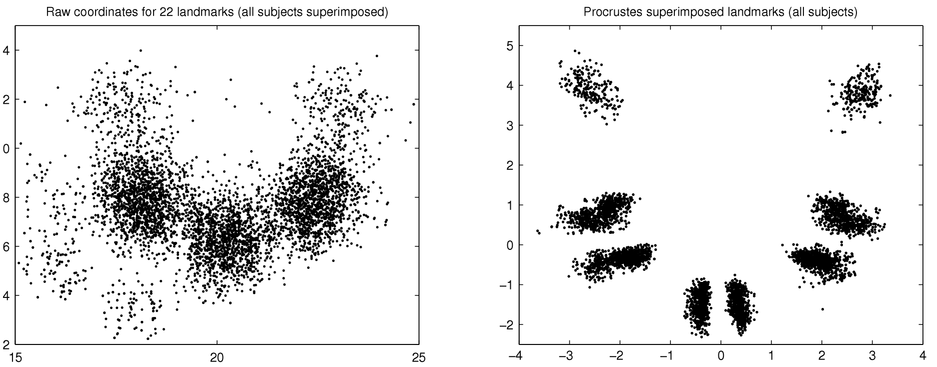

Fig. 14. Overhead view of the landmarks for all 240 configurations before (left) and after (right) Procrustes superimposition.

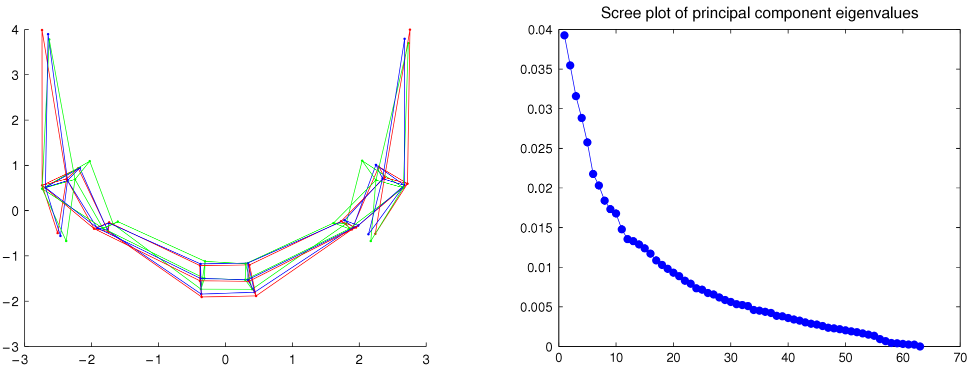

Fig. 15. On the left is an overhead view of the mean shapes calculated for the three treatment groups at time point four, with the bone-anchored treatment group in red, tooth-anchored in blue, and control in green. On the right is the scree plot for the principal component scores obtained from the tangent space coordinates. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

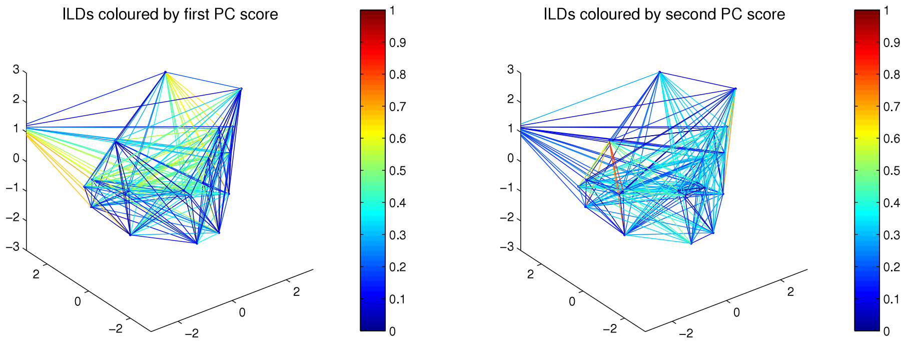

Fig. 16. Colour-coded plots of the correlations between the first (left) and second (right) PC scores with the inter-landmark distances.

Table 3

F statistics and *p*-values from Dryden and Mardia's adapted Hotelling T^2 test, comparing the groups pairwise at each time point.

