ESAIM: PS 20 (2016) 332–348
DOI: 10.1051/ps/2016015

ESAIM: Probability and Statistics
www.esaim-ps.org

A FULLY DATA-DRIVEN METHOD FOR ESTIMATING THE SHAPE OF A POINT CLOUD

A. Rodríguez-Casal[1] and P. Saavedra-Nieves[1]

Abstract. Given a random sample of points from some unknown distribution, we propose a new data-driven method for estimating its probability support S. Under the mild assumption that S is r-convex, the smallest r-convex set which contains the sample points is the natural estimator. The main problem for using this estimator in practice is that r is an unknown geometric characteristic of the set S. A stochastic algorithm is proposed for selecting its optimal value from the data under the hypothesis that the sample is uniformly generated. The new data-driven reconstruction of S is able to achieve the same convergence rates as the convex hull for estimating convex sets, but under a much more flexible smoothness shape condition.

Mathematics Subject Classification. 62G05, 62G20.

Received 11 August 2015. Revised 11 April 2016. Accepted May 23, 2016.

1. Introduction

Support estimation deals with the problem of reconstructing the compact and nonempty support $S \subset \mathbb{R}^d$ of an absolutely continuous random vector X assuming that a random sample $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ from X is given. A closely related topic is perimeter estimation. Also, under uniform assumptions on the distribution of the sample, Arias–Castro and Rodríguez–Casal [1] estimated the length of the boundary of the support.

The question of reconstructing the support has different but quite natural responses depending on the available information on S. For example, if no assumptions are made *a priori* on the shape of the support S, Chevalier [7] and Devroye and Wise [12] proposed a general purpose estimator which is just a sort of *dilated* version of $\mathcal{X}_n$. Specifically,

$$S_n = \bigcup_{i=1}^{n} B_{\epsilon_n}[X_i],$$

where $B_{\epsilon_n}[X_i]$ denotes the closed ball centered at X_i with radius ϵ_n, a sequence of smoothing parameters which must tend to zero but not too quickly in order to achieve consistency. See also Grenander [16], Cuevas [9], Korostelëv and Tsybakov [18], Cuevas and Rodríguez-Casal [11] or more recent references like Genovese *et al.* [15] where the boundary of this estimator is converted into an estimator of one-dimensional curves and the rates of convergence are determined. The main disadvantage of the Devroye–Wise estimator is its dependence on the unknown and influential radius of the balls ϵ_n. Small values of ϵ_n provide split estimators whereas for large

Keywords and phrases. Support estimation, r-convexity, uniformity, maximal spacing.

[1] Department of Statistics and Operations Research, University of Santiago de Compostela, Spain. paula.saavedra@usc.es

Article published by EDP Sciences

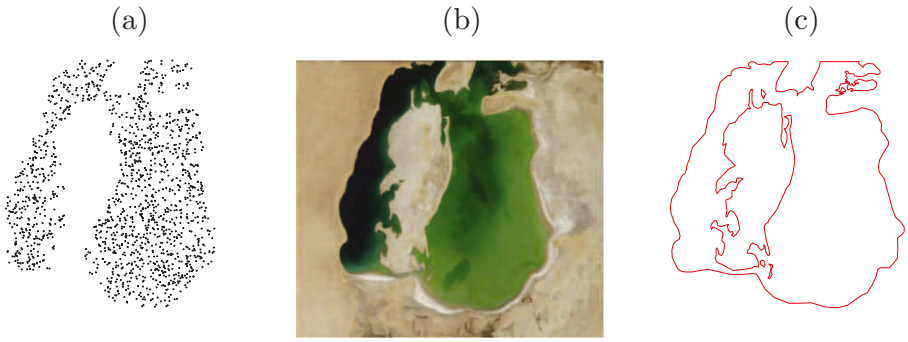

FIGURE 1. (a) $\mathcal{X}_{1500}$ on the Aral Sea. (b) Aral Sea's image from the Moderate Resolution Imaging Spectroradiometer on NASA's Terra satellite in 2000. (c) Aral Sea's boundary.

values of ϵ_n the estimator could considerably overestimate S. Baíllo *et al.* [5] and Baíllo and Cuevas [3] suggested two general methods for selecting the parameter ϵ_n assuming that S is connected and star-shaped, respectively.

However, more sophisticated alternatives, that can achieve better error rates, could be used if some *a priori* information about the shape of S is available. For instance, if the support is assumed to be convex then the convex hull of the sample points, $H(\mathcal{X}_n)$, provides a natural support estimator. This is just the intersection of all convex sets containing $\mathcal{X}_n$. For analyzing in depth this estimator, see Schneider [25, 26], Dümbgen and Walther [13] or Reitzner [22].

In practice, the convexity property may be too restrictive. In Figure 1, a uniform sample $\mathcal{X}_{1500}$ drawn inside the Aral Sea is shown. It was generated digitizing an image of the NASA's Terra satellite. Taking into account the shape of the contour of the Aral Sea in Figure 1, reconstructing it from the previous sample of points under convex shape assumptions does not seem convenient. So, it can be useful to introduce a more flexible shape condition such as r-convexity. A closed set $A \subset \mathbb{R}^d$ is said to be r-convex, for some $r > 0$, if $A = C_r(A)$, where

$$C_r(A) = \bigcap_{\{B_r(x):B_r(x)\cap A=\emptyset\}} (B_r(x))^c$$

denotes the r-convex hull of A and $B_r(x)$, the open ball with center x and radius r. The r-convex hull is closely related to the closing of A by $B_r(0)$ from the mathematical morphology (see Serra [27]). It can be shown that

$$C_r(A) = (A \oplus rB) \ominus rB,$$

where $B = B_1(0)$, $\lambda C = \{\lambda c : c \in C\}$, $C \oplus D = \{c + d : \ c \in C, d \in D\}$ and $C \ominus D = \{x \in \mathbb{R}^d : \ \{x\} \oplus D \subset C\}$, for $\lambda \in \mathbb{R}$ and sets C and D.

If it is assumed that S is r-convex, $C_r(\mathcal{X}_n)$ is the natural estimator for the unknown support. This estimator is well known in the computational geometry literature for producing good global reconstructions if the sample points are (approximately) uniformly distributed on the set S. See Edelsbrunner [14] for a survey on the subject. In fact, although the r-convexity is a more general restriction than the convexity, $C_r(\mathcal{X}_n)$ can achieve the same convergence rates than $H(\mathcal{X}_n)$ (see Rodríguez-Casal [24]). However, this estimator depends on the unknown parameter r. Figure 2 shows its influence by using the random sample on the Aral Sea presented in Figure 1. Small values of r provide estimators almost equal to $\mathcal{X}_n$. However, if large values of r are considered then $C_r(\mathcal{X}_n)$ practically coincides with $H(\mathcal{X}_n)$ (see Fig. 2d).

Most of the available results in literature about support estimation make special emphasis on asymptotic properties, especially consistency and convergence rates but they do not give any criterion for selecting the unknown parameters such as the parameter ϵ_n in S_n or r in $C_r(\mathcal{X}_n)$ from the sample. The aim of this paper is to overcome this drawback and present a method for selecting the parameter r for the r-convex hull estimator

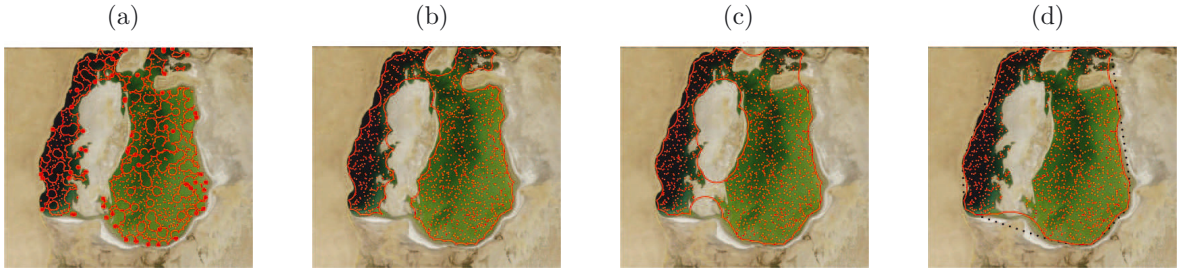

FIGURE 2. The boundary of $C_r(\mathcal{X}_{1500})$ is shown in red for (a) $r = 10$, (b) $r = 25$, (c) $r = 40$ and (d) $r = 90$. The boundary of $H(\mathcal{X}_{1500})$ is shown in dotted line in (d). (Color online)

from the available data. It should be noted that this problem, for the bidimensional case, has already been studied in literature by Mandal and Murthy [19]. They proposed a selector for r based on the concept of minimum spanning tree but only consistency of the method was provided. The optimality issues were not considered.

