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*

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.

Keywords: Functional data, Identifiability, Interpoint distance, Shape analysis, Volume function.

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,

[*] Corresponding address: Rúa Lope Gómez de Marzoa s/n. 15782 Santiago de Compostela. Spain

Email address: beatriz.pateiro@usc.es (Beatriz Pateiro-López)

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 (2001). Other researches have used the so-called *otholits*, 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. (2006). In Section 5 we will use this otolith example as an illustration for the methodology we propose.

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 data set 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 (1989). A general perspective of this theory can be found in Kendall (1989), Kendall et al. (1999) or Kendall and Le (2010).

We should mention however that other, less general, notions of shapes have been proposed. As Kent (1995) 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. (2002) 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 dissimi-*

larities 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 (2005) for a different use of interpoint distance distributions in the context of medical research.

In Kent (1994) 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 below, where the whole approach makes sense. Other related shape spaces can be found in the literature, in particular those based on "deformable templates": see Grenander (1976), Amit et al. (1991), Hobolt and Vedel-Jensen (2000), Hobolt et al. (2003).

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 below, 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 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 Cabo and Baddeley (1995, 2003).

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

Here ∂A denotes the topological boundary of the set A, $B(A, \epsilon)$ stands for the "parallel set" $B(A, \epsilon) = \{x : d(x, A) \leq \epsilon\}$ and $d(x, A) = \inf\{\|x - y\|, \ y \in A\}$ (when $A = \{x\}$ we will use the standard notation $B(x, \epsilon)$ instead of $B(\{x\}, \epsilon)$).

We assume that the space $\mathcal{C}_1$ is endowed with the metric,

$$d_{HH}(C, D) = d_H(C, D) + d_H(\partial C, \partial D),$$

where d_H stands for the ordinary Hausdorff metric between compact sets.

Let us now define on $\mathcal{C}_1$ the *isometry* equivalence relation: we will say that $C, D \in \mathcal{C}_1$ are *isometric* (and denote it by $C \sim D$) when there exists a isometry (i.e., a map $i : \mathbb{R}^2 \to \mathbb{R}^2$ satisfying $\|i(x) - i(y)\| = \|x - y\|$) such that $i(C) = D$. The family of all sets in $\mathcal{C}_1$ equivalent to a set C will be represented by $[C]$.

Finally, denote by $\mathcal{S}$ the family of equivalence classes and define in $\mathcal{S}$ the *quotient metric*, $\tilde{d}_{HH}$, using the standard definition method [see, for example, Burago et al. (2001, p. 62)],

$$\tilde{d}_{HH}([C], [D]) = \inf\{\sum_{i=1}^{n} d_{HH}(P_i, Q_i) : \ [P_1] = [C], \ [Q_n] = [D], \ n \in \mathbb{N}\}, \quad (1)$$

where the infimum is taken on all finite sequences such that $[Q_i] = [P_{i+1}]$ for $i = 1, \ldots, n-1$. In principle, the general method (1) to translate a metric to the quotient space defines only a semi-metric, but we will see below that in this case it provides a true metric; in fact, we will also see in Proposition 1 that (1) can be expressed in a much simpler way in our case.

The elements of the quotient metric space $\mathcal{S}$ will be called *shapes.* So the shapes are in fact classes of equivalence $[C]$ for $C \in \mathcal{C}_1$.

Some motivation

Regarding the intuitive meaning of the assumptions imposed on $\mathcal{C}_1$, let us note that they do not entail any serious restriction for the practical classification problems of pattern recognition. To explain the meaning of these assumptions let us identify our shapes with figures drawn with a sign painting marker:

Assumption (i) just states that, after re-scaling, our shapes must have a minimum "thickness", expressed in a minimum "drawing area" a.

Condition (ii) is usual in geometric probability models. Under this assumption, the set C cannot consist of a closed "central core" C_1 plus some "superfluous" parts P (such as rays or isolated points) with $\mu(P) = 0$.

Condition (iii) rules out involved drawings, with a very large boundary. To see this, let us briefly recall the notion of *(boundary) Minkowski content*, which is perhaps the simplest way (among several others, see e.g. Mattila (1995)) to define the "boundary measure" of a set $C \subset \mathbb{R}^d$. Of course, for the two-dimensional case, "boundary measure" is synonymous with "length perimeter". In precise terms, the $(d-1)$-dimensional (boundary) Minkowski content of C is defined by the limit

$$L_0(C) = \lim_{\epsilon \to 0} \frac{\mu(B(\partial C, \epsilon)}{2\epsilon}, \tag{2}$$

A closely related notion is the *one-sided (outer) Minkowski content*, defined by

$$L_0^+(C) = \lim_{\epsilon \to 0} \frac{\mu(B(C, \epsilon) \setminus C)}{\epsilon}, \tag{3}$$

See Ambrosio et al. (2008) for a comprehensive study of this notion, including conditions under which $L_0(C) = L_0^+(C)$. For statistical aspects related to the Minkowski content we refer to Cuevas et al. (2007) and Berrendero et al. (2014). Note that under condition (iii), $L_0(C) \leq m_1$ for all $C \in \mathcal{C}_1$.

A simpler, alternative expression for the distance between shapes.

While (1) gives the "canonical" expression for the distance in a quotient metric space, the effective calculation of this metric looks rather troublesome. The following proposition provides a simpler, more natural expression for (1) and shows that $\tilde{d}_{HH}$ is in fact a metric instead of just a semi-metric: this means that $\tilde{d}_{HH}([C], [D]) = 0$ implies $[C] = [D]$.

Proposition 1. *The semi-metric (1) can be expressed as*

$$\tilde{d}_{HH}([C], [D]) = \inf\{d_{HH}(C', D') : \ C' \in [C], \ D' \in [D]\}. \tag{4}$$

Moreover, this expression defines in fact a true metric.

Proof. This result follows from Th. 2.1 in Cagliari et al. (2014). In part (i) of this theorem it is proved that a expression of type (4) holds for the semi-distance (1) in the quotient space whenever the equivalence classes of this space are the orbits of the action of a group of isometries. This is the case here.

The fact that expression (1), or (4), defines a true metric is a consequence of conclusion (iv) in the aforementioned theorem where the authors prove that (4) is a metric if and only if the orbits of the action are closed sets. To see that $[C]$ is a closed set let us consider a convergent sequence $\{C_n\}$ of elements $C_n \in [C]$ with $n \geq 1$; denote by C_0 the limit, i.e., $d_{HH}(C_n, C_0) \to 0$. By definition of $[C]$, any C_n can be obtained as $C_n = t_n(C)$, where t_n is an isometry. Since $\|t_n(x) - t_n(y)\| = \|x - y\|$, it turns out that the sequence $\{t_n\}$ is equicontinuous; moreover, for each $x \in \mathbb{R}^2$ the sequence $\{t_n(x)\}$ is bounded; this is clearly true when $x \in C$, since the sequence $C_n = t_n(C)$ is d_H-convergent. Then, for a general $x \in \mathbb{R}^2$, $\{t_n(x)\}$ is also bounded (since, given $x_0 \in C$, $\|t_n(x) - t_n(x_0)\| = \|x - x_0\|$). So, from Ascoli-Arzelà Theorem [e.g., Folland (1999, p. 137)] we can ensure that there exists a subsequence of $\{t_n\}$, denoted again $\{t_n\}$, such that $t_n \to t$, uniformly on compacts, for some transformation t, which must be necessarily an isometry. We thus have $d_H(t_n(C), t(C)) \to 0$, but, since $t_n(C) = C_n$ and $d_H(C_n, C_0) \to 0$, we get $C_0 = t(C)$. Finally to see $C_0 \in [C]$ we only have to prove that C_0 fulfills conditions (i), (ii) and (iii) stated above in the definition of the class $\mathcal{C}_1$. But this a trivial consequence of the *Classification Theorem for Isometries on the Plane* [see, for example, Martin (1982, p. 65)] which states that each non-identity isometry on the plane is either a translation, a rotation, a reflection, or a glide-reflection (i.e., the composition of a reflection and a translation in the direction of the reflection axis). This shows that the plane isometries are "measure preserving" (i.e., $\mu(A) = \mu(t(A)))$ and "boundary preserving" (i.e., $\partial t(C) = t(\partial C)$ and therefore, (i)-(iii) hold also for $t(C) = C_0$. We conclude that $[C]$ is closed. □