Comparison	*F* statistic	*p*-value
B1 vs. C1	0.072	1.000
B1 vs. T1	0.100	0.997
C1 vs. T1	0.114	0.995
B2 vs. C2	0.209	0.966
B2 vs. T2	0.074	0.999
C2 vs. T2	0.456	0.853
B3 vs. C3	0.321	0.914
B3 vs. T3	0.122	0.993
C3 vs. T3	0.401	0.877
B4 vs. C4	0.182	0.975
B4 vs. T4	0.106	0.996
C4 vs. T4	0.159	0.983

5.3. Reconciling the results

With the above example, three different methods of analysis were performed, with somewhat differing results. The methods of Dryden and Mardia were able to detect changes within the entire group over time, but no differences between treatments and the control. As mentioned previously, this may be due to the fact that this method considers the entire configuration as a whole, so large overall within-group variability can obscure more specific between-group differences. The EDMA method can make pairwise comparisons between groups, but no repeated measures type of analysis. Differences are seen between each treatment group and the control group at time points 2 through 4, and these differences can be broken down further by analyzing each inter-landmark distance. Both tooth- and bone-anchored treatments were larger than the control group in ILDs related to maxillary width, but additionally the bone-anchored group was larger in some front-to-back dimensions between the maxillary bone and maxillary tooth landmarks.

After performing multidimensional scaling on the output from the persistent homology-based analysis (derived from the β_1 distances), two dimensions were found to be appropriate to explain the variability in the data, with the first displaying significant *time*, *group* and *time**group* effects. This dimension corresponded to maxillary expansion, and saw increases in the two treatment groups. The bone-anchored group increased at time point 2, and stayed high through time points 3 and 4, while the tooth-anchored group increased until peaking at time point 3, and then regressing slightly at time point 4. The differences between the treatment groups were never statistically significant however. The second MDS dimension was not statistically significant for *time*, *group* or *time**group* in the repeated measures MANOVA.

According to the orthodontists undertaking this study, the two treatment groups both saw maxillary expansion, whereas the control group did not. Clinically, they often notice that the tooth-anchored treatment pushes the teeth a bit further apart (sloping outwards), but that the teeth 'settle' back to a normal expanded position after appliance removal. Anecdotally, the bone-anchored subjects did not see this effect, and overall outcomes after appliance removal (i.e. at time point 4) were similar for both treatments. This extra expansion followed by slight relapse in the tooth-anchored group was in fact seen in the repeated measures MANOVA on the first MDS coordinate derived from the β_1 distances (recall Fig. 7 left), although differences between the treatment groups were never statistically significant.

6. Conclusions and future work

Methods incorporating persistent homology can be useful in analyzing landmark-based shape data. This special case, where specific points are matched across subjects and used as input to separate filtrations, is different than the typical set of point cloud data that is analyzed using persistent homology. Comparing the configurations by using a metric on the space of their persistence diagrams can glean useful information about the types of variability within the data set. By analyzing the distances between the β_0, β_1 and β_2 (or higher) persistence diagrams separately, different features of the configurations are visible. In some cases, if the structure of interest is known (as was the case in our example, with the U-shaped maxilla), then a choice can be made to analyze the set of persistence diagrams of the same dimension at the feature of interest. If the analysis is more exploratory, then each of dimensions will provide different information, and can help to discern the dimension of interesting features in the data or differences between groups.

The persistent homology-based analysis was most effective in distinguishing treatment effects when the β_1 persistence diagrams were used for the analysis. If more landmarks were used to represent the maxillary configuration, or a different type of filtration (such as a Čech) was used, then β_2 distances may have also been useful, but due to the reasonably small number of landmarks and the Rips filtration chosen, voids would appear less frequently than loops. β_0 persistence diagrams are likely most useful when there are areas that are spatially separate (or in the case of the filtration that we chose here, regions that are smaller than average separated by regions that are larger than average). The choice of filtration can also be tailored for different objectives, but the general theme is to fix a triangulation on all the configurations, perform persistent homology on each separately, and compare them using distances between the persistence diagrams (of appropriate dimension). The results obtained when the Delaunay triangulation was used in the persistent homology calculation appear to be richer than those obtained when the entire three-dimensional flag complex was used.

