Polynomial volume estimation and its applications

Antonio Cuevas, Beatriz Pateiro-López

FUNDING

Research has been funded by project INNPAR2D (reference code MTM2016-76969-P), from the Ministry of Economy and Competitiveness and the European Regional Development Fund (ERDF).

Polynomial volume estimation and its applications

Antonio Cuevas[a] and Beatriz Pateiro-López[b]

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

[b]Departamento de Estadística, Análisis Matemático y Optimización,
Universidad de Santiago de Compostela, Spain

Abstract

Given a compact set $S \subset \mathbb{R}^d$ we consider the problem of estimating, from a random sample of points, the Lebesgue measure of S, $\mu(S)$, and its boundary measure, $L(S)$ (as defined by the Minkowski content of ∂S). This topic has received some attention, especially in the two-dimensional case $d = 2$, motivated by applications in image analysis. A new method to simultaneously estimate $\mu(S)$ and $L(S)$ from a sample of points inside S is proposed.

The basic idea is to assume that S has a polynomial volume, that is, that $V(r) := \mu\{x : d(x, S) \leq r\}$ is a polynomial in r of degree d, for all r in some interval $[0, R)$. We develop a minimum distance approach to estimate the coefficients of $V(r)$ and, in particular $\mu(S)$ and $L(S)$, which correspond, respectively, to the independent term and the first degree coefficient of $V(r)$. The strong consistency of the proposed estimators is proved. Some numerical illustrations are given.

Keywords: Set estimation; volume estimation; boundary length estimation

1 Introduction

The general background. Set estimation

The theory of set estimation is closely linked to nonparametric statistics and stochastic geometry. The general goal of this theory is to estimate a compact set $S \subset \mathbb{R}^d$ from a random sample of points; see, e.g., Cuevas (2009) for a short overview. Some relevant applications appear in different areas, including ecology [estimation of the habitat of a species or the *home range* of typical individuals; see Getz and Wilmers (2004), Kie et al. (2010)], econometrics [estimation of the efficient boundary in productivity analysis; see Simar and Wilson (2000)], image analysis (Willett and Novak, 2007; Jang, 2006), nonparametric quality control (Baíllo and Cuevas, 2006), and clustering (Rinaldo and Wasserman, 2010).

In this setting, a natural aim is the estimation of some functionals of S, in particular the volume and boundary measure of S.

Some notations and basic definitions

Let us consider the d-dimensional Euclidean space $\mathbb{R}^d$, equipped with the usual inner product $\langle\cdot,\cdot\rangle$ and the corresponding norm $\|\cdot\|$. Given a set $A \subset \mathbb{R}^d$, we denote by A^c, $\mathrm{int}(A)$ and ∂A the complement, interior and boundary of A, respectively. We denote by $B(x, r)$ the closed ball with centre x and radius r. Also, for any compact set $A \subset \mathbb{R}^d$ we will denote (with a slight abuse of notation) by $B(A, \varepsilon)$ the closed ε-neighbourhood, or ε-parallel set, of A, $B(A, \varepsilon) = \{x \in \mathbb{R}^d : \ d(x, A) \leq \varepsilon\}$, where $d(a, C) = \inf\{\|a - c\| : c \in C\}$. If A and C are non-empty compact subsets of $\mathbb{R}^d$, the Hausdorff distance between A and C is defined by $d_H(A, C) = \inf\{\varepsilon > 0 : A \subset B(C, \varepsilon),\ C \subset B(A, \varepsilon)\}$. Denote by $\mu(S)$ the d-dimensional Lebesgue measure of S. Thus $\mu(S)$ is the volume of S for $d = 3$ and the area for $d = 2$. When no confusion is possible we will sometimes use these terms even for the general case $S \subset \mathbb{R}^d$.

The boundary measure of S (i.e., the "perimeter" or the "surface area" of S) is often defined in terms of the Minkowski content, $L_0(S)$, or its one-sided version, $L(S)$, given by

$$L_0(S) = \lim_{\varepsilon \to 0} \frac{\mu(B(\partial S, \varepsilon))}{2\varepsilon}, \text{ or } L(S) = \lim_{\varepsilon \to 0} \frac{\mu(B(S, \varepsilon) \setminus S)}{\varepsilon}, \tag{1}$$

provided that these limits do exist and are finite. Typically, the values $L_0(S)$ and $L(S)$ coincide for regular enough sets; see Ambrosio, Colesanti and Villa (2008) for details and additional references on the Minkowski content. In what follows, we will mostly use $L(S)$.

Let ν be a Borel measure on $\mathbb{R}^d$. Let A, C be Borel sets with finite ν-measure. The (pseudo) distance in measure between A and C is given by $d_\nu(A, C) = \nu(A \Delta C)$, where Δ denotes the symmetric difference between A and C, that is, $A \Delta C = (A \setminus C) \cup (C \setminus A)$. We often use either the Lebesgue measure μ or a probability measure in the role of ν.

Statement of the problem

Suppose we have a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ of iid observations from a random variable X with absolutely continuous probability distribution $P_X \equiv P$ and compact support $S \subset \mathbb{R}^d$. We want to estimate the volume (Lebesgue measure) of S, $\mu(S)$, and the surface measure, $L(S)$, as defined in (1).

A brief overview of boundary measure and volume estimation

To gain some perspective, and for comparison purposes with our proposal here, we list here some (mostly recent) contributions on the topic of volume and boundary measure estimation. These contributions can be organized according to different criteria depending on

i) the assumed model to generate the sample observations,

ii) the functional (volume or surface area) to be estimated,

iii) the type of estimator; in some cases the estimation is undertaken in a plug-in fashion, as a by-product of a set estimator $\hat{S}$ of S so that $\mu(S)$ and $L(S)$ are estimated by $\mu(\hat{S})$ and $L(\hat{S})$. In other cases direct estimators of the volume and boundary measure are constructed, not relying on any previous estimation of S.

Whatever the chosen approach some restrictions must be imposed on the set S. It is clear that the family of all compact sets with a finite boundary measure is huge and the task of estimating both $\mu(S)$ and $L(S)$ from a finite sample of points seems hopeless unless some additional shape conditions are imposed.

We now summarize some contributions on the topic according to the criteria i)-iii) listed above. First, let us consider the references dealing with estimation under convexity-type restrictions on the basis of the "inside model", i.e., all the sample points $X_1, \ldots, X_n$ are taken inside the target set S according to a distribution with support S. If S is assumed to be convex, then the natural estimator for $\mu(S)$ is $\mu(S_n)$, where S_n denotes the convex hull of the sample points. Some deep results concerning convergence rates and asymptotic distribution for this estimator can be found in Brunel (2016) and Pardon (2011). Of course the estimator $\mu(S_n)$ is typically biased since in general $S_n \subsetneq S$. The unbiased estimation of $\mu(S)$ when S is convex is addressed in Baldin and Reiss (2016).

Although the assumption of convexity for S is natural and appealing from different points of view, it is also quite restrictive for many practical applications. This has motivated the use of several extensions of the notion of convex set. One of them is α-convexity: a closed set S is said to be α-convex when it can be expressed as the intersection of the complements of a family of open balls of radius α. This definition is clearly inspired in the characterization of a closed convex set as an intersection of closed half-spaces; see Cuevas et al. (2012) for some background and references.