3. The interpoint distance distribution

As mentioned in the introduction, our approach is based on eventually identifying a shape $[C]$ with a density function, supported on $[0, 1]$. This is the density function of the distribution of the random variable defined as the distance between two points randomly chosen on C.

To be more precise, for each $C \in \mathcal{C}_1$, define the random variable

$$Y_C = \|X_1 - X_2\|, \tag{5}$$

where X_1, X_2 are iid random variables uniformly distributed on C. It is readily seen that Y_C is absolutely continuous with respect to the Lebesgue measure μ. Let us denote by f_C the density function of Y_C.

Theorem 1 below provides a partial mathematical motivation for the identification $[C] \simeq f_C$ by showing that the transformation $[C] \mapsto f_C$ is continuous (in fact it is Lipschitz), so that if two shapes are close enough then the corresponding interpoint distance densities must be also close together. The problem of analyzing to what extent f_C is helpful in order to identify C will be discussed in Section 4.

The Lipschitz property of the transformation $C \mapsto f_C$ will be established with respect to the standard L_1 metric between densities and also for the so-called Wasserstein (or Kantorovich) metric defined, for two cumulative distribution functions on the real line F and G, by

$$d_W(F,G) = \int_{\mathbb{R}} |F(x) - G(x)|dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt,$$

where F^{-1}, G^{-1} denote the corresponding quantile functions. This metric has a number of interesting properties and applications. It has been sometimes called "the earth mover distance", due to its connections with the transportation problem; see Villani, C. (2003). In Rubner et al. (2000) and Ling and Okada (2007) can be found some details on the use of this distance in image retrieval. Of course, when F and G are absolutely continuous (as it will always be the case in what follows), d_W can also be interpreted as a distance between the density functions.

The following result can be seen as a statement of "compatibility" between the distances $d_1(f,g) = \int_0^1 |f - g|d\mu$ or d_W (defined in the space of densities on $[0,1]$) and the "natural" distance $\tilde{d}_{HH}$ defined in our space of shapes. The whole point is to replace, in practice, the use of $\tilde{d}_{HH}$ (whose effective calculation is cumbersome) by the more convenient distances d_1 or d_W. In principle, the intuitive interpretation of $d_1(f,g)$ (as the area of the region between f and g) is perhaps more direct but, as we have already mentioned, d_W is also used in image analysis, Rubner et al. (2000). Our experimental results, see Section 5 and the Supplementary Material document, show a very similar behaviour for both distances with perhaps a slightly better performance for d_1.

Theorem 1. *Let $\mathcal{D}$ be the space of probability density functions (with respect to the Lebesgue measure) on $[0,1]$. Then*

(a) The transformation $T : \mathcal{C}_1 \to \mathcal{D}$ given by $T(C) = f_C$ fulfills the Lipschitz condition with respect to the L_1 metric, that is, $d_1(f_C, f_D) \leq m d_{HH}(C,D)$, for some constant $m > 0$.

(b) Also, if we denote by F_C and F_D the cumulative distribution functions of Y_C and Y_D, respectively, we have that $d_W(F_C, F_D) \leq \frac{m}{2} d_{HH}(C,D)$, where m is the same constant of statement (a).

(c) The transformation T induces another transformation $\tilde{T}([C]) = f_C$, defined in the quotient space, which is also Lipschitz, with constants m and $m/2$ respectively, for both considered metrics.

Proof. (a) From the relation between the L_1 metric and the total variation distance,

$$\int |f_C - f_D| d\mu = 2 \sup_A |P_C(A) - P_D(A)|, \tag{6}$$

where P_C and P_D are the probability measures associated with f_C and f_D and the supremum is taken on $\mathcal{B} = \mathcal{B}([0,1])$, the Borel sets of $[0,1]$ on the elements C, D chosen to represent $[C]$ and $[D]$. Now, observe that for all $A \in \mathcal{B}$, and using the notation introduced in expression (5),

$$\begin{aligned}&P_C(A) = \mathbb{P}(Y_C \in A) = \mathbb{P}(Y_C \in A|X_1, X_2 \in C \cap D)\mathbb{P}(X_1, X_2 \in C \cap D)\\ &+\mathbb{P}(Y_C \in A|X_1 \text{ or } X_2 \notin C \cap D)\mathbb{P}(X_1 \text{ or } X_2 \notin C \cap D),\end{aligned}$$

where X_1, X_2 are iid uniformly distributed on C. A similar expression holds for $P_D(A)$, except that C is replaced with D and X_1, X_2 are replaced with X_1^*, X_2^*, iid uniform on D, that is,

$$\begin{aligned}&P_D(A) = \mathbb{P}(Y_D \in A) = \mathbb{P}(Y_D \in A|X_1^*, X_2^* \in C \cap D)\mathbb{P}(X_1^*, X_2^* \in C \cap D)\\ &+\mathbb{P}(Y_D \in A|X_1^* \text{ or } X_2^* \notin C \cap D)\mathbb{P}(X_1^* \text{ or } X_2^* \notin C \cap D),\end{aligned}$$

Note that $\mathbb{P}(Y_C \in A|X_1, X_2 \in C \cap D) = \mathbb{P}(Y_D \in A|X_1^*, X_2^* \in C \cap D)$. Therefore,

$$\begin{aligned}|P_C(A) - P_D(A)| &\leq \mathbb{P}(Y_C \in A|X_1, X_2 \in C \cap D)\mathbb{P}(X_1 \text{ or } X_2 \notin C \cap D)\\ &+ \mathbb{P}(Y_C \in A|X_1, X_2 \in C \cap D)\mathbb{P}(X_1^* \text{ or } X_2^* \notin C \cap D)\\ &+ \mathbb{P}(Y_C \in A|X_1 \text{ or } X_2 \notin C \cap D)\mathbb{P}(X_1 \text{ or } X_2 \notin C \cap D)\\ &+ \mathbb{P}(Y_D \in A|X_1^* \text{ or } X_2^* \notin C \cap D)\mathbb{P}(X_1^* \text{ or } X_2^* \notin C \cap D).\end{aligned}$$

For the first term in the right-hand side of $|P_C(A) - P_D(A)|$ we have,

$$\begin{aligned}&\mathbb{P}(Y_C \in A|X_1, X_2 \in C \cap D)\mathbb{P}(X_1 \text{ or } X_2 \notin C \cap D)\\ &\leq \mathbb{P}(X_1 \text{ or } X_2 \in C \setminus D) \leq 2\mathbb{P}(X_1 \in C \setminus D) \leq \frac{2}{a}\mu(C \setminus D),\end{aligned}$$

where a is the minimal area of the elements of $\mathcal{C}$ defined in condition (i). The same holds for the third term. Similarly, we have that the second and fourth terms in $|P_C(A) - P_D(A)|$ are smaller than $\frac{2}{a}\mu(D \setminus C)$. Hence,

$$\sup_A |P_C(A) - P_D(A)| \leq \frac{4}{a}\mu(C\Delta D), \tag{7}$$

where $C\Delta D$ stands for the symmetric difference $C\Delta D = (C \setminus D) \cup (D \setminus C)$.