The automatic selection criterion which will be proposed in this work is based on a very intuitive idea. As it can be seen in Figures 2c or 2d, land areas are contained in $C_r(\mathcal{X}_n)$ if the selected r is too large. So, the estimator contains a big ball (or spacing) empty of sample points. Janson [17] calibrated the size of the maximal spacing when the sample distribution is uniform on S. Recently, Berrendero *et al.* [6] used this result to test uniformity when the support is unknown. Here, we will follow the somewhat opposite approach. We will assume that $\mathcal{X}_n$ follows a uniform distribution on S and if a big enough spacing is found in $C_r(\mathcal{X}_n)$ then it is concluded that r is too large. Here, it is proposed to select the largest value of r compatible with the uniformity assumption on $C_r(\mathcal{X}_n)$.

Once the parameter r is estimated, it is natural to go back to the support estimation problem. An automatic estimator for S, based on the estimator of r, is proposed in this paper. Two metrics between sets are usually considered in order to assess the performance of a support estimator. Let A and C be two closed, bounded, nonempty subsets of $\mathbb{R}^d$. The Hausdorff distance between A and C is defined by

$$d_H(A, C) = \max\left\{\sup_{a\in A} d(a, C), \sup_{c\in C} d(c, A)\right\},$$

where $d(a, C) = \inf\{\|a - c\| : c \in C\}$ and $\|\ \|$ denotes the Euclidean norm. On the other hand, if A and C are two bounded and Borel sets then the distance in measure between A and C is defined by $d_\mu(A, C) = \mu(A \triangle C)$, where μ denotes the Lebesgue measure and $\triangle$, the symmetric difference, that is, $A \triangle C = (A \setminus C) \cup (C \setminus A)$. Hausdorff distance quantifies the physical proximity between two sets whereas the distance in measure is useful to quantify their similarity in content. However, neither of these distances are completely useful for measuring the similarity between the shape of two sets. The Hausdorff distance between boundaries, $d_H(\partial A, \partial C)$, can be also used to evaluate the performance of the estimators, see Baíllo and Cuevas [3], Cuevas and Rodríguez-Casal [11], Rodríguez-Casal [24] or Genovese *et al.* [15].

This paper is organized as follows. In Section 2, the optimal smoothing parameter of $C_r(\mathcal{X}_n)$ to be estimated is formally defined. The new data-driven algorithm for selecting it is presented in Section 3. Consistency of this estimator is established in Section 4. In addition, a new estimator for the support S is proposed. It is showed that it is able to achieve the same convergence rates as the convex hull for estimating convex sets. The numerical questions involving the practical application of the algorithm are analyzed in Section 5. In Section 6, the performances of the new selector and Mandal and Murthy [19]'s method will be analyzed through a simulation study. Conclusions are exposed in Section 7. Finally, proofs are deferred to Section 8.

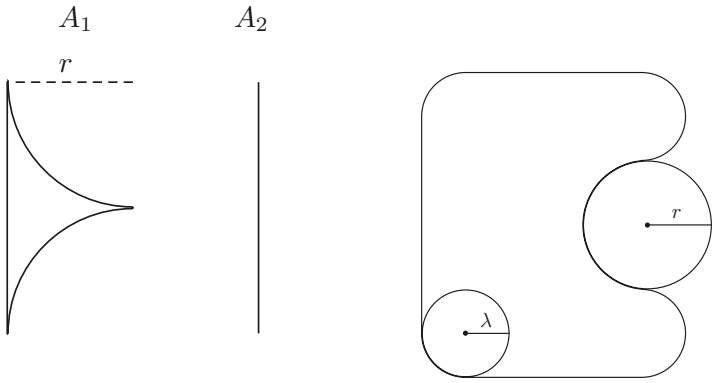

FIGURE 3. $A_1 \cup A_2$ fulfills the r-rolling condition $\not\Rightarrow$ $A_1 \cup A_2$ is r-convex (*left*). (R_λ^r) is a more general condition (*right*).

2. Optimal parameter and shape restrictions

The problem of reconstructing a r-convex support S using a data-driven procedure can be solved if the parameter r is estimated from a random sample of points $\mathcal{X}_n$ taken in S. Next, it will be presented an algorithm to do this. The first step is to determine precisely the optimal value of r to be estimated. It is established in Definition 2.1. We propose to estimate the highest value of r which verifies that S is r-convex.

Definition 2.1. Let $S \subset \mathbb{R}^d$ a compact, nonconvex and r-convex set for some $r > 0$. It is defined

$$r_0 = \sup\{\gamma > 0 : C_\gamma(S) = S\}. \tag{2.1}$$

For simplicity in the exposition, it is assumed that S is not convex. Of course, if S is convex r_0 would be infinity. In Proposition 2.4, it is proved that, under mild regularity conditions, the supreme established in (2.1) is a maximum, that is, S is r_0-convex too. Under this hypothesis, the optimality of the smoothing parameter defined in (2.1) can be justified. It is clear that S is r-convex for $r \leq r_0$ but if $r < r_0$, $C_r(\mathcal{X}_n)$ is a non admisible estimator since it is always outperformed by $C_{r_0}(\mathcal{X}_n)$. This is because, with probability one, $C_r(\mathcal{X}_n) \subset C_{r_0}(\mathcal{X}_n) \subset S$ and hence, $d_\mu(C_{r_0}(\mathcal{X}_n), S) \leq d_\mu(C_r(\mathcal{X}_n), S)$ (the same holds for the Hausdorff distance). It should also noted that, for $r > r_0$, $C_r(\mathcal{X}_n)$ would considerably overestimate S, specially if S has a big hole inside, see Figure 5a below. The mild regularity condition we need to prove Proposition 2.4 is slightly stronger than than r-convexity:

(R_λ^r) S fulfills the r-rolling property and S^c fulfills the λ-rolling condition.

Following Cuevas *et al.* [10], it is said A satisfies the (outside) r-rolling condition if each boundary point $a \in \partial A$ is contained in a closed ball with radius r whose interior does not meet A. There exist interesting relationships between this property and r-convexity. In particular, Cuevas *et al.* [10] proved that if A is compact and r-convex then A fulfills the r-rolling condition. According to Figure 3 (*left*), the reciprocal is not always true. Proposition 2.2 shows that (R_λ^r) is a (mild) sufficient condition to ensure the r-rolling condition implies r-convexity. However, we think that the equivalence between r-convexity and r-rolling is much more general and it can be proved under even milder conditions. Condition (R_λ^r) was essentially analyzed in Walther [28,29] where only the case $r = \lambda$ is taken into account. In this work, the radius λ can be different from r, see Figure 3 (*right*). Walther [28,29] proved that, under (R_r^r), S is r-convex. Proposition 2.2 extends this property since, for $\lambda < r$, Walther's result would only imply λ-convexity but not r-convexity. So, for sets satisfying (R_λ^r), r-convexity is ensured, even for very small values of λ.