Assuming that S is α-convex the volume $\mu(S_{n,\alpha})$ of the α-convex hull of the sample provides a natural, biased, estimator for $\mu(S)$; see Rodríguez-Casal (2007).

An improved (bias corrected) version of this estimator, has been proposed by Arias-Castro et al. (2016). It achieves the minimax convergence rate under the regularity assumption that both S and S^c are α-convex.

Regarding the estimation of the perimeter $L(S)$, let us mention the case in which $S \subset \mathbb{R}^2$ satisfies the above mentioned α-convexity condition for a given $\alpha > 0$. Now, a natural estimator of $L(S)$ from an inside sample, is the corresponding perimeter of the α-convex hull of the sample, $L(S_{n,\alpha})$. In Cuevas et al. (2012, Th. 6) it is proved, under mild conditions, that in this α-convex bivariate case we have $L(S_{n,\alpha}) \to L(S)$, almost surely (a.s.), as $n \to \infty$. The non-trivial computational aspects can be dealt with using the R-package *alphahull* by Pateiro-López and Rodríguez-Casal (2010). Other related interesting ideas on the estimation of the perimeter, relying on the use of the so-called α-shape, are analyzed in Arias-Castro and Rodríguez-Casal (2016).

Now, let us focus on the references dealing with the estimation of the boundary measure under the following "inside-outside" model: assume that the target set S fulfils $S \subset (0,1)^d$. Under the "inside-outside" model we have independent identically distributed (iid) observations $(X_1, \mathbb{I}_S(X_1)), \ldots, (X_n, \mathbb{I}_S(X_n))$ of a random variable $(X, \mathbb{I}_S(X))$ where X is uniformly distributed on $[0,1]^d$ and $\mathbb{I}_S$ stands for the indicator function of S. Thus, under this model we also have sample data outside S and we assume that for each X_i we know $\mathbb{I}_S(X_i)$, that is, we are able to decide whether $X_i \in S$ or $X_i \in S^c$.

A plug-in type estimator of the boundary Minkowski content, see Cuevas et al. (2007); Armendáriz et al. (2009), is:

$$L_n = \frac{\mu(T_n(\varepsilon_n))}{2\varepsilon_n}, \text{ with } T_n = \{z \in [0,1]^d : \exists X_i \in B(z,\varepsilon_n) \cap S, \text{ and } X_j \in B(z,\varepsilon_n) \cap S^c\}.$$

A k-NN version of this idea can be found in Cuevas et al. (2013).

A different estimator based on Delaunay triangulations has been proposed by Jiménez and Yukich (2011). In fact, the technique proposed by these authors allows for the estimation of more general surface integrals under quite general shape restrictions. Still, the considered sampling model requires to have data points inside and outside the target set S.

Practical motivations

The perimeter $L(S)$ and the area $\mu(S)$ are, obviously, basic functionals of primary interest in the analysis of a set $S \subset \mathbb{R}^2$. The so-called "compactness index"

$$CI(S) = \frac{\text{Perimeter}(S)^2}{\text{Area}(S)} = \frac{L(S)^2}{\mu(S)}$$

has been used in shape analysis (with applications in medicine). See Montero and Bribiesca (2009) for a survey. Roughly speaking, $CI(S)$ measures how "irregular" the shape of S is.

The square root of $CI(S)$ has been sometimes called "contour index"; see Canzonieri and Carbone (1998) for an application in oncology.

Even if S is completely known, we might want to have good estimators of $L(S)$ and $\mu(S)$ based on Monte Carlo samples. Thus, the estimators of $L(S)$ and $\mu(S)$ provide a sort of stochastic algorithms to approximate these quantities as well as the compactness index $CI(S)$.

2 The assumption of polynomial volume: its geometric meaning

As we have indicated in the introduction, when only an inside sample is available, most usual estimators of $\mu(S)$ and $L(S)$ use a geometric shape condition on S (such as convexity or α-convexity) which is incorporated to the estimator S_n, via the "hull-principle" (that is, S_n is defined as the minimal set fulfilling the imposed condition). Then $\mu(S)$ and $L(S)$ are estimated by $\mu(S_n)$ and $L(S_n)$, which entails additional problems for the practical evaluation of these quantities.

We will follow here a different strategy: first we will assume a condition on S expressed in "analytic" or "algebraic" terms: it simply consists on imposing that the volume of the parallel set $V(r) = \mu(B(S,r))$ is a polynomial on some domain $[0,R)$. Second, we will see that the coefficients of such polynomial have a direct interpretation in terms of the parameters of interest, $\mu(S)$ and $L(S)$. Finally, the target parameters will be obtained by minimizing the distance between the function $V(r)$ and a consistent estimator $V_n(r)$ of this function.

So we start by defining our crucial assumption.

Definition 1. *A compact set $S \subset \mathbb{R}^d$ is said to fulfil the polynomial volume property if there exist constants $\theta_0, \ldots, \theta_d \in \mathbb{R}$ and $R > 0$ such that*

$$\mu(B(S,r)) = \theta_0 + \theta_1 r + \ldots + \theta_d r^d, \text{ for all } r \in [0, R). \tag{2}$$

This condition has been recently employed in a statistical context by Berrendero et al. (2014). However these authors use a sample model quite different to that considered here.

Geometric aspects of the polynomial volume assumption

First, it is clear that under condition (2), the Lebesgue measure of S is $\mu(S) = \theta_0$ and the one-sided Minkowski content (1) of S is $L(S) = \theta_1$.

Second, we should mention that assumption (2) is closely related to the concept of convexity. The classical Steiner's theorem in convex geometry establishes that any compact convex set satisfies condition (2) for all $r \in [0, \infty)$; see e.g. Morvan (2008, Ch. 16). In Heveling et al. (2004) it is proved that, for the two-dimensional case $d = 2$, the validity of (2) in $[0, \infty)$ is in fact equivalent to the convexity of S. These authors also give counterexamples to prove that such equivalence does not hold for $d = 3$.

Third, and foremost, there is a class of sets, much wider than that of convex sets, which satisfies condition (2). This is the class of sets with positive reach. This notion, first introduced by Federer (1959), has a clear intuitive meaning. It is formally defined as follows. Let $\text{Unp}(S)$ be the set of points $x \in \mathbb{R}^d$ with a unique metric projection, denoted by $\xi_S(x)$, on S. This means that, for $x \in \text{Unp}(S)$, $\xi_S(x)$ is the unique point fulfilling $d(x, S) = \|x - \xi_S(x)\|$. For $x \in S$, let $\text{reach}(S, x) = \sup\{r > 0 : \text{int}(B(x, r)) \subset \text{Unp}(S)\}$. Then, the reach of S is defined by $r_0 := \text{reach}(S) = \inf\{\text{reach}(S, x) : x \in S\}$. We will say that S is a set with positive reach if $\text{reach}(S) > 0$. The condition that S has a positive reach is clearly a sort of smoothness assumption on S, expressed in purely geometric terms, with no direct use of differentiability properties. In simple, informal terms, the assumption $r_0 > 0$ rules out the existence of sharp inwards peaks in the boundary of S. Positive reach is a well-known extension of the crucial notion of convexity. Indeed it is not difficult to show that for a closed set S, $\text{reach}(S) = \infty$ if and only if S is convex.

In the pioneering paper by Federer (1959), the author proved the following result, establishing the relation between positive reach and polynomial volume.

