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.