Proposition 2.2. *Let $S \subset \mathbb{R}^d$ be a nonempty, compact support verifying (R^r_λ). Then, S is r-convex.*

Proposition 2.2 is the key for proving that r_0 is a maximum, see Proposition 2.4 below. To see this, let be $\{r_n\}$ a sequence converging to r_0 such that $C_{r_n}(S) = S$. This sequence always exists, see Definition 2.1. We know, using the results in Cuevas *et al.* [10], that S satisfies the r_n-rolling condition. But, by Proposition 2.3 this property is preserved in the limit, so S is r_0-rolling. We do not know if r-convexity is also preserved by taking the limit in the parameter r. Finally, under (R^r_λ), r_0-rolling implies that S is r_0-convex.

Proposition 2.3. *Let $A \subset \mathbb{R}^d$ be a closed set. Let $\{r_n\}$ be a sequence of positive terms converging to $\overline{r}$. If A fulfills the r_n-rolling condition, for all n, then A fulfills the $\overline{r}$-rolling condition.*

Proposition 2.4. *Let $S \subset \mathbb{R}^d$ be a nonempty, compact and nonconvex set verifying (R^r_λ) and let r_0 be the parameter defined in (2.1). Then, $C_{r_0}(S) = S$ and, as consequence, S fulfills the r_0-rolling condition.*

Remark 2.5. Under certain conditions of S (for instance, $\mathrm{Int}(H(S)) \neq \emptyset$), it is verified that $C_\infty(S) = H(S)$ where $C_\infty(S) = \lim_{r_n \to \infty} C_{r_n}(S)$. Therefore, if S is assumed to be convex, Proposition 2.4 remains true. For more details, see Walther [29].

3. Selection of the optimal smoothing parameter

The uniformity test proposed in Berrendero *et al.* [6] has been considered in order to estimate r_0 from $\mathcal{X}_n$. This test is based on the multivariate spacings, see Janson [17]. In the univariate case, spacings are defined as the length of gaps between sample points. For general dimension d, the maximal spacing of S is defined as

$$\Delta_n(S) = \sup\{\gamma : \exists x \text{ with } B_\gamma[x] \subset S \setminus \mathcal{X}_n\}.$$

The value of the maximal spacing depends only on S and on the sample points $\mathcal{X}_n$. The Lebesgue measure (volume) of the balls with radius $\Delta_n(S)$ is denoted by $V_n(S)$. Berrendero *et al.* [6] used the Janson [17] (1987)'s Theorem to introduce a uniformity test on the distribution of $\mathcal{X}_n$. They consider the problem of testing

$$H_0 : \ X \text{ is uniform with support } S.$$

With significance level α, H_0 will be rejected if

$$V_n(S) > \frac{a(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n}, \tag{3.1}$$

where $a = \mu(S)$, u_α denotes the $1-\alpha$ quantile of a random variable U with distribution

$$\mathbb{P}(U \leq u) = \exp(-\exp(-u)) \text{ for } u \in \mathbb{R} \tag{3.2}$$

and the value of the constant β that does not depend on S is explicitly given in Janson [17]. Concretely,

$$\beta = \frac{1}{d!}\left(\frac{\sqrt{\pi}\Gamma\left(\frac{d}{2}+1\right)}{\Gamma\left(\frac{d+1}{2}\right)}\right)^{d-1}.$$

In particular, for the bidimensional case, $\beta = 1$. If S is unknown this test can not be directly applicable. Under the (R^r_λ) condition with $\lambda = r$, Berrendero *et al.* [6] considered $S_n = C_r(\mathcal{X}_n)$ as the estimator of S, but no data-driven method was provided for selecting r. The maximal spacing of S is estimated by

$$\hat{\Delta}_n = \sup\{\gamma : \exists x \text{ with } B_\gamma[x] \subset S_n \setminus \mathcal{X}_n\}, \tag{3.3}$$

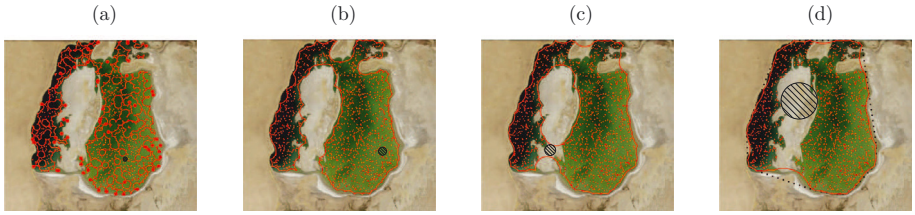

FIGURE 4. Maximal spacing of $C_r(\mathcal{X}_{1500})$ is shown in dashed lines for (a) $r = 10$, (b) $r = 25$, (c) $r = 40$ and (d) $r = 90$. The boundary of $H(\mathcal{X}_{1500})$ is shown in dotted line in (d).

and the critical region (3.1) can be replaced by

$$\hat{V}_{n,r} > \hat{c}_{n,\alpha,r} = \frac{a_n(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n},$$

where $a_n = \mu(C_r(\mathcal{X}_n))$ and $\hat{V}_{n,r}$ denotes the volume of the ball of radius $\hat{\Delta}_n$, see (3.3).

Figure 4 shows the maximal spacings for the estimators of the Aral Sea considered in Figure 2. A bad choice (a big value) of the smoothing parameter allows to detect a large gap, clearly incompatible with the uniformity hypothesis, see Figure 4d for $r = 90$. This means that the estimator contains a large spacing which is not contained in the Aral Sea. Since the sample is uniform on the original support, we can conclude that the smoothing parameter is too large. It must be selected smaller. The estimator of r_0 is based on this idea. If we assume that the distribution is uniform on S, and according to Definition 2.1, r_0 will be estimated by

$$\hat{r}_0 = \sup\{\gamma > 0 : H_0 \text{ is accepted on } C_\gamma(\mathcal{X}_n)\}. \tag{3.4}$$

The technical aspects for the estimator defined in (3.4) are considered in Sections 4 and 5.

4. Main results

The existence of the supreme defined in (3.4) must be guaranteed. Theorem 4.1 will show that this is the case and $\hat{r}_0$ consistently estimates r_0

Theorem 4.1. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R_λ^r) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (2.1) and $\hat{r}_0$ defined in (3.4). Let $\{\alpha_n\} \subset (0,1)$ be a sequence converging to zero verifying $\lim_{n\to\infty} \log(\alpha_n)/n = 0$. Then, $\hat{r}_0$ converges to r_0 in probability.*

Remark 4.2. We assume that S is not convex only for simplicity in the exposition. If S is convex it can be shown that $\hat{r}_0$ goes to infinity (which is the value of r_0 in this case) because, with high probability, the test is not rejected for all values of r.

Once the consistency of the estimator defined in (3.4) has been proved, it would be natural to study the behavior of the random set $C_{\hat{r}_0}(\mathcal{X}_n)$ as an estimator for the support S. In particular, if $\lim_{r\to r_0^+} d_H(S, C_r(S)) = 0$ then consistency of $C_{\hat{r}_0}(\mathcal{X}_n)$ can be proved easily from Theorem 4.1. However, the consistency can not be guaranteed if $d_H(S, C_r(S))$ does not go to zero as r goes to r_0 from above (as $\hat{r}_0$ does, see Proposition 8.2 below). This problem can be solved by considering the estimator $C_{r_n}(\mathcal{X}_n)$ where $r_n = \nu\hat{r}_0$ with $\nu \in (0,1)$ fixed. This ensures that, for n large enough, with high probability $C_{r_n}(\mathcal{X}_n) \subset S$. From the practical point of view the selection of ν is not a major issue because $\hat{r}_0$ is numerically approximated and the computed estimator always satisfies this property without multiplying by ν. In some sense, Theorem 4.3 gives the convergence rate of the numerical approximation of $\hat{r}_0$, see Section 5 for the details on the computation of the estimator.