Let us now prove that

$$\mu(C\Delta D) \leq 2m_1 d_{HH}(C, D), \tag{8}$$

where m_1 is the constant introduced in the definition on $\mathcal{C}_1$. To see this, put $d_{HH}(C, D) = r$ and take $x \in C \setminus D$. We must have $x \in B(D, r) \setminus D$ which entails $x \in B(\partial D, r) \subset B(\partial C, 2r)$. Similarly, if $x \in D \setminus C$ we have $x \in B(C, r) \setminus C$ so that $x \in B(\partial C, r)$.

Thus, using assumption (iii) we have obtained that

$$\mu(C\Delta D) \leq \mu(B(\partial C, 2r)) \leq 2m_1 r = 2m_1 d_{HH}(C, D).$$

This, together with (6), (7) and (8) proves the first statement (a).

(b) This directly follows from Theorem 4 in Gibbs and Su (2002). According to this result, if we consider probability measures defined on a space Ω with finite diameter, $\text{diam}(\Omega)$, we have $d_W \leq \text{diam}(\Omega) \cdot d_{TV}$. In our case, all the considered distributions are defined on the unit interval. This, together with $2d_{TV} = d_1$ leads to statement (b).

(c) This statement follows from parts (a) and (b) combined with the expression (4) of the quotient metric. □

Remark 1. *The search for a Lipschitz-type as that in Theorem 1 is quite natural in those situations where a set (or a shape) is replaced with a more convenient auxiliary function. For example, a result in a similar spirit can be found in Cabo and Baddeley (1995, Th. 5.4) but these authors consider the so-called covariogram function*, instead of the interpoint distance density, and the distance d_{HH} is replaced with another metric defined in terms of the so-called "linear scan transform".

The covariogram of a bounded Borel set $A \subset \mathbb{R}^d$ is defined by $K_A(y) = \mu(A \cap T_y A)$, where $y \in \mathbb{R}^d$, $T_y A = A - y = \{a - y : a \in A\}$ and μ is the Lebesgue measure in $\mathbb{R}^d$. This function is useful in different problems of stochastic geometry and stereology. Some references are Cabo and Baddeley (1995, 2003) and Galerne (2011). Using some results in these papers it is easy to prove (see the Supplementary Material document for details) that the random interpoint distance Y_C of a bounded Borel set C in the plane has a continuous density f_C with $f_C(0) = 0$ and $f_C(\rho_C) = 0$, where $\rho_C = \text{diam}(C)$.

4. The identifiability problem

In order to implement the idea of identifying a shape $[C]$ with the corresponding interpoint distance density f_C, we must still overcome a further obstacle. Even if we restrict to the space of shapes $[C]$ with $C \in \mathcal{C}_1$ (where the continuity of the transformation $[C] \mapsto f_C$ is warranted) one might have that $f_C = f_D$ for $[C] \neq [D]$. This follows as a consequence of a counterexample, due to Mallows and Clark (1970) [inspired by a question posed by Blaschke], showing two non-congruent polygons, C and D with the same *chord length* distribution. The chord length is the length of the segment intercepted in C by a random chord. Since the chord length distribution determines uniquely the interpoint distance distribution [see, Matern (1986, p. 25)] the mentioned counterexample applies also to the interpoint distance distribution.

The interpoint distance has been also used (with applications to crystallography and DNA mapping) in finite sets of points; see Caelli (1980) and Lemke et al. (2003) for further counterexamples, references and insights.

Thus, in summary, the interpoint distance distribution has not full capacity to discriminate shapes. Hence, we should further restrict our shape space to those sets $[C]$ such that C lives in an appropriate subset $\mathcal{C}_2 \subset \mathcal{C}_1$ fulfilling the identifiability condition

$$\text{(iv) For all } C,\ D \in \mathcal{C}_2 \text{ with } [C] \neq [D] \text{ we have } Y_C \overset{d}{\neq} Y_D, \tag{9}$$

where Y_C and Y_D denote the interpoint distances (5) on C and D and the notation $\overset{d}{\neq}$ means that both variables are not identically distributed.

Some identifiability problems similar to (9) have been considered in the stochastic geometry literature under different conditions. For example, Matheron (1986) formulated the following conjecture: *Every planar convex body is determined within all planar convex bodies by its covariogram, up to translations and reflections.* This conjecture was completely solved, in the affirmative by Averkov and Bianchi (2009).

In the following subsection we will show that the analogous problem (9) for the interpoint distance distribution can be solved under quite general conditions, which do not require convexity.

4.1. Interpoint distances and polynomial area

The main geometric assumption we will use to guarantee identifiability is defined as follows.

Definition 1. *A set $C \subset \mathbb{R}^2$ is said to have inner polynomial area* if there exist constant $R = R(C) > 0$ and $L = L(C) > 0$ such that

$$\mu(I_r(C)) = \mu(C) - L(C)r + \pi r^2, \text{ for } 0 \leq r < R, \tag{10}$$

where $I_r(C)$ denotes the *inner parallel set* $I_r(C) = \{x \in C : B(x, r) \subset C\}$.

For example, the circle $C = B(0, m)$ fulfills (10) with $L(C) = 2\pi m$, $R < m$ and $\mu(C) = \pi m^2$.

Remark 2. *It is clear that, if (10) holds, the quantity $L(C)$ could be obtained as a sort of inner Minkowski content,* $L_0^-(C)$ defined in a similar way to outer version $L_0^+(C)$ given in (3). Moreover, if the ordinary (two-sided) Minkowski content, $L_0(C)$ does exist [see (2)] then condition (10) clearly entails $L(C) = L_0(C) = L_0^+(C)$.

Now, our goal is to motivate this definition in a twofold way. First, we will relate it to some relevant mathematical concepts. Second, we will exhibit a broad class of sets satisfying (10). For this purpose, it will be useful to recall some notions, due to Federer (1959), from geometric measure theory: the *reach* of a closed set is defined as the supremum, reach(C), of those values such that any point x whose distance to C is smaller than reach(C) has only one closest point on C. This concept leads to a valuable generalization of the notion of convex set, which can be interpreted also as a geometric smoothness condition (not directly relying on differentiability assumptions). Figure 1 illustrates the nice intuitive meaning of this notion. It can be shown that C is convex if and only if $\operatorname{reach}(C) = \infty$. According to a result proved by Federer (1959) [which is a generalization of the classical Steiner's formula for convex sets], the sets of positive reach have a polynomial volume. More precisely [Federer (1959), Ths. 5.6 and 5.19]:

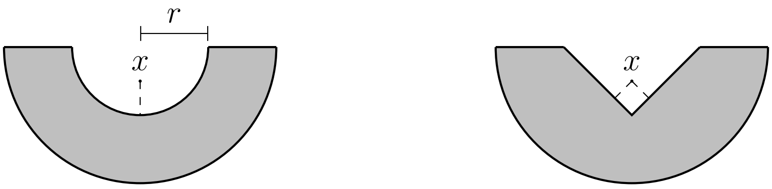

Figure 1: The set C in the left has positive reach r (any x whose distance to C is smaller than r has only one closest point on C). The set C in the right has not positive reach.