Theorem (Federer, 1959, Th. 5.6 and Th. 5.19). *Let $S \subset \mathbb{R}^d$ be a compact set with $r_0 = \mathrm{reach}(S) > 0$. Then there exist unique values $\Phi_0(S), \ldots, \Phi_d(S)$ such that*

$$\mu(B(S,r)) = \sum_{i=0}^{d} r^{d-i} \omega_{d-i} \Phi_i(S), \ \textit{for } 0 \leq r < r_0, \tag{3}$$

where $\omega_0 = 1$ and, for $j \geq 1$, ω_j denotes the j-dimensional measure of the unit ball in $\mathbb{R}^j$.

Furthermore, $\Phi_0(S)$ coincides with the Euler characteristic of S.

As a consequence, if S is a compact set with positive reach, then $\Phi_d(S) = \mu(S)$ and the one-sided Minkowski content, $L(S)$, in (1) always exists and corresponds to the coefficient of the first-degree term in (3).

Federer's result is in fact much deeper than stated here, since in particular the $\Phi_i(S)$ have an interpretation in terms of curvature measures. For our purposes, besides the independent term ($= \mu(S)$) and the coefficient of the first order term ($= L(S)$), it is important that the term $\Phi_0(S)$ in (3) equals the Euler characteristic of S. This is an integer valued topological invariant (i.e., it is preserved by homeomorphisms). For example, if S is a closed ball, $\Phi_0(S) = 1$, and the same holds for any other compact set S homeomorphic to the closed ball. In practice, this means that the highest order coefficient can be sometimes assumed to be known in (3). This amounts to impose a further, not too restrictive, geometric condition. In particular, the additional restriction $\theta_d = \mu(B(0,1))$ in (2) is meaningful and useful in many cases.

As a further appealing feature of the polynomial volume assumption, let us mention that many simple interesting sets not-fulfilling the positive reach property do in fact satisfy (2). This is the case of the "pac-man" figure obtained in $\mathbb{R}^2$ by erasing in the unit ball all the points in the first quadrant (see Figure 1), or two tangent spheres in $\mathbb{R}^3$, and many others; see Heveling et al. (2004) for details.

3 Estimation of $\mu(S)$ and $L(S)$ under the polynomial volume assumption

In this section we establish the theoretical basis of our method. In particular, it is shown how to consistently estimate the coefficients $\theta_0, \ldots, \theta_d$ in (2).

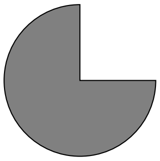

Figure 1: The set in gray has not positive reach but satisfies the polynomial volume property (2).

3.1 An auxiliary result

The following elementary result is mentioned here just as an auxiliary tool in the proof of Theorem 1. It establishes that the Hausdorff convergence of S_n to S entails the convergence in measure for the corresponding parallel sets, that is, $d_\mu(B(S_n, r), B(S, r)) \to 0$ for $r > 0$. No further assumption is needed. In intuitive terms, this means that the parallel set is always "regular enough" to ensure the convergence in measure. In particular, note that if we take the sample itself, i.e., $S_n = \mathcal{X}_n$, the assumption $d_H(S_n, S) \to 0$ is obviously true, with probability one, so that the estimation in measure of $B(S, r)$ can be done even using the simplest possible estimator, $B(\mathcal{X}_n, r) = \cup_{i=1}^n B(X_i, r)$; see Devroye and Wise (1980), Kim and Korostelev (2000), Cuevas and Rodríguez-Casal (2004).

Lemma 1. *Let S be a compact non-empty set in $\mathbb{R}^d$. Let S_n be a sequence of compact sets such that $d_H(S_n, S) \to 0$. Then, for $r > 0$,*

$$d_\mu(B(S_n, r), B(S, r)) \to 0. \tag{4}$$

Proof. We have to prove that $\mu(B(S_n, r)\Delta B(S, r)) \to 0$. Given $\varepsilon \in (0, r/2]$ we have, for n large enough,

$$\mu(B(S_n, r) \setminus B(S, r)) \leq \mu(B(S_n, r) \setminus B(S, r - \varepsilon)) \ \leq \mu(B(S, r + \varepsilon) \setminus B(S, r - \varepsilon)), \tag{5}$$

since by the Hausdorff convergence, for n large enough, $S_n \subset B(S, \varepsilon)$ and, therefore, $B(S_n, r) \subset B(S, r + \varepsilon)$. Second, we have also eventually

$$\mu(B(S, r) \setminus B(S_n, r)) \leq \mu(B(S, r) \setminus B(S_n, r - \varepsilon)) \ \leq \mu(B(S, r) \setminus B(S, r - 2\varepsilon)), \tag{6}$$

since, again by the Hausdorff convergence, for n large enough, $S \subset B(S_n, \varepsilon)$ and, therefore, $B(S, r - 2\varepsilon) \subset B(S_n, r - \varepsilon)$. Now, the result follows directly from (5) and (6), together with the fact that the volume function $V(r) = \mu(B(S,r))$ is continuous; see, e.g., Stachó (1976, Theorem 4 and Lemma 2 (i)). □

3.2 Our estimators: consistency results

The idea is that if, for a compact set S fulfilling the polynomial volume assumption (2), we are able to approximate $\mu(B(S,r))$ from a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ for different values of r, then we can consistently estimate the coefficients $\theta = (\theta_0, \ldots, \theta_d)$.

Theorem 1. *Let $S \subset \mathbb{R}^d$ be a compact set fulfilling the polynomial volume assumption introduced in Definition 1. Given an interval $[a,b] \subset (0,R)$, let $V(r) = V(r,\theta) = \mu(B(S,r))$, as defined in (2), where $\theta = (\theta_0, \ldots, \theta_d)$ and $r \in [a,b]$. Let $S_n(r) = \cup_{i=1}^n B(X_i, r)$ be the "offset" estimator of $B(S,r)$ based on a sample $X_1, \ldots, X_n$ drawn from a distribution P with support S. Denote also $V_n(r) = \mu(S_n(r))$.*

Then,

(a)

$$\|V_n - V\|_\infty := \sup_{r\in[a,b]} |V_n(r) - V(r)| \to 0, \text{ almost surely, as } n \to \infty. \tag{7}$$

(b) The minimum-distance estimator of θ, given by

$$\hat{\theta} = \text{argmin}_{\tau\in\mathbb{R}^{d+1}} \|V_n(\cdot) - V(\cdot, \tau)\|_\infty \tag{8}$$

is uniquely defined and (componentwise) almost surely consistent for θ, that is

$$\hat{\theta} \to \theta, \text{ almost surely, as } n \to \infty. \tag{9}$$

(c) An analogous strong consistency result holds if the estimator $\hat{\theta}$ is replaced with the estimator $\tilde{\theta}$, obtained as indicated in (8), but with the supremum norm $\|\cdot\|_\infty$ replaced with the L_p norm $\|f\|_p = \left(\int_a^b |f|^p\right)^{1/p}$, for any $p > 1$.

Proof. (a) From Lemma 1, $V_n(r) \to V(r)$ a.s. for all $r \in [a,b]$, since $d_H(\mathcal{X}_n, S) \to 0$, a.s. Now, observe that V_n and V are continuous *increasing* functions. Therefore, given $\varepsilon > 0$ there is a partition $a = r_0 < \ldots < r_{k+1} = b$ such that $V(r_{i+1}) - V(r_i) < \varepsilon$ for all $i = 0, \ldots, k$. Then, the uniform convergence (7) follows directly from an elementary ε, n-argument involving the a.s. pointwise convergence $V_n(r_i) \to V(r_i)$ and the monotonicity of the V_n.