Theorem 4.3. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in the (2.1) and $\hat{r}_0$ defined in (3.4). Let $\{\alpha_n\} \subset (0,1)$ be a sequence converging to zero under the conditions of Theorem 4.1. Let be $\nu \in (0,1)$ and $r_n = \nu \hat{r}_0$. Then,*

$$d_H(S, C_{r_n}(\mathcal{X}_n)) = O_P\left(\frac{\log n}{n}\right)^{\frac{2}{d+1}}.$$

The same convergence order holds for $d_H(\partial S, \partial C_{r_n}(\mathcal{X}_n))$ and $d_\mu(S \triangle C_{r_n}(\mathcal{X}_n))$.

Remark 4.4. Theorem 4.3 shows that $C_{r_n}(\mathcal{X}_n)$ achieves the same convergence rates as the convex hull of the sample for reconstructing convex sets.

Remark 4.5. The selector proposed by Mandal and Murthy [19], r_n^{MM}, goes to zero in probability. In Pateiro-López and Rodríguez-Casal [21] it is proved that, for a deterministic sequence of parameters d_n ($d_n \leq r_0$ and $d_n^2 n/\log(n) \to \infty$), the convergence rate (in probability) for the distance in measure is, for the bidimensional case, $d_n^{-1/3} n^{-2/3}$. This is the convergence rate of the new proposal plus a penalizing term $d_n^{-1/3}$ which goes to infinity if $d_n \to 0$. It is expected that this penalizing factor, $(r_n^{MM})^{-1/3}$ also appears for the the Mandal and Murthy's proposal.

5. Numerical aspects of the algorithm

Although a fully data-driven method for estimating the optimal parameter established in (2.1) has been proposed from a theoretical point of view, its practical implementation depends on the specification of two parameters that must be selected by the practitioner.

Next, the main numerical aspects are considered in order to detail the algorithm completely. With probability one, for n large enough, the existence of the estimator defined in (3.4) is guaranteed under the hypotheses of Theorem 4.1. However, in practice, this estimator might not exist for a specific sample $\mathcal{X}_n$ and a given value of the significance level α. Therefore, the influence of α must be taken into account. The null hypothesis will be (incorrectly) rejected on $C_r(\mathcal{X}_n)$ for $0 < r \leq r_0$ with probability α approximately. This is not important from the theoretical point of view, since we are assuming that $\alpha = \alpha_n$ goes to zero as the sample size increases. But, what should be done, for a given sample, if H_0 is rejected for *all* r (or at least *all* reasonable values of r)? In order to fix a minimum acceptable value of r, it is assumed that S (and, hence, the estimator) will have no more than C connected components. Too split estimators will not be considered even in the case that we reject H_0 for all r. The minimum value that ensures a number of connected components not greater than C will be taken in this latter case, see below. Therefore, this parameter C can be interpreted as a geometric stopping criteria that does not appear in theoretical results because the sequence α_n tends to zero.

Dichotomy algorithms can be used to compute $\hat{r}_0$. The practitioner must select a maximum number of iterations I and two initial points r_m and r_M with $r_m < r_M$ such that the null hypothesis of uniformity is rejected and accepted on $C_{r_M}(\mathcal{X}_n)$ and $C_{r_m}(\mathcal{X}_n)$, respectively. According to the previous comments, it is assumed that the number of connected components of $C_{r_m}(\mathcal{X}_n)$ must not be greater than C. Choosing a value close enough to zero is usually sufficient to select r_m. However, if selecting this r_m is not possible because, for very low and positive values of r, the hypothesis of uniformity is still rejected on $C_r(\mathcal{X}_n)$ then r_0 is estimated as the positive closest value to zero r such that the number of connected components of $C_r(\mathcal{X}_n)$ is smaller than or equal to C. On the other hand, if the hypothesis of uniformity is accepted even on $H(\mathcal{X}_n)$ then we propose $H(\mathcal{X}_n)$ as the estimator for the support.

To sum up, the following inputs should be given: the significance level $\alpha \in (0,1)$, a maximum number of iterations I, a maximum number of connected components C and two initial values r_m and r_M. Given these parameters $\hat{r}_0$ will be computed as follows:

(1) In each iteration and while the number of them is smaller than I:
   (a) $r = (r_m + r_M)/2$.

(b) If the null hypothesis is not rejected on $C_r(\mathcal{X}_n)$ then $r_m = r$.

(c) Otherwise, $r_M = r$.

(2) Then, $\hat{r}_0 = r_m$.

According to the correction of the bias proposed by Ripley and Rasson [23] for the convex hull estimator, Berrendero *et al.* [6] suggested rejecting the null hypothesis of uniformity when

$$\hat{V}_{n,r} > \frac{\mu(S_n)(u_\alpha + \log n + (d-1)\log\log n + \log\beta)}{n - v_n},$$

where v_n denotes the number of vertices of $S_n = C_r(\mathcal{X}_n)$ (points of $\mathcal{X}_n$ that belong to ∂S_n). In this work, it is proposed to redefine the critical region as

$$\hat{V}_{n,r} > \hat{c}^*_{n,\alpha,r},$$

where $\hat{c}^*_{n,\alpha,r}$ is equal to

$$\frac{\mu(S_n)(u_\alpha + \log(n - v_n) + (d-1)\log\log(n - v_n) + \log\beta)}{n - v_n},$$

that is, we suggest to replace n by $n - v_n$ in the definition of $\hat{c}_{n,\alpha,r}$ elsewhere not only in the denominator. Although the main theoretical results in Section 4 are established in terms of $\hat{c}_{n,\alpha,r}$ instead of $\hat{c}^*_{n,\alpha,r}$, the proofs are completely analogous in both cases since v_n is negligible with respect to n. See, for instance, the upper bound for the expected number of vertices in Theorem 3 by Pateiro–López and Rodríguez–Casal [21].

Some technical aspects related to the computation of the maximal spacings must be also considered. Testing the null hypothesis of uniformity is a procedure repeated I times in this algorithm. This may seem to be very computing intensive since the test involves calculating the maximal spacing. However, we do not need to know the exact value of the maximal spacing since we are not interested in computing the test statistic. In fact, it is only necessary to check if, for a fixed r, $C_r(\mathcal{X}_n)$ contains an open ball, that does not intersect the sample points with volume greater than the test's critical value $\hat{c}^*_{n,\alpha,r}$. In other words, we will simply check if an open ball of radius equal to $\hat{c}^*_{n,\alpha,r}$ and center x is contained in $C_r(\mathcal{X}_n) \setminus \mathcal{X}_n$. If this disc exists then $x \notin B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$ where

$$B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n) = \bigcup_{X_i \in \mathcal{X}_n} B_{\hat{c}^*_{n,\alpha,r}}(X_i)$$

is the dilation of radius $\hat{c}^*_{n,\alpha,r}$ of the sample. Therefore, the centers of the possible maximal balls necessarily lie outside $B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$. Following Berrendero *et al.* [6], to check if the null hypothesis of uniformity is rejected on $C_r(\mathcal{X}_n)$, we will follow the next steps:

(1) Determine the set $D(r) = C_r(\mathcal{X}_n) \cap \partial B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$. Notice that, if $x \in D(r)$ then $B_{\hat{c}^*_{n,\alpha,r}}(x) \cap \mathcal{X}_n = \emptyset$.
(2) Calculate $M(r) = \max\{d(x, \partial C_r(\mathcal{X}_n) : x \in D(r)\}$.
(3) If $M(r) \leq \hat{c}^*_{n,\alpha,r}$ then the null hypothesis of uniformity is not rejected.

It should be noted that $\partial C_r(\mathcal{X}_n)$ and $\partial B_{\hat{c}^*_{n,\alpha,r}}(\mathcal{X}_n)$ can be easily computed (at least for the bidimensional case), see Pateiro-López and Rodríguez-Casal [20].

6. Simulation study

The performances of the algorithm proposed in this paper and Mandal and Murthy's method [19] will be analyzed in this section. They will be denoted by RS and MM, respectively. A total of 1000 uniform samples of four different sizes n have been generated on three support models in the Euclidean space $\mathbb{R}^2$, see Figure 5.