If $S \subset \mathbb{R}^d$ is a compact set with $r_0 = \operatorname{reach}(S) > 0$, then there exist unique values $\Phi_0(S), \ldots, \Phi_d(S)$ over such that

$$\mu(B(S,r)) = \sum_{i=0}^{d} r^{d-i}\omega_{d-i}\Phi_i(S), \textit{ for } 0 \leq r < r_0, \tag{11}$$

where ω_j is the j-dimensional measure of a unit ball in $\mathbb{R}^j$.

Remark 3. *The above result has some connections with other important geometric notions. Some are almost immediate: for example, if S is a compact set with positive reach, then $\Phi_d(S) = \mu(S)$ and the outer Minkowski content defined in (2) always exists and corresponds to the first-degree term in (11). Another, not so obvious, deep geometric connection of (11) is as follows: the coefficient $\Phi_0(S)$ coincides with the Euler characteristic of S. This is an integer-valued topological invariant with deep geometric implications, far beyond the scope of this paper; see, e.g., Hatcher (2002) for details. In the following remark we show an example which, in addition to recall the intuitive meaning of $\Phi_0(S)$, will also serve for further generalizations.*

On the other hand, note that $\operatorname{reach}(S) = r_0 > 0$ *is just a sufficient condition for polynomial volume in the interval $[0, r_0)$. Many other sets, which do not satisfy* $\operatorname{reach}(S) > 0$ *(such as that of the right panel in Figure 1), might fulfill a polynomial volume property of type (11).*

Remark 4. *Let us consider the annulus $D = B(0,M) \setminus \operatorname{int}(B(0,m))$, with $m < M$. A direct calculation shows that $\mu(B(D,r)) = 2\pi(M+m)r + \pi(M^2 - m^2)$. Moreover, it is clear that* $\operatorname{reach}(D) = m$*. As a conclusion, the annulus D fulfills $\Phi_0(D) = 0$ in (11). By the way, the same holds for any set, of positive reach, homeomorphic to the annulus (as the Euler characteristic is a topological invariant).*

Now, we are ready to show that in fact (10) applies to a broad class of sets under a quite general condition (expressed in terms of the classical positive reach property).

Proposition 2. *The class $\mathcal{P}(R)$ of sets which fulfill condition (10) contains all regular sets C such that for some closed ball B_1, with $C \subset \operatorname{int}(B_1)$, the set $E = B_1 \setminus \operatorname{int}(C)$ has positive reach R and it is homeomorphic to an annulus (as that considered in Remark 4).*

Proof. Note that $\mu(B(E,r)) = \mu(E)+\mu(B(B_1,r))-\mu(B_1)+\mu(C)-\mu(I_r(C))$. Now, E has positive reach R and, by (11), $\mu(B(E,r)) = rL_0^+(E) + \mu(E)$. Note also that $\Phi_0(E) = 0$ since $B_1 \setminus \operatorname{int}(C)$ is homeomorphic to an annulus D (for which $\Phi_0(D) = 0$, according to Remark 4). Therefore,

$$\mu(I_r(C)) = \mu(C) - L(C)r + \pi r^2, \text{ with } L(C) = L_0^+(E) - L_0(B_1).$$

□

As a conclusion, we have that the class of sets fulfilling (10) includes many relevant sets found in practice. See Berrendero et al. (2014) for further information and statistical applications of the notion of polynomial volume.

We are now ready to establish the main result of this section which provides a large class $\mathcal{R}$ of sets which can be distinguished from each other according to the distribution of the respective interpoint distances. In other words, if $C, D \in \mathcal{R}$ then $f_C \neq f_D$, where f_C denotes the density function of the interpoint distance Y_C.

Theorem 2. *(a) Suppose that C is a compact set in $\mathbb{R}^2$ fulfilling condition (10) of inner polynomial area. Denote by Y_C the interpoint distance in C. Then*

$$\mathbb{P}(Y_C \leq \rho) = \frac{\pi\rho^2}{\mu(C)} - \frac{\pi\rho^3 L(C)}{\mu(C)^2} + \frac{\pi^2\rho^4}{\mu(C)^2} + \frac{1}{\mu(C)^2}\int_{C\setminus I_\rho(C)} \mu(B(x,\rho)\cap C)dx, \quad (12)$$

for $\rho > 0$ be small enough so that $\rho < R$ in (10) and $I_\rho(C) \neq \emptyset$, where $I_\rho(C)$ denotes the inner parallel set $I_\rho(C) = \{x \in C : B(x,\rho) \subset C\}$.

(b) Let C, D be compact sets, with diameter 1, in $\mathbb{R}^2$ fulfilling the polynomial inner area condition (10). If $\mu(C) \neq \mu(D)$, then the respective interpoint distance have different distributions, that is, $Y_C \overset{d}{\neq} Y_D$.

Proof. (a) Let X_1, X_2 bee iid random variables uniformly distributed on C. Denote by P_C the probability distribution uniform on C.

$$\begin{aligned}
\mathbb{P}(Y_C \leq \rho) &= \int_C \mathbb{P}\left(X_1 \in B(x,\rho)\right) dP_C(x) = \int_C P_C(B(x,\rho))dP_C(x) \\
&= \int_{I_\rho(C)} P_C(B(x,\rho))dP_C(x) + \int_{C\setminus I_\rho(C)} P_C(B(x,\rho))dP_C(x) \\
&= \frac{1}{\mu(C)^2}\int_{I_\rho(C)} \mu(B(x,\rho))dx + \frac{1}{\mu(C)^2}\int_{C\setminus I_\rho(C)} \mu(B(x,\rho)\cap C)dx
\end{aligned}$$

$$
\begin{aligned}
&= \pi\rho^2 \frac{\mu(I_\rho(C))}{\mu(C)^2} + \frac{1}{\mu(C)^2} \int_{C \setminus I_\rho(C)} \mu(B(x,\rho) \cap C) dx \\
&= \pi\rho^2 \frac{\mu(C) - L(C)\rho + \pi\rho^2}{\mu(C)^2} + \frac{1}{\mu(C)^2} \int_{C \setminus I_\rho(C)} \mu(B(x,\rho) \cap C) dx \\
&= \frac{\pi\rho^2}{\mu(C)} - \frac{\pi\rho^3 L(C)}{\mu(C)^2} + \frac{\pi^2\rho^4}{\mu(C)^2} + \frac{1}{\mu(C)^2} \int_{C \setminus I_\rho(C)} \mu(B(x,\rho) \cap C) dx
\end{aligned}
$$

(b) This result readily follows from (a). First note that the integral $\int_{C \setminus I_\rho(C)} \mu(B(x,\rho) \cap C)dx$ in the last term of (12) is of order ρ^3 as $\rho \to 0$ since the integrand is of type $O(\rho^2)$ and the measure of the integration set is $O(\rho)$, from the polynomial area assumption. Therefore the main term in (12) is $\frac{\pi\rho^2}{\mu(C)}$. Now, If $\mu(C) \neq \mu(D)$, the main terms $\frac{\pi\rho^2}{\mu(C)}$ in the respective expressions (12) for the distribution functions of Y_C and Y_D are different. Hence, these distribution functions must be different for ρ small enough. □

5. An application to fish family identification from otolith images

The AFORO database (http://www.icm.csic.es/aforo/) offers an open online catalogue of fish otolith images. As defined by Tuset et al. (2008), otoliths are "acellular concretions of calcium carbonate and other inorganic salts that develop over a protein matrix in the inner ear of vertebrates". The application of otoliths research has developed significantly over the last years, see Begg et al. (2005). Fish species identification, age and growth determination or stock and hatchery management are some of the most common and important applications of otolith data.