(b) From (a), given $\varepsilon > 0$ we may ensure that, with probability one,

$$\|V_n - V\|_\infty < \varepsilon, \quad \text{for } n \text{ large enough, say } n \geq n_0. \tag{10}$$

On the other hand, from the classical Chebyshev's Theorem in approximation theory [see, e.g., DeVore (1986, Ths. 2.1 and 2.2)], since V_n is a continuous function on $[a, b]$ there is a unique best approximation P_n for V_n in $[a, b]$ within the space Π_d of polynomials of degree at most d. Note that, by construction, since P_n is unique, we must have $P_n(r) = V(r, \hat{\theta})$, for $r \in [a, b]$, where $\hat{\theta}$ is given by (8). Since, by assumption, $V \in \Pi_d$ we must have, with probability one that for $n \geq n_0$ the best approximant P_n of V_n in Π_d must fulfill $\|P_n - V_n\|_\infty \leq \|V_n - V\|_\infty < \varepsilon$ a.s. for $n > n_0$. As a consequence, $\|P_n - V\|_\infty \to 0$, a.s.

Now observe that Π_d is a finite dimensional vector space. Hence, all norms defined on Π_d are equivalent. In particular, the norm in Π_d defined (for each polynomial) as the maximum of the absolute values of the polynomial coefficients is equivalent to the norm defined as the restriction of the supremum norm to Π_d.

Therefore, since $P_n, V \in \Pi_d$, the consistency (9) follows directly from $\|P_n - V\|_\infty \to 0$ a.s., in view of the equivalence of the two norms mentioned above.

(c) The proof is almost identical to that for the case $\|\cdot\|_\infty$. We will need to use the fact that for the L_p-norm, with $1 < p < \infty$, we also have that there is a unique L_p-best approximant for V since the unit ball in L_p is strictly convex; see DeVore (1986, Th. 1.2 and subsequent remarks). Now, we get $\|P_n - V_n\|_p \to 0$ a.s. and the proof goes along the same lines as in (b) since $\|V_n - V\|_\infty \to 0$, a.s. on $[a, b]$ also implies $\|V_n - V\|_p \to 0$, a.s. □

4 Numerical experiments

The goal of this section is to show the practical performance of the method based on the polynomial volume assumption (2), combined with the L_2-based minimum distance estimates presented in Subsection 3.2; see part (c) of Theorem 1. An alternative version of the method, somewhat simpler computationally, will be also considered in Subsection 4.3. As a simple strategy to evaluate the results, we will compare them with those obtained with the plug-in estimators based on other stronger shape assumptions. For example, if we assume that S is convex (resp. α-convex), we can estimate $\mu(S)$ and $L(S)$ by $\mu(S_n)$ and $L(S_n)$ where S_n is the convex hull (resp. $S_n \equiv S_{n,\alpha}$ is the α-convex hull) of the sample data. Note, however, that these plug-in approaches can be either very risky (for example, convexity is a very restrictive condition) or difficult to implement (to our knowledge, there is no available algorithm to calculate $\mu(S_{n,\alpha})$ and $L(S_{n,\alpha})$ when $S_{n,\alpha}$ is the α-convex hull of the sample in dimension larger than 2). By contrast, the polynomial volume assumption is reasonably easy to handle, especially for small values of d.

4.1 Some two-dimensional examples: the disk and the annulus

We first consider the closed unit disk $S = B(0,1)$ in $\mathbb{R}^2$. Note that S satisfies (2) for all $r \in \mathbb{R}$. For different sample sizes, we generate $B = 500$ samples from the uniform distribution on S. For each sample, we calculate $V_n(r)$ for 50 equally spaced values of $r \in [1,2]$. The estimators $\tilde{\theta}_i$, $i = 0,1,2$ are then obtained as indicated in part (c) of Theorem 1 with the L_2 norm. Table 1 shows the mean value and standard deviation of $|\tilde{\theta}_i - \theta_i|/\theta_i$ over the B repeats. Taking into account that $\theta_0 = \mu(S)$ and $\theta_1 = L(S)$, we can compare the results in Table 1 with those obtained if we estimate θ_0 and θ_1 with $\hat{\theta}_0 = \mu(S_n)$ and $\hat{\theta}_1 = L(S_n)$, S_n being the convex hull of the sample and $\mu(S_n)$ and $L(S_n)$ its area and perimeter, respectively. Results are summarized in Table 2, case (a). For convex support estimation, the convex hull of the sample is asymptotically optimal in minimax sense; see Korostelëv and Tsybakov (1993). This is not, however, the case if we consider the volume of the convex hull of the sample as an estimator of the volume of a convex support. Baldin and Reiss (2016) propose a minimax optimal estimator for the volume of a convex set, based on a Poisson point process model; see also Ripley and Rasson (1977) and Moore (1984). Such estimator coincides with the volume of a dilation of the convex hull of the sample from its barycentre. Therefore, this dilated hull could be considered as estimator, S_n, for the set S itself and $\hat{\theta}_0 = \mu(S_n)$ and $\hat{\theta}_1 = L(S_n)$ as estimators for the area and perimeter of S, respectively. Some simulation results are summarized in Table 2, case (b).

	$\|\tilde{\theta}_0 - \theta_0\|/\theta_0$		$\|\tilde{\theta}_1 - \theta_1\|/\theta_1$		$\|\tilde{\theta}_2 - \theta_2\|/\theta_2$	
Sample size	Mean	SD	Mean	SD	Mean	SD
$n = 500$	0.06698	0.00817	0.01537	0.00320	0.00201	0.00036
$n = 1000$	0.04259	0.00483	0.00967	0.00187	0.00128	0.00020
$n = 5000$	0.01466	0.00131	0.00331	0.00048	0.00044	0.00005

Table 1: Mean value and standard deviation of $|\tilde{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,1)$, where $\tilde{\theta}_i$, $i = 0, 1, 2$, are obtained as indicated in Theorem 1 (c) for $p = 2$. For each sample, we calculate $V_n(r)$ for 50 equally spaced values of $r \in [1, 2]$.

		$\|\hat{\theta}_0 - \theta_0\|/\theta_0$		$\|\hat{\theta}_1 - \theta_1\|/\theta_1$	
Sample size	S_n	Mean	SD	Mean	SD
$n = 500$	(a)	0.05262	0.00735	0.01999	0.00314
	(b)	0.00790	0.00578	0.00735	0.00390
$n = 1000$	(a)	0.03337	0.00437	0.01263	0.00185
	(b)	0.00451	0.00366	0.00456	0.00227
$n = 5000$	(a)	0.01149	0.00116	0.00433	0.00048
	(b)	0.00122	0.00092	0.00149	0.00065

Table 2: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,1)$, $i = 0, 1$. Now $\hat{\theta}_0 = \mu(S_n)$ and $\hat{\theta}_1 = L(S_n)$, where S_n is (a) the convex hull of the sample and (b) the dilation of the convex hull of the sample from its barycentre, as proposed in Baldin and Reiss (2016).