The first set, $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$, is a circular ring with $r_0 = 0.15$. The other two ones are two interesting sets, $S = \mathbf{C}$ and $S = \mathbf{S}$ with $r_0 = 0.2$ and $r_0 = 0.0353$, respectively. The values of n considered are $n = 100$, $n = 500$, $n = 1000$ and $n = 1500$. In addition, four values for α have been taken into account, $\alpha_i = 10^{-i}$, $i = 1, \ldots, 4$. The maximum number of connected components C was fixed equal to 4.

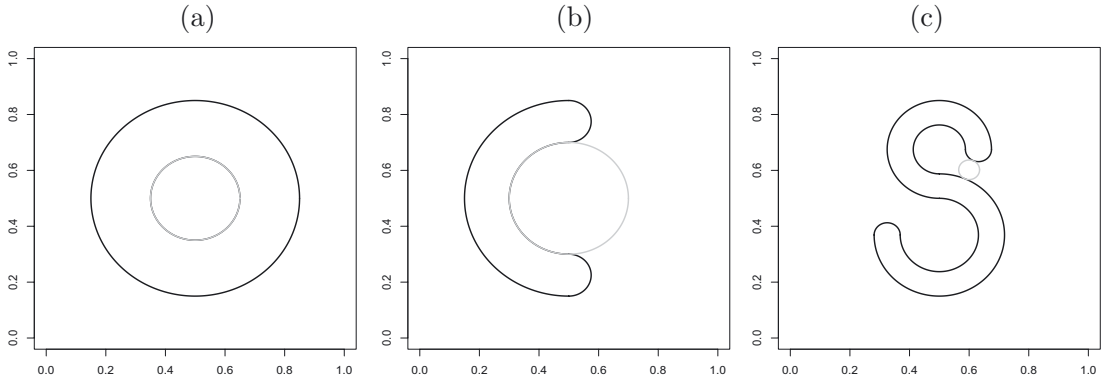

FIGURE 5. (a) $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$ with $r_0 = 0.15$. (b) $S = \mathbf{C}$ with $r_0 = 0.2$. (c) $S = \mathbf{S}$ with $r_0 = 0.0353$. Circles in gray have radius equal to r_0.

TABLE 1. Empirical means of 1000 RS and MM estimations for the smoothing parameter of the r-convex hull with $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$. In this case, $r_0 = 0.15$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.1592	0.1456	0.1438	0.1410
	$\alpha_2 = 10^{-2}$	0.1592	0.1509	0.1499	0.1495
	$\alpha_3 = 10^{-3}$	0.1592	0.1516	0.1507	0.1503
	$\alpha_4 = 10^{-4}$	0.1592	0.1517	0.1507	0.1504
MM		0.1969	0.1295	0.1084	0.0977

TABLE 2. Empirical means of 1000 RS and MM estimations for the smoothing parameter of the r-convex hull with $S = \mathbf{C}$. In this case, $r_0 = 0.2$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.2724	0.2007	0.1903	0.1888
	$\alpha_2 = 10^{-2}$	0.2929	0.2150	0.2056	0.2032
	$\alpha_3 = 10^{-3}$	0.2982	0.2188	0.2089	0.2055
	$\alpha_4 = 10^{-4}$	0.2988	0.2226	0.2105	0.2068
MM		0.1636	0.1072	0.0897	0.0809

TABLE 3. Empirical means of 1000 RS and MM estimations for the smoothing parameter of the r-convex hull with $S = \mathbf{S}$. In this case, $r_0 = 0.0353$.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.0954	0.0833	0.0637	0.0548
	$\alpha_2 = 10^{-2}$	0.0954	0.0878	0.0695	0.0602
	$\alpha_3 = 10^{-3}$	0.0958	0.0886	0.0736	0.0631
	$\alpha_4 = 10^{-4}$	0.1077	0.0887	0.0778	0.0659
MM		0.1644	0.1055	0.088	0.0792

For each fixed random sample, both estimators of the smoothing parameter of the r-convex hull have been calculated. So, one thousand estimations have been obtained for each algorithm, fixed a model and the values of n and α. The empirical means of these one thousand estimations are showed in Tables 1–3 for the RS and

TABLE 4. Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.9288	0.3293	0.2085	0.1623
	$\alpha_2 = 10^{-2}$	0.9288	0.3143	0.1970	0.1492
	$\alpha_3 = 10^{-3}$	0.9294	0.3123	0.1957	0.1484
	$\alpha_4 = 10^{-4}$	0.9288	0.3122	0.1957	0.1483
MM		1.4165	0.3378	0.2316	0.1837
		0.9337	0.2956	0.1819	0.1364

TABLE 5. Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = \mathbf{C}$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.6041	0.1472	0.0920	0.0712
	$\alpha_2 = 10^{-2}$	0.6677	0.1589	0.0833	0.0640
	$\alpha_3 = 10^{-3}$	0.6820	0.1953	0.0832	0.0631
	$\alpha_4 = 10^{-4}$	0.6837	0.2440	0.0865	0.0626
MM		0.4145	0.1681	0.1125	0.0885
		0.3727	0.1277	0.0800	0.0606

TABLE 6. Empirical means of 1000 estimations (multiplied by 10) obtained for the distance in measure between $S = \mathbf{S}$ and the resulting support estimators for RS and MM methods. The last row contains the benchmarks (multiplied by 10) for each sample size.

	n	100	500	1000	1500
RS	$\alpha_1 = 10^{-1}$	0.6389	0.2591	0.1842	0.1485
	$\alpha_2 = 10^{-2}$	0.6389	0.2537	0.1821	0.1455
	$\alpha_3 = 10^{-3}$	0.6411	0.2530	0.1821	0.1464
	$\alpha_4 = 10^{-4}$	0.6797	0.2529	0.1816	0.1476
MM		1.2319	0.4851	0.2445	0.1514
		1.0794	0.3320	0.2038	0.1541

MM methods. We should mention that MM method is included in these table only for illustrative purposes. The results of these two algorithms are not directly comparable since the goal of MM is not to estimate the parameter r_0 defined in (2.1). However, comparing the behavior of the two resulting support estimators can be really interesting. Tables 4–6 contain the empirical means (multiplied by 10) of one thousand Monte Carlo estimations for the distance in measure between the RS and MM support estimators and the corresponding theoretical models, respectively. In addition, we have estimated the distance in measure between the r_0-convex hull of each sample and its corresponding support model for the different sample sizes. The means of these estimations can be considered as a benchmark. They are showed (multiplied by 10) in the last row of Tables 4–6. A grid of 334^2 points was considered in the unit square for estimating the distance in measure. The parameter ν was fixed equal to 0.95 for calculating the RS support estimator.

Figure 6 contains the boxplots for the estimations of the distance in measure between the resulting support estimators for the RS and MM methods when $n = 1500$.

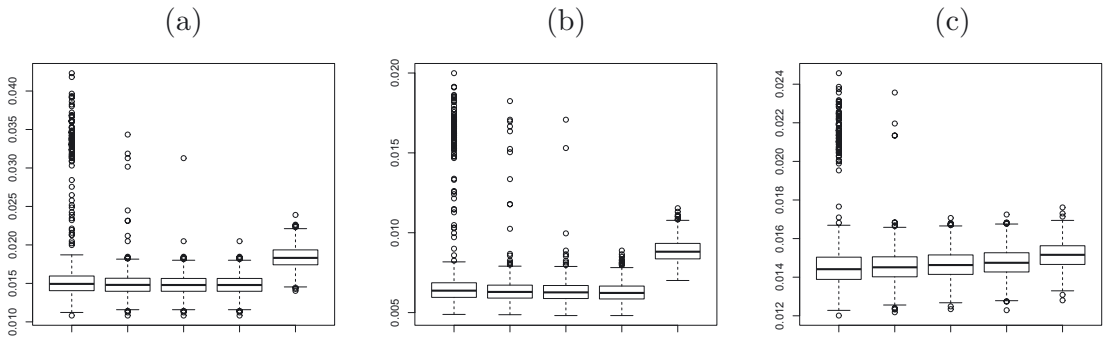

FIGURE 6. Boxplots of the estimations for the distance in measure for RS and MM methods when $n = 1500$ for (a) $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$, (b) $S = \mathbf{C}$ and (c) $S = \mathbf{S}$. From left to right, RS considering α_1, α_2, α_3 and α_4 and MM.