The AFORO database contains at present more than 4500 high resolution images corresponding to 1382 species and 216 families from the Mediterranean Sea and the Antarctic, Atlantic, Indic and Pacific Oceans. For this study, we have considered fishes belonging to three families: *Soleidae*, *Labridae* and *Scombridae*. There are important features of otoliths that can be used for species identification. The otolith shape (outline), the inner groove and the otolith margins, among others, are important characteristics in the morphological description of otoliths. According to the characterization in Tuset et al. (2008), the terms that better describe the shape of the otolith's outline in the family *Soleidae* are discoidal, elliptic and bullet-shaped (and intermediate shapes between these three). For the family *Labridae*, the otolith's outlines are mainly cuneiform, oval and rectangular (and intermediate shapes). For the family *Scombridae*, the otoliths are characterized by their serrate margins. See Figure 2 for examples of otoliths from these three families.

Interpoint distance: estimated distribution and density functions. We have 240 high resolution images of otoliths and their corresponding contours (70 *Soleidae*, 125 *Labridae* and 45 *Scombridae*). For the practical implementation of the method in this example, we need to generate pairs of uniform

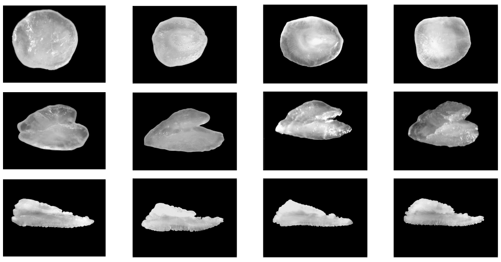

Figure 2: High resolution images of otoliths. First row: *Soleidae.* Second row: *Labridae.* Third row: *Scombridae.*

points within the otholiths (area in black in the filled-in contour images, see Supplementary material). For this purpose, we can use the standard acceptance-rejection method, generating uniform points on a rectangle containing the otolith and accepting those points belonging to the black area. This procedure will be slow on images with a small percentage of black pixels with respect to the bounding rectangle. Another possibility, faster than the acceptance-rejection method, is to select pixels in black randomly and, for each pixel, generate a uniformly distributed random point within that pixel. Other issues about sampling generation in more general situations, such as 3D shapes, are discussed in Osada et al. (2002). For each otolith, we compute the empirical cumulative distribution function of the interpoint distance using the distances (rescaled by the estimated diameter) between 50000 pairs of random points on the otolith. Figure 3 shows the empirical cumulative distribution functions (left) and the estimated interpoint distance densities (right) corresponding to the 240 otoliths (*Soleidae, Labridae and Scombridae in dark, medium and light gray, respectively).*

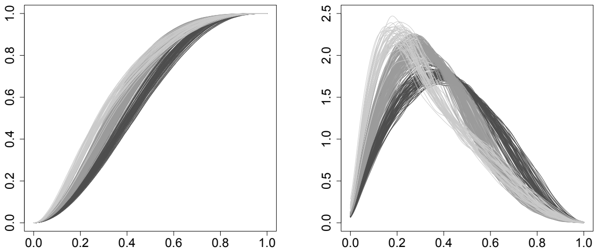

Figure 3: Left, empirical distribution functions of the interpoint distance on the otoliths. Right, estimated densities. In dark gray, *Soleidae.* In medium gray, *Labridae.* In light gray, *Scombridae.*

Hierarchical clustering. First, we apply an agglomerative hierarchical clustering procedure for each pair of families, considering both the L_1 distance between densities and the Wasserstein distance between cumulative distribution functions as the dissimilarity criterion. As linkage method, we have considered single-linkage, complete-linkage and average-linkage. For the sake of brevity, we only discuss here the average-linkage method, which gives the best results.

Let us first discuss the results on the dataset consisting of *Soleidae* and *Labridae* otoliths (dataset A). Figure 4 shows the dendrogram based on the L_1 distance between the estimated densities. We can consider the otoliths divided in two big groups (represented in dark and light gray). We observe, see Table 1 (left), that one cluster is dominated by *Soleidae* otoliths (94.29% of *Soleidae* otoliths belong to cluster 1) and the other contains mainly *Labridae* otoliths (98.40% of *Labridae* otoliths belong to cluster 2). The results of the clustering procedure based on the Wasserstein distance between distribution functions are quite similar, see Table 1 (right).

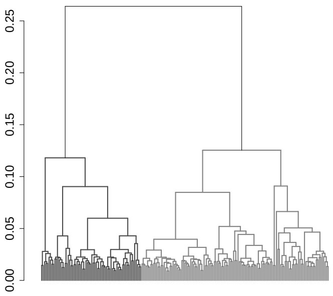

Figure 4: Dendrogram using the L_1 distance between interpoint distance densities for the dataset consisting of *Soleidae* and *Labridae* otoliths (dataset A). The tree is cut into two groups, represented in dark and light gray.

Table 1: Hierarchical clustering on three datasets of otoliths. For each dissimilarity criterion, count and row percent of the true family labels versus the group labels for a partition into two clusters.

		L_1 distance		Wasserstein distance	
		Cluster 1	Cluster 2	Cluster 1	Cluster 2
Dataset A	*Soleidae*	66	4	67	3
		94.29%	5.71%	95.71%	4.29%
	Labridae	2	123	2	123
		1.60%	98.40%	1.60%	98.40%
Dataset B	*Soleidae*	69	1	69	1
		98.57%	1.43%	98.57%	1.43%
	Scombridae	0	45	0	45
		0.00%	100.00%	0.00%	100.00%
Dataset C	*Labridae*	123	2	123	2
		98.40%	1.60%	98.40%	1.60%
	Scombridae	2	43	2	43
		4.44%	95.56%	4.44%	95.56%

Now, let us consider the dataset consisting of *Soleidae* and *Scombridae* otoliths (dataset B). We apply again an agglomerative hierarchical clustering procedure using both the L_1 distance and the Wasserstein distance as the dissimilarity criterion. We split the corresponding dendrograms into two groups. The results are summarized in Table 1 (dataset B). We found that all but one of the *Soleidae* otoliths belong to the first cluster and all the *Scombridae* otoliths belong to the other cluster.

