Journal of Multivariate Analysis 146 (2016) 237–247

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

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

# Shape classification based on interpoint distance distributions

José R. Berrendero [a], Antonio Cuevas [a], Beatriz Pateiro-López [b,*]

[a] *Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain*

[b] *Departamento de Estadística e Investigación Operativa, Universidad de Santiago de Compostela, Spain*

ARTICLE INFO

*Article history:*
Received 24 January 2015
Available online 9 October 2015

*AMS subject classifications:*
62H30
62H35

*Keywords:*
Functional data
Identifiability
Interpoint distance
Shape analysis
Volume function

ABSTRACT

According to Kendall (1989), in shape theory, *The idea is to filter out effects resulting from translations, changes of scale and rotations and to declare that shape is "what is left"*. While this statement applies in principle to classical shape theory based on landmarks, the basic idea remains also when other approaches are used. For example, we might consider, for every shape, a suitable associated function which, to a large extent, could be used to characterize the shape. This finally leads to identify the shapes with the elements of a quotient space of sets in such a way that all the sets in the same equivalence class share the same identifying function. In this paper, we explore the use of the interpoint distance distribution (i.e. the distribution of the distance between two independent uniform points) for this purpose. This idea has been previously proposed by other authors [e.g., Osada et al. (2002), Bonetti and Pagano (2005)]. We aim at providing some additional mathematical support for the use of interpoint distances in this context. In particular, we show the Lipschitz continuity of the transformation taking every shape to its corresponding interpoint distance distribution. Also, we obtain a partial identifiability result showing that, under some geometrical restrictions, shapes with different planar area must have different interpoint distance distributions. Finally, we address practical aspects including a real data example on shape classification in marine biology.



## 1. Introduction

We are concerned here with the problem of classifying *shapes*, where, in informal terms, a shape is the family of all plane figures that can be obtained from a basic template figure (e.g., a square) by applying isometry transformations (rigid movements + symmetries) together with changes of scale. Also, we would like to include all the "deformed versions" (within some limits) of these basic elements, subject again to isometry transformations and/or scale changes. So, to mention just a very simple example, one could think that we want to automatically discriminate between two capital letters, say "B" and "D", manually drawn with a thick line marker, whatever their size or their orientation.

In marine biology, one might be interested on classifying fish species using shape analysis techniques. In some cases the basis for the recognition method is the fish image itself; see Storbeck and Daan [46]. Other researches have used the so-called *otoliths*, small pieces present in the inner ear of the fish, which can be considered as "microfossils" whose shapes are useful in species recognition, among other applications; see Lombarte et al. [36]. In Section 5 we will use this otolith example as an illustration for the methodology we propose.

* Correspondence to: Rúa Lope Gómez de Marzoa s/n, 15782 Santiago de Compostela, Spain.
*E-mail address:* beatriz.pateiro@usc.es (B. Pateiro-López).

http://dx.doi.org/10.1016/j.jmva.2015.09.017

Whatever the practical problem at hand, we need to define, in precise mathematical terms, what we mean for "shapes" in our setting. Then we will be ready to use the statistical methods for classification, either supervised (discrimination) or unsupervised (clustering) from the available dataset of shapes. In the example of Section 5 we will focus on clustering but discrimination methods could be considered as well.

The classical theory of shape analysis is largely based on the use of "landmarks" (i.e., finite vectors of coordinates characterizing the shapes). It was developed, to a large extent, by D. Kendall who expressively referred to shape analysis studies in the following terms: *The idea is to filter out effects resulting from translations, changes of scale and rotations and to declare that shape is "what is left"*; see Kendall [29]. A general perspective of this theory can be found in Kendall [29], Kendall et al. [30] or Kendall and Le [31].

We should mention however that other, less general, notions of shapes have been proposed. As Kent [33] points out, *"... statistical models for shapes may be based on underlying models for the landmarks themselves, or they may be constructed directly within shape space. In some special cases specialized models may be constructed"*. Our approach here could be understood as one of these specialized models: roughly speaking, we propose to identify a shape with the corresponding *interpoint distance distribution*, that is, the distribution of the distance (normalized to 1) between two randomly chosen points in the figure.