Conclusions of the simulation study. According to the results showed in Tables 1–3, RS presents a good global behavior for estimating the smoothing parameter r_0. Only when $S = \mathbf{C}$ and $n = 100$, MM provides better results, see Table 2. In this particular case, the estimations of RS are specially greater than 0.2, the real value of parameter r_0. In general, MM provides too small estimations, mainly for high values of the sample size, see Tables 1 and 2.

The influence of the level of significance α must be also discussed. Taking low values of α reduces the number of outliers considerably for the three support models presented. In addition, if the model considered is not too complex then small values of α provide better results for n large enough reducing the risk of rejecting the null hypothesis of uniformity when it is satisfied, see for instance $S = B_{0.35}[(0.5, 0.5)] \setminus B_{0.15}((0.5, 0.5))$ or $S = \mathbf{C}$ in Tables 1 and 2. Therefore, excessively low values of r will not be selected. However, if the support model is not so simple then choosing large values of α provides better estimations for the smoothing parameter, see Table 3 for $S = \mathbf{S}$. Anyway, for moderate and large values of the sample size the dependence on α of RS method is small.

Although the results are not shown, the role of the parameter C has been studied too. Concretely, simulations for C equal to 2, 4, 6 and 8 were analyzed. The influence of C is not relevant since the consideration of different values practically provides the same results.

Finally and according to Tables 4–6, RS always provides the smallest estimation errors for the criteria considered except when $S = \mathbf{C}$ with $n = 100$ or even $n = 500$ if the value of α is too large (see Tab. 5). Therefore, RS support estimator is more competitive than MM algorithm. According to the previous comments, it can be seen that the number of outliers for RS increases if large values of α are considered for the three support models (see Fig. 6).

7. Conclusions and discussion

A theoretical automatic estimator for the optimal parameter defined in (2.1) has been proposed under (R^r_λ). But, its practical implementation depends on the specification of the significance level α for the uniformity test and the maximum number of connected components C for the resulting support estimator. According to Section 6, their influence on estimations is not so important. However, both of them must be selected for avoiding too small values of the smoothing parameter or, equivalently, split estimators for the support. As practical recommendation, small values of α and large values of C are preferred. Choosing a small α reduces the number of too low estimations and provides, in general, the best results. As for the parameter C, a larger value than the real number of connected components of the support should be selected. Due to this number is

rarely known and the value of C has not a strong influence on estimations, selecting a large value for it will be enough.

Finally, natural extensions of this work will be discussed. Although the geometric property (R_λ^r) is very flexible, it would be very interesting to consider different families of sets. For instance, the parameter ρ for the family of ρ cone-sets could be estimated using similar ideas, see Cholaquidis *et al.* [8]. Of course, under this new shape condition, consistency results for the new procedure should be investigated. Another important achievement would be to generalize this work for non-uniform distributions. Janson [17] derived the asymptotic distribution of the maximal spacing under uniform assumptions. Aaron *et al.* [1] obtained the asymptotic distribution of the maximal spacing when data are generated from a positive, bounded support Lipchitz continuous density function. However, the definition of a spacing in Aaron *et al.* [1] depends on the density function that, in practice, is unknown. A plug-in estimator of the maximal spacing could be proposed using a kernel density estimator.

8. Proofs

In this section the proofs of the stated theorems are presented.

Proof of Proposition 2.2. An auxiliary result is necessary. For $a \in \partial A$, Lemma 8.1 relates the uniqueness of a unit vector $\eta(a)$ and the existence of some $x \in A$ such that a coincides metric projection of x onto A.

Lemma 8.1. *Let $A \subset \mathbb{R}^d$ be a nonempty and closed set and $a \in \partial A$. Let us assume that there exists $x \notin A$ such that $\rho = \|x - a\| = d(x, A)$, that is, the point a is a metric projection of x onto A. If exists $\lambda > 0$ and a unit vector $\eta(a)$ such that $B_\lambda[a - \lambda\eta(a)] \subset A$, then $x = a + \rho\eta(a)$.*

Proof. To see this suppose the contrary, that is, let us suppose that exists x verifying the required conditions with $x \neq a + \rho\eta(a)$.

Then, x, a and $a - \lambda\eta(a)$ can not lie on the same line and hence,

$$\|a - \lambda\eta(a) - x\| < \|a - \lambda\eta(a) - a\| + \|a - x\| = \lambda + \rho. \tag{8.1}$$

Let $z \in \partial B_\lambda[a - \lambda\eta(a)] \cap [x, a - \lambda\eta(a)]$, where $[x, a - \lambda\eta(a)]$ denotes the line segment with endpoints x and $a - \lambda\eta(a)$ (see Fig. 7, *left*). Then,

$$\|a - \lambda\eta(a) - x\| = \|a - \lambda\eta(a) - z\| + \|z - x\| = \lambda + \|z - x\|.$$

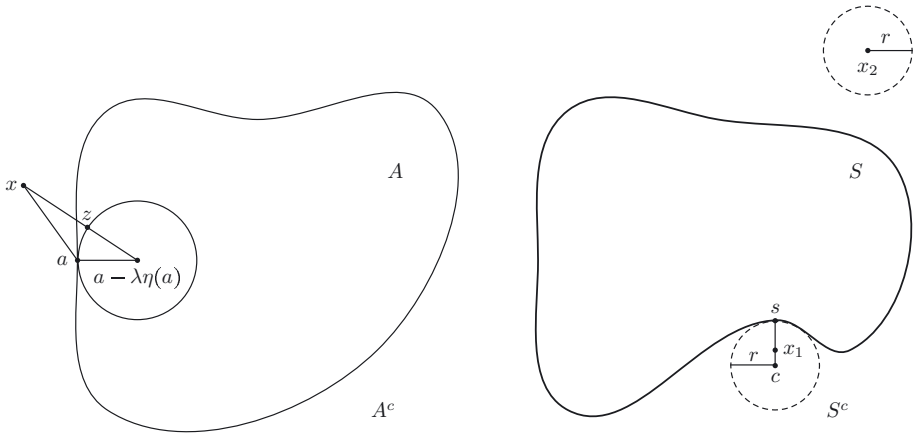

Figure 7. Elements of Lemma 8.1 (*left*). Elements of Proposition 2.2 with $d(x_1, S) < r$ and $d(x_2, S) > r$ (*right*).

According to (8.1), $\|z-x\| = \|a-\lambda\eta(a)-x\| - \lambda < \lambda+\rho-\lambda = \rho$, which is a contradiction since $z \in A$ and $\rho = d(x,A)$. □

Now, we are in position to conclude the proof of Proposition 2.2. Let us prove that $S = C_r(S)$. Since $S \subset C_r(S)$, it is enough to check if $C_r(S) \subset S$. Equivalently, it will be checked that for all $x \in S^c$ there exists an open ball of radius r containing x. This ball will not intersect S. Let us fix $x \notin S$. If $d(x,S) \geq r$ then $x \in B_r(x)$ and $B_r(x) \cap S = \emptyset$.

Otherwise, if $d(x,S) < r$, let s be a projection of x on S and let us define $\rho = d(x,S) = \|x-s\|$. The λ-rolling property allow to prove easily that, $B_\lambda[s-\lambda\eta(s)] \subset S$ where $\eta(s) = (s-x)/\rho$. According to Lemma 8.1, $x = s+\rho\eta(s)$. In addition, $s \in \partial S$ and, according to the imposed conditions, S fulfills the r-rolling property. So,

$$\exists c \in \mathbb{R}^d \text{ such that } s \in B_r[c] \text{ and } B_r(c) \cap S = \emptyset.$$

According to Lemma 8.1, $c = s + r\eta(s)$ since s is a projection of c on S. We are supposing that $\rho < r$. So, $\|x-c\| = \|(\rho-r)\eta(s)\| = r-\rho < r$. Then, $x \notin C_r(S)$ since that $x \in B_r(c)$ and $B_r(c) \cap S = \emptyset$.