Finally, we consider the complete dataset consisting of otoliths from the three families and apply the agglomerative hierarchical clustering procedure using the L_1 distance. If we cut the corresponding tree into three groups, we obtain that 94.29% of *Soleidae* otoliths belong to the first cluster, 96.80% of *Labridae* otoliths belong to the second cluster and 95.56% of *Scombridae* otoliths belong to the third cluster. The dendrogram and the complete table of results based on the L_1 distance and the Wasserstein distance can be found in the Supplementary Material.

k-means clustering. Now, we investigate the performance of the k-means clustering algorithm. We apply the k-means algorithm to each pair of families of otoliths ($k = 2$). Here we briefly describe the results based on the L_1 distance (the complete table of results based on the L_1 distance and the Wasserstein distance is provided as Supplementary Material). For the dataset consisting of *Soleidae* and *Labridae* images, we obtain a 96.92% of correctly clustered otoliths. For the dataset consisting of *Soleidae* and *Scombridae* images, we obtain a 99.13% of correctly clustered otoliths. For the dataset consisting of *Labridae* and *Scombridae* images, we obtain a 97.64% of correctly clustered otoliths.

Final remarks. (a) We observe that both clustering methods (hierachical clustering and k-means) perform reasonably well.

We would also like to note that the main reason to choose the families *Soleidae*, *Labridae* and *Scombridae* was that the AFORO database contains a large number of images of each of these families. At the beginning of the study, we had also considered two other large families: *Gobiidae* and *Serranidae* (see the Supplementary Material for examples of otoliths in these two families). As might be expected, the clustering methods did not perform well, for example, for the dataset containing *Gobiidae* and *Soleidae* otoliths since the shape of some of the *Gobiidae* otoliths resembles that of the *Soleidae* otoliths. The same occurs for the dataset containing *Serranidae* and *Labridae* otoliths.

(b) As a referee pointed out to us, the use of interpoint distance distributions can be extended to more general (not necessarily planar) situations. Thus, otholits are in fact three-dimensional structures, one might consider also the 3D extension of our technique. Likewise, one might think of incorporating possibly non-uniform choices of the random points defining the interpoint distances. This would entail additional theoretical and computational challenges; see Tebaldi et al. (2011) for computational aspects related to interpoint distance distributions.

6. Discussion. Connections with FDA

The study of those problems where the "sample elements" and/or the target "parameters" are members of an infinite-dimensional space is today a mainstream topic in statistical research. Of course, the classical nonparametric curve estimation theory (developed since the 1960's) is an important precedent but perhaps the excellent book by Grenander (1981) is one of the pioneering references in putting together these ideas in a more or less systematic fashion. As it often happens in the beginnings of a new scientific theory, the terminologies are not unified. Grenander's proposal *abstract inference*, has been later be replaced by the non-exactly equivalent, *infinite-dimensional statistics* (Bongiorno et al. (2014)) or *functional statistics*. Recently, the overview paper Marron and Alonso (2014) proposes the name Object Oriented Data Analysis (OODA) to refer to *"statistical analysis of populations of complex objects"*; In that paper, classical Kendall's Shape Analysis (SA) is explicitly included in the OODA framework, alongside *Functional Data Analysis (FDA)*, the study of statistical methods (regression, classification, principal components, etc.) suitable for those situations in which the sample data $x_1, \dots, x_n$ are functions, typically (but not necessarily) depending of one real variable, $x_i : [a, b] \to \mathbb{R}$.

If we take the number of publications as a hint of the popularity of a scientific topic, FDA is perhaps the most successful chapter in the field of infinite-dimensional statistics. Since the popular textbook by Ramsay and Silverman (1997), several other well-known monographs have contributed to the popularization of FDA; see Ferraty and Vieu (2006), Ferraty and Romain (2011) and Horváth and Kokoszka (2012), among others. See also, Cuevas (2014) for a recent overview.

We think that Marron and Alonso (2014) make a good point in bringing together shape analysis and FDA as two particular instances of OODA. In fact, the conceptual relation between both topics is quite obvious at a formal level, since shapes can be ultimately identified with functions of some kind (or equivalence classes of functions). However, the connection holds true from, at least, two other more relevant aspects:

(a) We have shown that (under some restrictions) shapes can be identified with *density functions* (those of the corresponding interpoint distance distributions). Hence, following our approach, a statistical problem with shapes can be recast as a FDA problem in which the available data are density functions. See Delicado (2011) for an account of this topic. Many interesting issues can be considered in such a setup: for example, principal components analysis and other techniques of dimension reduction.

(b) Still, considering SA from the FDA point of view suggest to study the adaptation of the increasing literature on FDA *variable selection* (or *feature selection*), to the SA framework; see, for example Berrendero et al.

(2015) and references therein for some recent theoretical and practical insights on this subject. In particular, it seems worthwhile to analyze the possible connections between some of these variable selection and the classical landmarks theory in shape analysis.

Acknowledgement

This work has been partially supported by Spanish Grants MTM2013-44045-P (Berrendero and Cuevas) and MTM2013-41383-P (Pateiro-López).

Supplementary material

The "Supplementary material" document provides additional figures and tables for Section 5. It includes also a short discussion on the relation between the covariogram function and the interpoint distances distribution.

References

Ambrosio, L., Colesanti, A. and Villa, E. (2008). Outer Minkowski content for some classes of closed sets. *Math. Ann.* 342, 727–748.

Amit, Y., Grenander, U. and Piccioni, M. (1991). Structural image restoration trough deformable templates. *J. Amer. Statist. Assoc.* 86, 376–387.

Averkov, G. Bianchi, G. (2009). Confirmation of Matheron's conjecture on the covariogram of a planar convex body. *J. Eur. Math. Soc.* 11, 1187-1202.

Begg, G. A., Campana, S. E., Fowler, A. J., and Suthers, I. M. (2005) Otolith research and application: current directions in innovation and implementation. *Marine and Freshwater Research*, 56, 477-483.

Berrendero, J.R., Cholaquidis, A., Cuevas, A. and Fraiman, R. (2014). A geometrically motivated parametric model in manifold estimation. *Statistics* 48, 983–1004.

Berrendero, J.R., Cuevas, A. and Torrecilla, J.L. (2015). Variable selection in functional data classification: a maxima-hunting proposal. *Statistica Sinica*, to appear.

Bonetti, M. and Pagano, M. (2005). The interpoint distance distribution as a descriptor of point patterns, with an application to cluster detection. *Statistics in Medicine* 24, 753-773.

Bongiorno, E.G., Goia, A., Salinelli, E. and Vieu, P. (2014). *Contributions in infinite-dimensional statistics and related topics.*. S.E. Esculapio, Bologna.

Burago, D., Burago, Y. and Ivanov, S. (2002). *A Course in Metric Geometry.* American Mathematical Society.

Cabo, A.J. and Baddeley, A.J. (1995). Line transects, covariance functions and set convergence. *Adv. Appl. Prob.* 27, 585-605.

Cabo, A.J. and Baddeley, A.J. (2003). Estimation of mean particle volume using the set covariance function. *Adv. Appl. Prob.* 35, 27-46.

Caelli, T. (1980). On generating spatial configurations with identical interpoint distance distributions. Combinatorial mathematics, VII (Proc. Seventh Australian Conf., Univ. Newcastle, Newcastle, 1979), pp. 69-75, Lecture Notes in Math., 829, Springer, Berlin.