*Related literature*

In fact, the idea of using the interpoint distance distribution to identify the shapes has been previously proposed by other authors, with different applications in mind. For example, the very much cited paper by Osada et al. [43] explores the practical aspects of using the interpoint distance in the problem of discriminating shapes in image analysis. As these authors point out, *"The primary motivation for this approach is to reduce the shape matching problem to the comparison of probability distributions, which is simpler than traditional shape matching methods that require pose registration, feature correspondence, or model fitting. We find that the dissimilarities between sampled distributions of simple shape functions (e.g., the distance between two random points on a surface) provide a robust method for discriminating between classes of objects (e.g., cars versus airplanes) in a moderately sized database, despite the presence of arbitrary translations, rotations, scales, mirrors, tessellations, simplifications, and model degeneracies"*. See also Bonetti and Pagano [7] for a different use of interpoint distance distributions in the context of medical research.

In Kent [32] interpoint distances (between landmarks) are used, via multi-dimensional scaling, in shape analysis. Our approach here is somewhat different as it avoids the use of landmarks at the expense of some loss in generality.

Let us finally mention that the use of interpoint distance distributions entails the precise definition of a corresponding, suitable "space of shapes"; see Section 2, where the whole approach makes sense. Other related shape spaces can be found in the literature, in particular those based on "deformable templates": see [2,23,26,27].

*The purpose and contents of this paper*

On the theoretical side, we will provide some support for the use of interpoint distance distributions to characterize shapes: first, we relate, in Theorem 1, the distance between interpoint distance distributions with a natural, geometrically motivated, distance between shapes defined in Section 2. Second, we consider the problem of providing a sufficient condition on the sets in the Euclidean plane in order to ensure that two different sets fulfilling this condition must necessarily have different interpoint distance distributions. Theorem 2 provides a quite general identifiability criterion, which is in fact the most general result of this type we are aware of. In the Supplementary material section (see Appendix A) we also briefly consider the connection between the interpoint distance distribution and the covariogram (sometimes called "set covariance"), another popular function which has been used sometimes to characterize sets and shapes; see [10,11].

Finally, in Section 5 our methodology based on interpoint distance distributions is used in a problem of fishes otoliths classification, via hierarchical clustering.

## 2. The space of shapes

In what follows we will mainly focus on the case of shapes in the plane $\mathbb{R}^2$ (the most important, by far, in practical applications). However, some of the ideas we will develop can be also adapted to more general, multivariate cases. Our starting point will be the family $\mathcal{C}$ of compact non-empty sets in $\mathbb{R}^2$ with diameter 1; this means that $\text{diam}(C) = \max\{\|x-y\|,\ x, y \in C\} = 1$, for all $C \in \mathcal{C}$, where $\|\cdot\|$ stands for the Euclidean norm. We may think that the family $\mathcal{C}$ is the result of transforming the set of all possible plane images by a uniform change of scale (where "uniform" means that the same transformation scale is applied in both coordinates) in such a way that all of them have a common diameter. We will define our space of shapes as the quotient space obtained from a natural equivalence relation in $\mathcal{C}$. However, the family $\mathcal{C}$ is too large to work with (in particular, to define a meaningful, tractable distance between shapes). So we will need to restrict ourselves to a smaller subset $\mathcal{C}_1 \subset \mathcal{C}$ which, still, will include most "black-and-white" images arising in practical applications.

To be more specific, given two positive constants $a$ and $m_1$, we define $\mathcal{C}_1$ as the class of sets $C \in \mathcal{C}$ fulfilling the following conditions:

(i) $\mu(C) \geq a$, where $\mu$ denotes the Lebesgue measure in $\mathbb{R}^2$.

(ii) All the sets in $\mathcal{C}_1$ are regular, that is, every $C \in \mathcal{C}_1$ fulfills $C = \overline{\text{int}(C)}$.

(iii) $\mu(B(\partial C, \epsilon)) < m_1\epsilon,\ \forall \epsilon \in (0, 1]$.