Next, we consider a non-convex set, the annulus $S = B(0, 5) \setminus \text{int}(B(0, 4))$. Note that S satisfies the polynomial volume assumption introduced in Definition 1 for $r < 4$. Moreover, the highest order coefficient of the polynomial in (2) is $\theta_2 = 0$ (the Euler characteristic of S is zero). Again, for each sample size, we generate $B = 500$ samples from the uniform distribution on S. For each sample, we calculate $V_n(r)$ for 50 equally spaced values of $r \in [2, 3.5]$. The results are summarized in Table 3. These results slightly improve when the Euler characteristic is assumed to be known, see Table 4.

	$\vert\tilde{\theta}_0 - \theta_0\vert/\theta_0$		$\vert\tilde{\theta}_1 - \theta_1\vert/\theta_1$		$\vert\tilde{\theta}_2 - \theta_2\vert$	
Sample size	Mean	SD	Mean	SD	Mean	SD
$n = 500$	0.19841	0.01469	0.01904	0.00285	0.15253	0.02006
$n = 1000$	0.12501	0.00886	0.01183	0.00160	0.09470	0.01111
$n = 5000$	0.04251	0.00218	0.00400	0.00041	0.03229	0.00274

Table 3: Mean value and standard deviation of $|\tilde{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,5) \setminus \mathrm{int}(B(0,4))$, where $\tilde{\theta}_i$, $i = 0, 1, 2$, are obtained as indicated in Theorem 1 (c) for $p = 2$ (since $\theta_2 = 0$, we report the mean value and standard deviation of $|\tilde{\theta}_2 - \theta_2|$). For each sample, we calculate $V_n(r)$ for 50 equally spaced values of $r \in [2, 3.5]$.

	$\vert\hat{\theta}_0 - \theta_0\vert/\theta_0$		$\vert\hat{\theta}_1 - \theta_1\vert/\theta_1$	
Sample size	Mean	SD	Mean	SD
$n = 500$	0.15867	0.01235	0.00421	0.00134
$n = 1000$	0.10033	0.00736	0.00262	0.00077
$n = 5000$	0.03410	0.00187	0.00086	0.00021

Table 4: Mean value and standard deviation of $|\tilde{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,5) \setminus \mathrm{int}(B(0,4))$, $i = 0, 1$ (assumming $\theta_2 = 0$ known). The estimations $\tilde{\theta}_i$ are obtained as indicated in Theorem 1 (c) for $p = 2$. For each sample, we calculate $V_n(r)$ for 50 equally spaced values of $r \in [2, 3.5]$.

In order to compare the results in Tables 3 and 4 with those obtained with other procedures based on a more specific information on the set of interest, we use that S is α-convex for $\alpha = 4$. Then, we can estimate θ_0 and θ_1 with $\hat{\theta}_0 = \mu(S_{n,r})$ and $\hat{\theta}_1 = L(S_{n,r})$, being $S_{n,r}$ the r-convex hull of the sample and $\mu(S_{n,r})$ and $L(S_{n,r})$ its area and perimeter, respectively. The outputs are summarized in Table 5 for two different values of r (the value $r = 4$ corresponds to the case when α is assumed to be known and r is chosen accordingly, whereas $r = 2$ corresponds to a more conservative estimation of S for the case when α is assumed to be unknown).

Estimation by the interpolation method

In view of Lemma 1, another natural way of estimating the coefficients θ_i, $i = 0, \ldots, d$ under the polynomial volume assumption (2) would be just to choose some values $r_j \in [a, b] \subset (0, R)$

	$r = 4$				$r = 2$			
	$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$		$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$	
Sample size	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$n = 500$	0.10708	0.01083	0.00310	0.00096	0.18060	0.01320	0.00915	0.00169
$n = 1000$	0.06692	0.00626	0.00209	0.00054	0.11392	0.00794	0.00568	0.00095
$n = 5000$	0.02223	0.00157	0.00080	0.00014	0.03872	0.00197	0.00190	0.00025

Table 5: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,5) \setminus \mathrm{int}(B(0,4))$, $i = 0, 1$. Now $\hat{\theta}_0 = \mu(S_{n,r})$ and $\hat{\theta}_1 = L(S_{n,r})$, $S_{n,r}$ being the r-convex hull of the sample with $r = 4$ and $r = 2$.

for $j = 1, \ldots, d+1$ and solve in $\theta = (\theta_0, \ldots, \theta_d)$ the system of equations

$$V_n(r_j) = V(r_j), \ j = 1, \ldots, d+1, \tag{11}$$

(we keep the notations as in Theorem 1).

In principle, this “interpolation method” looks a bit reminiscent of the classical method of moments in parametric estimation. It could be also considered as a sort of surrogate of the “minimum distance method” whose consistency is proved in Theorem 1. Hence, one could perhaps expect a slightly worse performance for these “interpolation estimators”, when compared with those obtained by minimum distance between V_n and V. The reason would be very much the same why the maximum likelihood estimators are generally preferable to those obtained by the methods of moments: they make a more extensive use of the available information. However, the limited experimental results we show in the paper suggest in fact that both estimators are very close in performance. See for example, the outputs in Tables 1 and 4 compared with those, obtained by (11), shown in Tables 6 and 7. This is an interesting practical conclusion as the interpolation methodology in (11) is computationally faster (it only requires to compute $V_n(r)$ for $d+1$ values of r).

Moreover, in some simple cases (11) could lead to explicit expressions for the estimators. Let us assume, for instance, that $S \subset \mathbb{R}^2$ is a compact set with $r_0 = \mathrm{reach}(S) > 0$ such that the Euler characteristic of S is one (so $\theta_2 = \pi$). Then,

$$V(r) = \pi r^2 + L(S) r + \mu(S), \ \text{for } 0 \leq r < r_0,$$

so that $L(S)$ and $\mu(S)$ can be estimated as the solutions of the system $V_n(r_j) = V(r)$ for $j = 1, 2$ which can easily expressed in a explicit way in terms of the values $V_n(r_j)$.

	$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$		$\lvert\hat{\theta}_2 - \theta_2\rvert/\theta_2$	
Sample size	Mean	SD	Mean	SD	Mean	SD
$n = 500$	0.06673	0.00825	0.01544	0.00336	0.00203	0.00076
$n = 1000$	0.04242	0.00494	0.00972	0.00194	0.00129	0.00046
$n = 5000$	0.01460	0.00135	0.00333	0.00052	0.00044	0.00015

Table 6: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0, 1)$, $i = 0, 1, 2$. For each sample, the estimations $\hat{\theta}_i$ are obtained using the *interpolation method* from $V_n(r_1), V_n(r_2)$ and $V_n(r_3)$, where r_1, r_2 and r_3 are randomly selected in the interval $[1, 2]$.

	$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$	
Sample size	Mean	SD	Mean	SD
$n = 500$	0.15705	0.01577	0.00411	0.00206
$n = 1000$	0.09960	0.00903	0.00262	0.00122
$n = 5000$	0.03385	0.00257	0.00086	0.00040

Table 7: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0, 5) \setminus \text{int}(B(0, 4))$, $i = 0, 1$ (assumming $\theta_2 = 0$ known). For each sample, the estimations $\hat{\theta}_i$ are obtained using the *interpolation method* from $V_n(r_1)$ and $V_n(r_2)$, where r_1 and r_2 are randomly selected in the interval $[2, 3.5]$.