Cagliari, F., Di Fabio, B. and Landi, C. (2014). The natural pseudo-distance as a quotient pseudo-metric, and applications. *Forum Mathematicum*, to appear. DOI: 10.1515/forum-2012-0152.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *J. Statist. Plann. Inference* 147, 1–23.

Cuevas, A., Fraiman, R. and Rodríguez-Casal, A. (2007). A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* 35, 1031-1051.

Delicado, P. (2011). Dimensionality reduction when data are density functions. *Comp. Stat. Data Anal.* 55, 401–420.

Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* 93, 418–491.

Ferraty, F. and Romain, Y., eds. (2011). The Oxford Handbook of Functional Data Analysis. Oxford University Press, Oxford.

Ferraty, F. and Vieu, P. (2006). Nonparametric Functional Data Analysis: Theory and Practice. Springer.

Folland, G.B. (1999). *Real Analysis. Modern Techniques and Their Applications.* Wiley, New York.

Galerne, B. (2011). Computation of the perimeter of measurable sets via their covariogram. Applications to random sets. *Image Anal. Stereol.* 30, 39-51.

Gibbs, A.L. and Su, F.E. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.*, 70, 3, 419-435.

Grenander, U. (1976). *Pattern Synthesis. Lectures in Pattern Theory. Volume 1.* Springer-Verlag, New York.

Grenander, U. (1981). *Abstract Inference.* Wiley, New York.

Hatcher, A. (2002). *Algebraic Topology.* Cambridge University Press.

Hobolt, A. and Vedel-Jensen, E.B. (2000). Modelling stochastic changes in curve shape, with an application to cancer diagnostics. *Adv. Appl. Prob.* 32, 344–362.

Hobolt, A., Pedersen, J. and Vedel-Jensen, E.B. (2003). A continuous parametric shape model. *Ann. Inst. Statist. Math.* 55, 227–242.

Horváth, L. and Kokoszka, P. (2012). Inference for Functional Data with Applications. Springer, New York.

Kendall, D.G. (1989). A survey of the statistical theory of shape. *Statist. Sci.* 4, 87–120.

Kendall, D.G., Barden, D., Carne, T.K. and Le, H. (1999). *Shape and Shape Theory.* Wiley.

Kendall, W.S. and Le, H. (2010). Statistical shape theory. In *New Perspectives in Stochastic Geometry,*W.S. Kendall and I. Molchanov, eds., pp. 348-373. Oxford University Press.

Kent, J.T. (1994). The complex Bingham distribution and shape analysis. *J. Royal Statist. Soc. B* 56, 285–299.

Kent, J.T. (1995). Current issues for statistical inference in shape analysis. In *Proceedings in Current Issues in Statistical Shape Analysis*, K.V. Mardia and C.A. Gill eds., pp. 167–175. Leeds University Press.

Lemke, P., Skiena, S.S. and Smith, W.D. (1995). Reconstructing sets from interpoint distances. In *Discrete and Computational Geometry. Algorithms and Combinatorics* Volume 25, B. Aronov, S. Basu, J. Pach y M. Sharir, eds., pp. 597-631. Springer, New York.

Ling, H., Okada. K. (2007). An Efficient Earth Mover's Distance Algorithm for Robust Histogram Comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 840-853.

Lombarte, A., Chic, O., Parisi-Barabad, V., Olivella, R., Piera, J., García-Ladona, E. (2006). A web-based environment for shape analysis of fish otoliths. The AFORO database. *Sci. Mar.* 70, 147-152.

Mallows, C.L. and Clark, J.M.C. (1970). Linear-intercept distributions do not characterize plane sets. *J. Appl. Probability* 7, 240-244.

Marron, J.S. and Alonso, A.M. (2014). Overview of object oriented data analysis. *Biometrical Journal* 56, 732–753.

Martin, G.E. (1982). *Transformation Geometry. An Introduction to Symmetry.* Springer-Verlag. New York.

Matern, B. (1986). *Spatial Variation.* Lecture Notes in Statistics 36, Springer. New York.

Matheron, G. (1986). Le covariogramme gometrique des compacts convexes des $\mathbb{R}^2$. *Technical report N- 2/86/G*, Centre de Gostatistique, Ecole Nationale Suprieure des Mines de Paris.

Mattila, P. (1995). Geometry of Sets and Measures in Euclidean Spaces: Fractals and Rectifiability. Cambridge University Press. Cambridge.

Osada, R., Funkhouser, T., Chazelle, B. and Dobkin, D. (2002). Shape Distributions. *ACM Transactions on Graphics* 21, 807-832.

Ramsay, J. O. and Silverman, B. W. (1997). Functional Data Analysis. Springer, New York.

Rubner, Y., Tomasi, C. and Guibas, L.J. (2000). The Earth Movers Distance as a metric for image retrieval. *Intl J. Computer Vision* 40, 99–121.

Storbeck, F. and Daan, B. (2001). Fish species recognition using computer vision and a neural network. *Fisheries Research* 51, 11-15.

Tebaldi, P., Bonetti, M., and Pagano, M. (2011). M statistic commands: interpoint distance distribution analysis. *The Stata Journal* 11, 271–289.

Tuset, V. M., Lombarte, A. and Assis, C. A. (2008) Otolith atlas for the western Mediterranean, north and central eastern Atlantic. *Scientia Marina*, 72, 7-198.

Villani, C. (2003). *Topics in Optimal Transportation.* Graduate Studies in Mathematics, 58. American Mathematical Society, Providence, RI.

Supplementary Material for "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*

An application to fish family identification from otolith images

Filled-in contour images. The AFORO database contains both high resolution and filled-in contour images of otoliths, see Figure 1. The results in the study are obtained from the filled-in contour images.

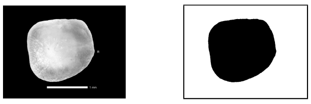

Figure 1: High resolution image (left) and filled-in contour image (right) of a *Soleidae* otolith.

Hierarchical clustering. We consider the complete dataset consisting of otoliths from three families of fishes (*Soleidae*, *Labridae* and *Scombridae*) and apply an agglomerative hierarchical clustering procedure using both the L_1 distance and the Wasserstein distance as dissimilarity criterion. In Figure 2 we show the dendrogram obtained using the L_1 distance. If we cut the corresponding tree into three groups, we obtain that 94.29% of *Soleidae* otoliths belong to the first cluster, 96.80% of *Labridae* otoliths belong to the second cluster and 95.56% of *Scombridae* otoliths belong to the third cluster. See Table 1 for the complete table of results.

k-means clustering. In this section, we investigate the performance of the k-means clustering algorithm. We apply the k-means algorithm to each pair of families of otoliths (k=2). We present the results based on the L_1 distance between densities and the Wasserstein distance between distributions, see Table 2.

We observe that k-means performs reasonably well, except perhaps on the dataset consisting on *Labridae* and *Scombridae* otoliths (dataset C), see Table 2. The k-means algorithm highly depends on the initial centroids and this may be the reason for the not so good results in this dataset.

[*] Corresponding address: Rúa Lope Gómez de Marzoa s/n. 15782 Santiago de Compostela. Spain

Email address: beatriz.pateiro@usc.es (Beatriz Pateiro-López)

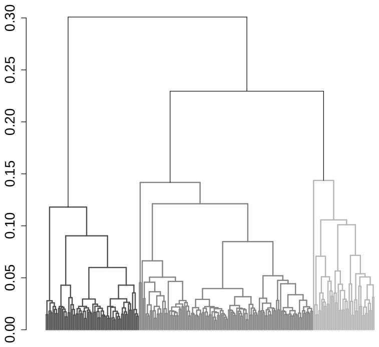

Figure 2: Dendrogram using the L_1 distance between the estimated interpoint distance densities of the otoliths in the families *Soleidae*, *Labridae* and *Scombridae*. The tree is cut into three groups, represented in different tones of gray.