Figure 7 (*right*) shows the elements used in the proof of Proposition 2.2. □

Proof of Proposition 2.3. It is verified that

$$\forall a \in \partial A \text{ and } \forall n \in \mathbb{N}\ \exists x_n \text{ such that } a \in B_{r_n}[x_n] \text{ and } B_{r_n}(x_n) \cap A = \emptyset.$$

For each $a \in \partial A$, let us consider the sequence of closed balls $\{B_{r_n}[x_n]\}$. It is clear that it can be assumed with no loss of generality $\{x_n\}$ converges to some point x_a since $\{x_n\}$ is a bounded sequence and it contains a convergent subsequence which we will denoted by $\{x_n\}$ again.

Since $a \in \partial A \subset A$, $a \in B_{r_n}[x_n]$ and $B_{r_n}(x_n) \cap A = \emptyset$, it follows that $d(x_n,A) = r_n$ and $\|x_n - a\| = r_n$. Hence,

$$r_n = d(x_n,A) \leq \|x_n - x_a\| + d(x_a,A).$$

Therefore, $d(x_a,A) \geq \overline{r}$. In particular, this implies $B_{\overline{r}}(x_a) \cap A = \emptyset$ and $\|x_a - a\| \geq \overline{r}$. On the other hand,

$$\|x_a - a\| \leq \|x_a - x_n\| + \|x_n - a\| = \|x_a - x_n\| + r_n.$$

So $\|x_a - a\| \leq \overline{r}$, that is, $a \in B_{\overline{r}}[x_a]$. This implies that A fulfills the $\overline{r}$-rolling condition. □

Proof of Proposition 2.4. It will be proved that $r_0 \in \{\gamma > 0 : C_\gamma(S) = S\}$. According to the properties of the supreme,

$$\exists \{r_n\} \subset \{\gamma > 0 : C_\gamma(S) = S\} \text{ such that } \lim_{n\to\infty} r_n = r_0.$$

Then, $C_{r_n}(S) = S$, for all $n \in \mathbb{N}$. Proposition 2 in Cuevas *et al.* [10] ensures that S fulfills the r_n-rolling property for all n. Then, S fulfills the r_0-rolling property, see Proposition 2.3. Taking into account the imposed restrictions, it is verified that S^c satisfies the λ-rolling condition. So, it is possible to guarantee that S is under $(R_\lambda^{r_0})$. According to Proposition 2.2, S is r_0-convex. Using again Proposition 2 in Cuevas *et al.* [10], it is possible to guarantee that S fulfills the r_0-rolling property. □

Proof of Theorem 4.1. Some auxiliary results are necessary. First we will prove that, with probability tending to one, $\hat{r}_0$ is at least as big as r_0.

Proposition 8.2. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R_λ^r) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (2.1) and $\{\alpha_n\} \subset (0,1)$ a sequence converging to zero. Then,*

$$\lim_{n\to\infty} \mathbb{P}(\hat{r}_0 \geq r_0) = 1.$$

Proof. From the definition of $\hat{r}_0$, see (3.4), it is clear that

$$\mathbb{P}(\hat{r}_0 \geq r_0) \geq \mathbb{P}(\hat{V}_{n,r_0} \leq \hat{c}_{n,\alpha_n,r_0}),$$

where, remember, $\hat{V}_{n,r_0}$ denotes the volume of the maximal spacing in $C_{r_0}(\mathcal{X}_n)$, $\hat{c}_{n,\alpha_n,r_0} = \mu(C_{r_0}(\mathcal{X}_n))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta)\cdot n^{-1}$ and u_{α_n} satisfies $\mathbb{P}(U \leq u_{\alpha_n}) = 1-\alpha_n$ and U is the random variable defined in (3.2). Since, with probability one, $C_{r_0}(\mathcal{X}_n) \subset S$, we have $\hat{V}_{n,r_0} \leq V_n(S)$ where $V_n(S)$ denotes de volume of a ball with radius the maximal spacing of S. Hence,

$$\mathbb{P}(\hat{r}_0 \geq r_0) \geq \mathbb{P}(V_n(S) \leq \hat{c}_{n,\alpha_n,r_0}) = \mathbb{P}\left(\frac{u_{\alpha_n}}{A_n}U_n \leq u_{\alpha_n}\right),$$

where

$$U_n = \frac{nV_n(S)}{\mu(S)} - \log n - (d-1)\log\log n - \log\beta$$

and

$$A_n = \frac{n\hat{c}_{n,\alpha_n,r_0}}{\mu(S)} - \log n - (d-1)\log\log n - \log\beta.$$

According to the Janson's Theorem [17], $U_n \xrightarrow{d} U$. In addition, it can be easily proved that $u_{\alpha_n}/A_n \xrightarrow{P} 1$. This can be done by taking into account that

$$\mu(C_{r_0}(\mathcal{X}_n))/\mu(S) = 1 + O_P((\log(n)/n)^{2/(d+1)}),$$

see Theorem 3 in Rodríguez-Casal [24]. Now, according to the Slutsky's Lemma, $(u_{\alpha_n}/A_n)U_n \xrightarrow{d} U$. Notice that U has a continuous distribution, so convergence in distribution implies that

$$\sup_u |\mathbb{P}\left((u_{\alpha_n}/A_n)U_n \leq u\right) - \mathbb{P}(U \leq u)| \to 0.$$

Since $\mathbb{P}(U \leq u_{\alpha_n}) = 1-\alpha_n$ and $\alpha_n \to 0$, this ensures that

$$\mathbb{P}\left((u_{\alpha_n}/A_n)U_n \leq u_{\alpha_n}\right) \to 1.$$

Therefore, $\mathbb{P}(\hat{r}_0 \geq r_0) \to 1$. □

It remains to prove that $\hat{r}_0$ cannot be arbitrarily larger that r_0. The following lemma ensures that, for a given $\gamma > r_0$, there exists an open ball contained in $C_\gamma(S)$ which does not meet S.

Lemma 8.3. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and let be $\gamma > 0$ such that $S \subsetneq C_\gamma(S)$. Then, there exists $\epsilon > 0$ and $x \in C_\gamma(S)$ such that $B_\epsilon(x) \subset C_\gamma(S)$ and $B_\epsilon(x) \cap S = \emptyset$.*

Proof. Let us assume, for a moment, that we can find $s \in \partial S$ such that $s \in \mathrm{Int}(C_\gamma(S))$. In this case, there exists $\rho > 0$ satisfying that $B_\rho(s) \subset C_\gamma(S)$. On the other hand, by assumption, S is r_0-convex which implies, by Proposition 2 in Cuevas *et al.* [10], that S fulfills the r_0-rolling condition. This ensures that there exists a ball $B_{r_0}(y)$ such that $s \in B_{r_0}[y]$ and $B_{r_0}(y) \cap S = \emptyset$. It is clear that we can find an open ball $B_\epsilon(x)$ such that $B_\epsilon(x) \subset B_{r_0}(y) \cap B_\rho(s)$. By construction $B_\epsilon(x) \subset B_{r_0}(y)$ and, hence, $B_\epsilon(x) \cap S = \emptyset$. Finally, $B_\epsilon(x) \subset B_\rho(s)$ and, therefore, $B_\epsilon(x) \subset C_\gamma(S)$. This would finished the proof in this case.