4.2 Some three-dimensional examples: the ball and the torus

We now show the practical performance of our method in the three-dimensional Euclidean space. We have considered two different sets $S \subset \mathbb{R}^3$, that satisfy the polynomial volume assumption (2). The estimations of the coefficients θ_i, $i = 0, \ldots, 3$, were obtained using the interpolation method described in (11) since, as commented before, it is computationally faster and the results obtained in the bidimensional case with this methodology were comparable to those obtained with the methodology of Theorem 1.

Thus, let us first consider the unit ball $S = B(0, 1)$ in $\mathbb{R}^3$. Note that S satisfies (2) for all $r \in \mathbb{R}$. In Table 8 we show, for different sample sizes, the mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ samples from the uniform distribution on S, $i = 0, \ldots, 3$. The estimations $\hat{\theta}_i$ are obtained by solving the system of equations in (11) where r_j, $j = 1, \ldots, 4$, are randomly selected in the interval $[10, 15]$. We can compare the results in Table 8 with those obtained if we estimate θ_0 and θ_1 with $\hat{\theta}_0 = \mu(S_n)$ and $\hat{\theta}_1 = L(S_n)$, where S_n is the convex hull of the sample and $\mu(S_n)$ and $L(S_n)$ its volume and surface area, respectively, see Figure 2. Results are summarized in Table 9.

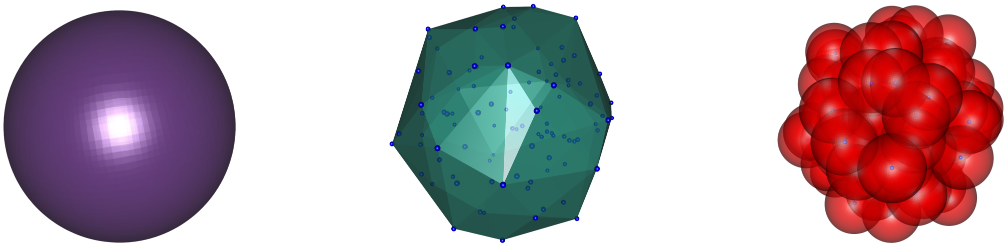

Figure 2: Left, $S = B(0,1)$ in $\mathbb{R}^3$. Middle, uniform sample of size $n = 100$ in S and convex hull of the sample in green. Right, in red, $S_n(r)$ for a given value of r.

	$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$		$\lvert\hat{\theta}_2 - \theta_2\rvert/\theta_2$		$\lvert\hat{\theta}_3 - \theta_3\rvert/\theta_3$	
Sample size	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$n = 2000$	0.13073	0.00505	0.05617	0.00264	0.02290	0.00119	0.00000	0.00000
$n = 5000$	0.08381	0.00264	0.03578	0.00133	0.01453	0.00060	0.00000	0.00000
$n = 10000$	0.05974	0.00153	0.02537	0.00078	0.01027	0.00034	0.00000	0.00000

Table 8: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0,1)$ in $\mathbb{R}^3$, $i = 0, \ldots, 3$. For each sample, the estimations $\hat{\theta}_i$ are obtained using the *interpolation method* from $V_n(r_j)$, with $j = 1, \ldots, 4$, where r_j are randomly selected in the interval $[10, 15]$.

As an example of non-convex set in $\mathbb{R}^3$, we consider a torus S with major radius 5 and minor radius 1, see Figure 3. In Table 10, we show the mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for S (for simplicity we only show the results corresponding

	$\lvert\hat{\theta}_0 - \theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1 - \theta_1\rvert/\theta_1$	
Sample size	Mean	SD	Mean	SD
$n = 2000$	0.09688	0.00426	0.05642	0.00264
$n = 5000$	0.06213	0.00218	0.03594	0.00133
$n = 10000$	0.04422	0.00129	0.02549	0.00078

Table 9: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = B(0, 1)$ in $\mathbb{R}^3$, $i = 0, 1$. Now $\hat{\theta}_0 = \mu(S_n)$ and $\hat{\theta}_1 = L(S_n)$, where S_n stands for the convex hull of the sample and $\mu(S_n)$ and $L(S_n)$ denote its volume and surface area, respectively.

to $\hat{\theta}_0$ and $\hat{\theta}_1$). Again, the estimations $\hat{\theta}_i$ are obtained by solving the system of equations in (11) where now r_j are randomly selected in the interval $[3, 4]$. It is important to note that, unlike the two-dimensional non-convex case, we cannot compare our results in Table 10 to any other procedure based on a more specific information on the set of interest (we are not aware of any implementation that supports the computation of the α-convex hull of the sample for $d = 3$).

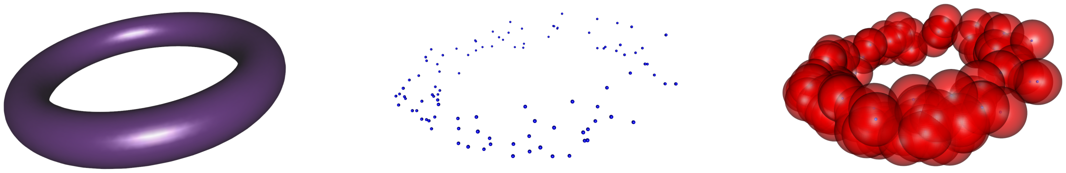

Figure 3: Left, torus S with major radius 5 and minor radius 1. Middle, uniform sample of size $n = 100$ in S. Right, in red, $S_n(r)$ for a given valur of r.

4.3 Possible extensions of the method

Lemma 1 establishes that the Hausdorff convergence of S_n to S entails the convergence in measure for the corresponding parallel sets. Since $S_n = \mathcal{X}_n$ is d_H-consistent, our proposal is

	$\lvert\hat{\theta}_0-\theta_0\rvert/\theta_0$		$\lvert\hat{\theta}_1-\theta_1\rvert/\theta_1$	
Sample size	Mean	SD	Mean	SD
$n=2000$	0.48674	3.49394	0.14970	1.41482
$n=5000$	0.21870	0.19926	0.06006	0.09791
$n=10000$	0.14743	0.03044	0.03932	0.01578

Table 10: Mean value and standard deviation of $|\hat{\theta}_i-\theta_i|/\theta_i$ over $B=500$ repeats for a torus S with major radius 5 and minor radius 1, $i=0,1$. For each sample, the estimations $\hat{\theta}_i$ are obtained using the *interpolation method* from $V_n(r_j)$, with $j=1,\ldots,4$, where r_j are randomly selected in the interval $[3,4]$.

to estimate $V(r)$ using this very simple estimator. But of course, the volume of the parallel set might be also approximated using more sophisticated estimators S_n. For example, if we assume in advance that S is convex (or α-convex), this information can be incorporated by estimating the "true" polynomial function $V(r)$ of the target set S in terms of the volume functions of the sample convex hull (or the sample α-convex hull). These estimators of $V(r)$ should be, in principle, more accurate, as they carry additional information on the set S. However, the computation of the parallel volumes for such sets is far from simple, even in dimension $d=2$, and especially for the α-convex hull, whose structure is considerably involved. So, in those cases the direct plug-in estimators considered in Tables 2 and 9 look as a quite natural alternative.