One area for future research would be in experimental and theoretical justification for which choices of filtration are most appropriate for different data types and objectives, and how the $\beta_0, \beta_1, \beta_2, \ldots$ distances can be used and combined to elucidate various types of data structures. The application to landmark configurations is one more way in which the exciting methods of persistent homology can be used to aid in the analysis of data, which warrants further exploration.

Acknowledgments

The authors would like to thank Drs. Paul Major and Manuel Lagravere for allowing the use of their orthodontic data set. They also thank the two anonymous referees who gave helpful comments which improved the clarity and presentation of the paper. Funding was provided by the University of Alberta Orthodontic Division's McIntyre Memorial Fund. The authors are also indebted to many of the participants at the Banff International Research Station workshop on Data Analysis using Computational Topology and Geometric Statistics for their insightful comments and ideas.

References

[1] I. Borg, J.F. Croenen, Modern Multidimensional Scaling, 2nd ed., Springer, 2005.
[2] P. Bubenik, G. Carlsson, P. Kim, Z. Luo, Statistical topology via morse theory, persistence, and nonparmetric estimation, Contemp. Math. (in press).
[3] P. Bubenik, P. Kim, A statistical approach to persistent homology, Homology Homotopy Appl. 9 (2) (2007) 337–362.
[4] G. Carlsson, Topology and data, Bull. Amer. Math. Soc. 46 (2) (2009) 255–308.
[5] M. Chung, P. Bubenik, P. Kim, Persistence diagrams of cortical surface data, in: LNCS: Proceedings of IPMI 2009, vol. 5636, 2009, pp. 386–397.
[6] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, Y. Mileyko, Lipschitz functions have L_p-stable persistence, Found. Comput. Math 10 (2) (2010) 127–139.
[7] M.-L. Dequéant, S. Ahnert, H. Edelsbrunner, T.M.A. Fink, E.F. Glynn, et al., Comparison of pattern detection methods in microarray time series of the segmentation clock, PLoS ONE 3 (8) (2008) e2856.
[8] V. de Silva, et al. Plex, online, 2000–2003. Available at: http://comptop.stanford.edu/programs/plex/.
[9] I.L. Dryden, K.V. Mardia, Statistical Shape Analysis, 1st ed., John Wiley and Sons, 1998.
[10] H. Edelsbrunner, J. Harer, Persistent homology—a survey, Contemp. Math. 453 (2008) 257–282.
[11] H. Edelsbrunner, D. Letscher, A. Zomorodian, Topological persistence and simplification, Discrete Comput. Geom. 28 (4) (2002) 511–533.
[12] R. Ghrist, Barcodes: the persistent topology of data, Bull. Amer. Math. Soc. 45 (1) (2007) 61–75.
[13] R. Ghrist, V. de Silva, Homological sensor networks, Notic. Amer. Math. Soc. 54 (1) (2007) 10–17.
[14] A. Hatcher, Algebraic Topology, 1st ed., Cambridge University Press, 2001.
[15] G. Heo, C. Small, Form representations and means for landmarks: a survey and comparative study, Comput. Vis. Image Underst. 102 (2) (2006) 188–203.
[16] S. Huckemann, T. Hotz, A. Munk, Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions, Statist. Sinica 20 (1) (2010) 1–100.
[17] S.R. Lele, J.T. Richtsmeier, An Invariant Approach to Statistical Analysis of Shapes, 1st ed., Chapman and Hall/CRC Press, 2001.
[18] L. Lovasz, M.D. Plummer, Matching Theory, 1st ed., Elsevier Science Publishers B.V., 1986.
[19] A. Sacan, O. Ozturk, H. Ferhatosmanoglu, Y. Wang, LFM-Pro: a tool for detecting significant local structural sites in proteins, Bioinform. 23 (6) (2007) 709–716.
[20] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.
[21] A. Zomorodian, Topology for Computing, 1st ed., Cambridge University Press, 2005.
[22] A. Zomorodian, G. Carlsson, Computing persistent homology, Discrete Comput. Geom. 33 (2) (2005) 249–274.