Table 1: Results of the hierarchical clustering procedure for *Soleidae*, *Labridae* and *Scombridae* otoliths. For each dissimilarity criterion, count and row percent of the true family labels versus the group labels for a partition into three clusters.

	L_1 distance			Wasserstein distance		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Soleidae	66	4	0	58	12	0
	94.29%	5.71%	0.00%	82.85%	17.14%	0.00%
Labridae	2	121	2	0	123	2
	1.60%	96.80%	1.60%	0.00%	98.40%	1.60%
Scombridae	0	2	43	0	2	43
	0.00%	4.44%	95.56%	0.00%	4.44%	95.56%

Table 2: Results of the k-means algorithm ($k = 2$). For each distance, contingency table (count and row percent) of the true family labels versus the group labels.

		L_1 distance		Wasserstein distance	
		Cluster 1	Cluster 2	Cluster 1	Cluster 2
Dataset A	*Soleidae*	68	2	65	5
		97.14%	2.86%	92.86%	7.14%
	Labridae	4	121	0	125
		3.20%	96.80%	0.00%	100.00%
Dataset B	*Soleidae*	69	1	67	3
		98.57%	1.43%	95.71%	4.29%
	Scombridae	0	45	0	45
		0.00%	100.00%	0.00%	100.00%
Dataset C	*Labridae*	123	2	101	24
		98.40%	1.60%	80.80%	19.20%
	Scombridae	2	43	5	40
		4.44%	95.56%	11.11%	88.89%

Gobiidae and Serranidae otoliths. At the beginning of the study, we had also considered two other large families: *Gobiidae* and *Serranidae* (see Figure 3 for examples of otoliths in these two families). As might be expected, the clustering methods did not perform well, for example, for the dataset containing *Gobiidae* and *Soleidae* otoliths. Note that the shape of some of the *Gobiidae* otoliths resembles that of the *Soleidae* otoliths. The same occurs for the dataset containing *Serranidae* and *Labridae* otoliths.

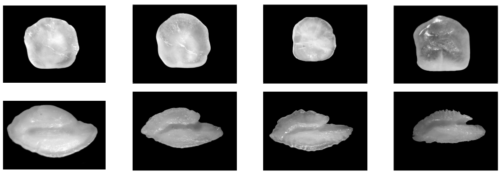

Figure 3: High resolution images of otoliths. First row: *Gobiidae*. Second row: *Serranidae*

Interpoint distances and covariogram

In this section we will establish some simple relationships between the interpoint distance and the covariogram, a well-known tool in stochastic geometry. As a consequence, some properties of the interpoint distance distribution will result.

The covariogram of a bounded Borel set $A \subset \mathbb{R}^d$ is defined by

$$K_A(y) = \mu(A \cap T_y A),$$

where $y \in \mathbb{R}^d$, $T_y A = A - y = \{a - y : a \in A\}$ and μ denotes the Lebesgue measure in $\mathbb{R}^d$.

This function has proven to be useful in different problems of stochastic geometry and stereology. Some references are Cabo and Baddeley (1995, 2003) and Galerne (2011). First note that

$$K_A(y) = \int_{\mathbb{R}^d} \mathbb{I}_A(x)\mathbb{I}_A(x-y)dx,$$

so that, K_A can be alternatively expressed in terms of a convolution of two indicator functions,

$$K_A = \mathbb{I}_A * \mathbb{I}_{-A}, \tag{1}$$

where $-A$ denotes the symmetric set $-A = \{-x : x \in A\}$.

Note that (1) is, up to a multiplicative constant, the density function of $X_1 - X_2$, where X_1, X_2 are iid random variables uniform on A. As a conclusion, K_A fully determines the distribution of the interpoint distance Y_A.

Let us now briefly summarize some other relevant properties of this function; see, e.g. Lemmas 1.2, 1.3 and 1.4 in Cabo and Baddeley (1995) and Proposition 2 in Galerne (2011).

Lemma 1. *Let $A \in \mathbb{R}^d$ be a bounded Borel set with covariogram K_A.*

(i) For all $y \in \mathbb{R}^d$, $0 \leq K_A(y) \leq K_A(0) = \mu(A)$. Moreover, $K_A(y) = 0$ whenever $\|y\| \geq \operatorname{diam}(A)$, $K_A(y) = K_A(-y)$ for all $y \in \mathbb{R}^d$ and K_A is uniformly continuous on $\mathbb{R}^d$.

(ii) For any integrable $f : [0, \infty) \to \mathbb{R}$,

$$\int_A \int_A f(\|x-y\|)dxdy = \int_{\mathbb{R}^d} f(\|w\|)K_A(w)dw.$$

This is the so-called "Borel's overlap formula". Two interesting particular cases are obtained for $f \equiv 1$ and $f(t) = \mathbb{I}_{[0,\rho]}(t)/\mu(A)^2$, leading respectively to

$$\int_{\mathbb{R}^d} K_A(y)dy = \mu(A)^2 \tag{2}$$

and

$$\mathbb{P}\{Y_A \leq \rho\} = \frac{1}{\mu(A)^2}\int_{B(0,\rho)} K_A(y)dy, \text{ for } \rho > 0, \tag{3}$$

where $Y_A = \|X_1 - X_2\|$, X_1 and X_2 being independent random variables uniformly distributed on A.

The following property of the interpoint distance distribution follows directly from Lemma 1.

Proposition 1. *Let X_1, X_2 be independent random variables uniformly distributed on C. Denote $Y_C = \|X_1 - X_2\|$. Then, Y_C has a continuous density f_C with $f_C(0) = 0$ and $f_C(\rho_C) = 0$, where $\rho_C = diam(C)$.*

Proof. Performing a change of variables to polar coordinates in (3) we have

$$\mathbb{P}\{\|X_1 - X_2\| \leq \rho\} = \frac{1}{\mu(C)^2}\int_0^\rho \int_0^{2\pi} rK_C(r\cos\theta, r\sin\theta)d\theta dr$$

Since K_C is continuous, we can differentiate under the integral sign to get that the distribution of the interpoint distance has the following continuous density

$$f_C(\rho) = \frac{1}{\mu(C)^2}\int_0^{2\pi} \rho K_C(\rho\cos\theta, \rho\sin\theta)d\theta, \text{ for all } \rho \in [0,1].$$

In particular, for $\rho = 0$ we get $f_C(0) = 0$. Also, from result (i) in Lemma 1, $f_C(\rho_C) = 0$. □

References

Cabo, A.J. and Baddeley, A.J. (1995). Line transects, covariance functions and set convergence. *Adv. Appl. Prob.* 27, 585-605.

Cabo, A.J. and Baddeley, A.J. (2003). Estimation of mean particle volume using the set covariance function. *Adv. Appl. Prob.* 35, 27-46.

Galerne, B. (2011). Computation of the perimeter of measurable sets via their covariogram. Applications to random sets. *Image Anal. Stereol.* 30, 39-51.