If no shape restriction is assumed, one could also think of estimating S by using $S_n = B(\mathcal{X}_n,\varepsilon_n)$, where ε_n is a sequence of smoothing parameters, see Devroye and Wise (1980). In that case, $V(r)$ might be estimated from $V_n(r)=\mu(B(\mathcal{X}_n,\varepsilon_n+r))$. To illustrate these ideas, we consider a regular polytope. Let $S=[0,1]^2$ be the unit square in $\mathbb{R}^2$. For different sample sizes, we generate $B=500$ samples from the uniform distribution on S. For each sample, we applied the "interpolation method" described in Subsection 4.1 with $V_n(r_j)$ computed as $\mu(B(S_n,r_j))$, where S_n is (a) the sample $\mathcal{X}_n$, (b) the convex hull of the sample and (c) the Devroye-Wise estimator $B(\mathcal{X}_n,\varepsilon_n)$ for different values of the smoothing parameter ε_n. Results are summarized in Table 11. The best results are obtained using the Devroye-Wise estimator, for which the corresponding calculations can be made quite efficiently (at least for dimensions 2 and 3). Note that, however, the results in case (c) depend largely on an adequate choice of the smoothing parameter ε_n. The dilated convex hull (b) slightly outperforms our proposal (as a drawback, it is harder to compute in higher dimensions). The overall conclusions suggest a good performance of our estimator, in terms of both, economy of assumptions and computational simplicity. The results of the comparison of (a) and (c) in the unit cube $S = [0,1]^3 \subset \mathbb{R}^3$ can be found in the Supplementary material document. We also include in the Supplementary material a brief simulation study with non uniform samples.

			$\|\hat{\theta}_0 - \theta_0\|/\theta_0$		$\|\hat{\theta}_1 - \theta_1\|/\theta_1$		$\|\hat{\theta}_2 - \theta_2\|/\theta_2$	
Sample size			Mean	SD	Mean	SD	Mean	SD
$n = 500$	(a)		0.05758	0.00782	0.04363	0.01136	0.00085	0.00032
	(b)		0.03197	0.00872	0.04704	0.01112	0.00016	0.00042
	(c)	$\varepsilon_n = 0.005$	0.03833	0.00794	0.03580	0.01136	0.00085	0.00031
		$\varepsilon_n = 0.01$	0.01892	0.00805	0.02799	0.01132	0.00084	0.00031
		$\varepsilon_n = 0.04$	0.10083	0.00889	0.01949	0.01050	0.00080	0.00029
$n = 1000$	(a)		0.03612	0.00485	0.03123	0.00786	0.00057	0.00019
	(b)		0.01842	0.00669	0.03356	0.00811	0.00020	0.00081
	(c)	$\varepsilon_n = 0.005$	0.01664	0.00493	0.02339	0.00786	0.00057	0.00019
		$\varepsilon_n = 0.01$	0.00472	0.00344	0.01563	0.00771	0.00056	0.00019
		$\varepsilon_n = 0.04$	0.12414	0.00558	0.03148	0.00786	0.00053	0.00018
$n = 5000$	(a)		0.01219	0.00133	0.01436	0.00347	0.00021	0.00007
	(b)		0.00506	0.00449	0.01539	0.00380	0.00018	0.00006
	(c)	$\varepsilon_n = 0.005$	0.00761	0.00136	0.00655	0.00341	0.00021	0.00006
		$\varepsilon_n = 0.01$	0.02757	0.00139	0.00306	0.00212	0.00021	0.00006
		$\varepsilon_n = 0.04$	0.15062	0.00165	0.04843	0.00347	0.00020	0.00006

Table 11: Mean value and standard deviation of $|\hat{\theta}_i - \theta_i|/\theta_i$ over $B = 500$ repeats for $S = [0,1]^2$, $i = 0,1,2$. For each sample, the estimations $\hat{\theta}_i$ are obtained using the *interpolation method* from $V_n(r_1), V_n(r_2)$ and $V_n(r_3)$, where r_1, r_2 and r_3 are randomly selected in the interval $[1,2]$. $V_n(r_j)$ is computed as $\mu(B(S_n, r_j))$, where S_n is (a) the sample $\mathcal{X}_n$, (b) the convex hull of the sample and (c) the Devroye-Wise estimator $B(\mathcal{X}_n, \varepsilon_n)$, for different values of ε_n.

4.4 Some technical details

All computations were carried out in R, (R Core Team (2017)). In the computation of the volume of the union of a family of 3D balls, we have used the Structural Bioinformatics Library (SBL), a C++/Python API by Cazals and Dreyfus (2016). The convex hull of the sample in $\mathbb{R}^3$, as well as its volume and surface area is computed using the R-package *geometry* by Habel et al. (2015). For the dilated convex hull in $\mathbb{R}^2$ we used the R-package *polyclip* by Johnson and Baddeley (2017).

5 Discussion

The assumption of polynomial volume, as stated in Definition 1 is reasonably general for practical purposes, including applications in image analysis. As discussed above, condition (2) not only applies to the broad class of sets with positive reach but also covers other sets with inward non-smooth peaks in their boundaries.

Unlike other methods for boundary measure estimation (see Cuevas et al. (2007), Jiménez and Yukich (2011), among others) the polynomial volume (PV) method outlined here only requires a sample inside the set $S \subset \mathbb{R}^d$: no external sample points are needed. Also, no auxiliary set estimator S_n of S is required, except for the sample itself. Likewise, the PV method can perform the simultaneous estimation of $\mu(S)$ and $L(S)$ and we do not require uniformity over S for the distribution of the sample points.

The role of the parameter R in our PV method is worth of some comments. Such parameter concerns the set S, that is, the target of the estimation. Hence it should be seen as a regularity parameter (or a smoothness parameter) very much in the same way as the number of derivatives or the Hölder exponent that we assume in an underlying density function to be estimated. Therefore R should not be confused with the tuning or smoothing parameters commonly found in nonparametrics (e.g. in density estimation). The crucial difference lies in the fact that such smoothing or tuning parameters appear only in the expressions of the estimators: they usually measure the extent at which such estimators are "close" to the data and are typically chosen to asymptotically achieve efficiency in the estimation. Note that in our case we must also fix the interval $[a, b] \subset (0, R)$ where the minimum distance will be performed but a and b are not properly smoothing parameters to be chosen in an optimal way.

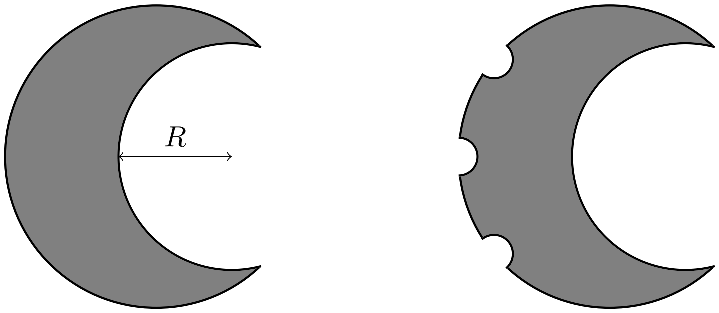

Figure 4: For the set in gray in the left, (2) holds for $[0, R)$. For the set in gray in the right (2) only holds for a small interval $[0, R/6)$, where $R/6$ is radius of the small indentations in the boundary.

If R is large enough, any choice of b close to R and a “not too close” to 0 will do the job.

However, it should be noted that the value of R in assumption (2) is relevant. If R is too small, the PV method will not work in practice, unless extremely large sample sizes are considered. This is quite intuitive, see Figure 4: if (2) only holds for a small interval $[0, R)$ then, S will be relatively irregular and, consequently, $\mu(S)$ and $L(S)$ will be harder to estimate.

The PV method is flexible enough to incorporate additional geometric information on the target set S, whenever such information can be translated in terms of the coefficients of $V(r)$. This is the case, of the Euler Poincaré characteristic of S which, as explained in Section 2, is directly related to the higher order coefficient θ_d. In the numerical experiments of Section 4 we have shown the effect of incorporating the knowledge of this coefficient to the PV estimation process.

In general, the empirical results shown in Section 4 suggest that the method is competitive, even if we compare it with other procedures based on a more specific information on the set S. Again, we should stress the ease of implementation even for case $S \subset \mathbb{R}^3$ which is very difficult to address with other approaches.

The main theoretical open problem connected to the PV method is the study of the asymptotic distribution and convergence rates of the estimates $\hat{\theta}_j$. This is a non-trivial issue which could be perhaps tackled with techniques similar to those employed in the study of minimum distance estimators.

Supplementary material. Supplementary material file includes some additional numerical experiments with data in $\mathbb{R}^3$ and non-uniform samples.

Acknowledgements. This work has been partially supported by Spanish Grants MTM2016-78751-P (A. Cuevas) and MTM2016-76969-P (B. Pateiro-López).

References

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

Amenta, N., Choi, S., Dey, T.K., Leekha, N. (2002). A simple algorithm for homeomorphic surface reconstruction. *Internat. J. Comput. Geom. Appl.* 12, 125-141.

Arias Castro, E., Rodríguez-Casal, A. (2016). On estimating the perimeter using the alpha-shape. *Ann. Inst. H. Poincaré Probab. Statist.* 53, 1051-1068.

Arias-Castro, E., Pateiro-López, B. and Rodríguez-Casal, A. (2016). Minimax estimation of the volume of a set with smooth boundary. Manuscript arXiv:1605.01333v1.

Armendáriz, I., Cuevas, A. and Fraiman, R. (2009). Nonparametric estimation of boundary measures and related functionals: asymptotic results. *Adv. in Appl. Probab.* 41, 311–322.

Baíllo, A. and Cuevas, A. (2006). Parametric versus nonparametric tolerance regions in detection problems. *Computational Statistics* 21, 523–536.

Baldin, N. and Reiss, M. (2016). Unbiased estimation of the volume of a convex body. *Manuscript*, *Stoch. Proc. Appl.* 126, 3716–3732.

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

Brunel, V. E. (2016). Adaptive estimation of convex and polytopal density support. *Probability Theory and Related Fields* 164, 1-16.

Canzonieri, V. and Carbone, A. (1998). Clinical and biological applications of image analysis in non-Hodgkin's lymphomas. *Hematol. Oncol.* 16, 15–28.

Cazals, F. and Dreyfus, T. (2016) The Structural Bioinformatics Library: modeling in biomolecular science and beyond. *Research Report RR-8957, Inria.* http://sbl.inria.fr

Cuevas, A. and Rodríguez-Casal, A. (2004). On boundary estimation. *Adv. in Appl. Probab.* 36, 340–354.

Cuevas, A. (2009). Set estimation: Another bridge between statistics and geometry. *BEIO* 25, 71–85.

Cuevas, A., Fraiman, R. and Pateiro-López, B. (2012). On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.* 44, 311–329.

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

Cuevas, A., Fraiman, R. and Györfi, L. (2013) Towards a universally consistent estimator of the Minkowski content. *ESAIM Probab. Stat.* 17, 359–369.

DeVore, R. A. (1986). Approximation of functions. In *Approximation Theory*, C. de Boor, ed., pp. 1-20. Proceedings of Symposia in Applied Mathematics, 36. AMS Short Course Lecture Notes. American Mathematical Society, Providence, RI.

Devroye, L. and Wise, G. (1980). Detection of abnormal behavior via nonparametric, estimation of the support. *SIAM J. Appl. Math.* 3, 480–488.

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

Getz, W. M. and Wilmers, C. C. (2004). A local nearest-neighbor convex-hull construction of home ranges and utilization distributions. *Ecography* 27, 489–505.

Habel, K., Grasman, R., Gramacy, R. B., Stahel, A. and Sterratt, D. C. (2015) geometry: Mesh Generation and Surface Tesselation. R package version 0.3-6. https://CRAN.R-project.org/package=geometry

Heveling, M., Hug, D. and Last, G. (2004). Does polynomial parallel volume imply convexity? *Math. Ann* 328, 469–479.

Jang, W. (2006). Nonparametric density estimation and clustering in astronomical sky survey. *Comp. Statist. & Data Anal.* 50, 760–774.

Johnson, A. and Baddeley, A. (2017). polyclip: Polygon Clipping. R package version 1.6-1. https://CRAN.R-project.org/package=polyclip

Jiménez, R. and Yukich, J. E. (2011). Nonparametric estimation of surface integrals. *Ann. Statist.* 39, 232–260.

Kie, J.G., Matthiopoulos, J., Fieberg, J. Powell, R. A., Cagnacci, F., Mitchell, M. S., Gaillard, J.-M. and Moorcroft, P. R. (2010). The home-range concept: are traditional estimators still relevant with modern telemetry technology? *Phil. Trans. R. Soc. B* 365, 2221–2231.

Kim, J. C. and Korostelëv, A. P. (2000). Estimation of smooth functionals in image models. *Math. Methods Statist.* 9, 140–159.

Korostelëv, A. P. and A. B. Tsybakov (1993). *Minimax theory of image reconstruction*, Volume 82 of *Lecture Notes in Statistics.* Springer-Verlag, New York.

Laha, R. G. and Rohatgi, V. K. (1979). *Probability Theory.* Wiley, New York.

Montero, R.S. and Bribiesca, E. (2009). State of the art of compactness and circularity measures. *International Mathematical Forum* 4, 1305–1335.

Moore, M. (1984). On the estimation of a convex set. *Ann. Statist.* 12, 1090-1099.

Morvan, J. M. (2008). *Generalized Curvatures.* Springer-Verlag, Berlin

Pardon, J. (2011). Central limit theorems for random polygons in an arbitrary convex set. *Ann. Probab.* 39, 881-903.

Pateiro-López, B. and Rodríguez-Casal, A. (2010) Generalizing the convex hull of a sample: The R package alphahull. *J. Statist. Softw.* 5, 1–28.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rinaldo, A. and Wasserman, L. (2010). Generalized density clustering. *Ann. Statist.* 38, 2678–2722.

Ripley, B. D., and Rasson, J.-P. (1977). Finding the edge of a Poisson forest. *J. Appl. Probability* 14, 483–491.

Rodríguez-Casal, A. (2007). Set estimation under convexity-type assumptions. *Ann. Inst. H. Poincaré Probab. Statist.* 43, 763–774.

Simar, L. and Wilson, P. (2000). Statistical inference in nonparametric frontier models: The state of the art. *J. Prod. Anal.* 13, 49–78.

Stachó, L. L. (1976). On the volume function of parallel sets. *Acta Sci. Math.* 38, 365–374.

Willett, R. M. and Novak, R. D. (2007). Minimax optimal level set estimation. *IEEE Trans. Image Process.* 16, 2965–2979.