It only remains to show that $\partial S \subset \partial C_\gamma(S)$ leads to a contradiction. First, the hypothesis $\partial S \subset \partial C_\gamma(S)$ imply that S satisfy the γ-rolling condition. This is a straightforward consequence of Proposition 2 in Cuevas *et al.* [10] since $C_\gamma(S)$ is γ-convex. But the γ-rolling condition imply, under the (R^r_λ) shape restriction, γ-convexity, see Proposition 2.2. This is a contradiction since we are assuming that $S \subsetneq C_\gamma(S)$. □

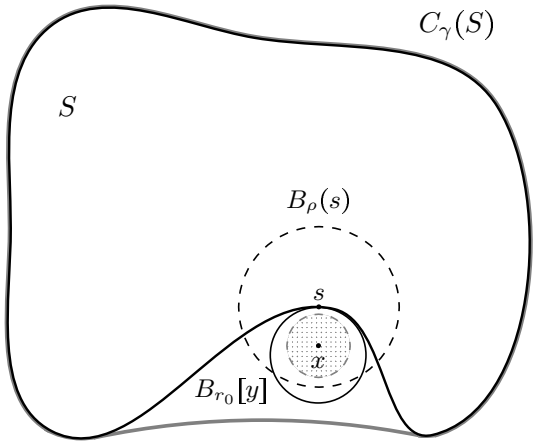

FIGURE 8. Elements of proof in Lemma 8.3. ∂S in black, $\partial C_\gamma(S)$ in gray, $B_\rho(s)$, $B_{r_0}[y]$ and $B_\epsilon(x)$ in gray.

Lemma 8.4. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (2.1). Then, for all $r > r_0$, there exists an open ball $B_\rho(x)$ such that $B_\rho(x) \cap S = \emptyset$ and*

$$\mathbb{P}\left(B_\rho(x) \subset C_r(\mathcal{X}_n),\ \textit{eventually}\right) = 1.$$

Proof. Let be r^* such that $r > r^* > r_0$. Since $C_{r_0}(S) = S \subsetneq C_{r^*}(S)$, according to Lemma 8.3,

$$\exists B_\epsilon(x) \text{ such that } B_\epsilon(x) \subset C_{r^*}(S) \text{ and } B_\epsilon(x) \cap S = \emptyset.$$

It can be assumed, without loss of generality, that $r \leq \frac{\epsilon}{2} + r^*$. If this is not the case then it would be possible to replace r^* by $r^{**} > r^*$ satisfying $r^{**} < r \leq \frac{\epsilon}{2} + r^{**}$. For this r^{**},

$$B_\epsilon(x) \subset C_{r^*}(S) \subset C_{r^{**}}(S) \text{ and } B_\epsilon(x) \cap S = \emptyset.$$

Now, we can apply Lemma 3 in Walther [28] in order to ensure that

$$\mathbb{P}\left(S \oplus r^*B \subset \mathcal{X}_n \oplus rB,\ \text{eventually}\right) = 1.$$

If $S \oplus r^*B \subset \mathcal{X}_n \oplus rB$ then $(S \oplus r^*B) \ominus r^*B \subset (\mathcal{X}_n \oplus rB) \ominus r^*B$, that is, $C_{r^*}(S) \subset (\mathcal{X}_n \oplus rB) \ominus r^*B$. This imply that

$$C_{r^*}(S) \ominus (r - r^*)B \subset ((\mathcal{X}_n \oplus rB) \ominus r^*B) \ominus (r - r^*)B.$$

In addition,

$$((\mathcal{X}_n \oplus rB) \ominus r^*B) \ominus (r - r^*)B = (\mathcal{X}_n \oplus rB) \ominus rB = C_r(\mathcal{X}_n),$$

where we have used that, for sets A, C and D, $(A \ominus C) \ominus D = A \ominus (C \oplus D)$. Finally, since $B_\epsilon(x) \subset C_{r^*}(S)$ and $\epsilon/2 \geq (r - r^*)$, we have $B_{\epsilon/2}(x) \subset C_{r^*}(S) \ominus (\epsilon/2)B \subset C_{r^*}(S) \ominus (r - r^*)B \subset C_r(\mathcal{X}_n)$. This concludes the proof of the lemma by taking $\rho = \epsilon/2$. □

Proposition 8.5. *Let $S \subset \mathbb{R}^d$ be a compact, nonconvex and nonempty set verifying (R^r_λ) and $\mathcal{X}_n$ a uniform and i.i.d sample on S. Let r_0 be the parameter defined in (2.1) and $\{\alpha_n\} \subset (0, 1)$ a sequence converging to zero such that $\log(\alpha_n)/n \to 0$. Then, for any $\epsilon > 0$,*

$$\mathbb{P}\left(\hat{r}_0 \leq r_0 + \epsilon,\ \textit{eventually}\right) = 1.$$

Proof. Given $\epsilon > 0$ let be $r = r_0 + \epsilon$. According to Lemma 8.4, there exists $x \in \mathbb{R}^d$ and $\rho > 0$ such that $B_\rho(x) \cap S = \emptyset$ and

$$\mathbb{P}\left(B_\rho(x) \subset C_r(\mathcal{X}_n), \text{ eventually}\right) = 1.$$

Since, with probability one, $\mathcal{X}_n \subset S$ we have $B_\rho(x) \cap \mathcal{X}_n = \emptyset$. Hence, if $B_\rho(x) \subset C_r(\mathcal{X}_n)$, we have $\hat{V}_{n,r} \geq \mu(B_\rho(x)) = c_\rho > 0$. Similarly, $\hat{V}_{n,r'} \geq \hat{V}_{n,r} \geq c_\rho$ for all $r' \geq r$. On the other hand, since $-u_{\alpha_n}/\log(\alpha_n) = \log(-\log(1-\alpha_n))/\log(\alpha_n) \to 1$, we have, with probability one,

$$\sup_{r'} \hat{c}_{n,\alpha_n,r'} \leq \mu(H(S))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta) \cdot n^{-1}$$

and

$$\mu(H(S))(u_{\alpha_n} + \log n + (d-1)\log\log n + \log\beta) \cdot n^{-1} \to 0$$

where $H(S)$ denotes the convex hull of S. This means that, with probability one, there is n_0 such that if $n \geq n_0$ we have $\sup_{r'} \hat{c}_{n,\alpha_n,r'} < c_\rho$. Therefore, if $B_\rho(x) \subset C_r(\mathcal{X}_n)$, we get $\hat{r}_0 \leq r$. This last statement follows from $\hat{V}_{n,r'} > \hat{c}_{n,\alpha_n,r'}$ for all $r' \geq r$ and the definition of $\hat{r}_0$, see (3.4). □

Theorem 4.1 is, then, a straightforward consequence of Propositions 8.2 and 8.5. □

Proof of Theorem 4.3. For the uniform distribution on S, Theorem 3 of Rodríguez-Casal [24] ensures that, under $(R^{\widetilde{r}}_{\widetilde{r}})$, then $\mathbb{P}(\mathcal{E}_n) \to 1$, where

$$\mathcal{E}_n = \left\{ d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leq A \left(\frac{\log n}{n}\right)^{2/(d+1)} \right\},$$

and A is some constant. Under the hypothesis of Theorem 4.3 this holds for any $\widetilde{r} \leq \min\{r, \lambda\}$. Fix one $\widetilde{r} \leq \min\{r, \lambda\}$ such that $\widetilde{r} < \nu r_0$ and define $\mathcal{R}_n = \{\widetilde{r} \leq r_n \leq r_0\}$. Since, by Theorem 4.1, $r_n = \nu\hat{r}_0$ converges in probability to νr_0 and $\widetilde{r} < \nu r_0 < r_0$, we have that $\mathbb{P}(\mathcal{R}_n) \to 1$. If the events $\mathcal{E}_n$ and $\mathcal{R}_n$ hold (notice that $\mathbb{P}(\mathcal{E}_n \cap \mathcal{R}_n) \to 1$) we have $C_{\widetilde{r}}(\mathcal{X}_n) \subset C_{r_n}(\mathcal{X}_n) \subset S$ and, therefore,

$$d_H(S, C_{r_n}(\mathcal{X}_n)) \leq d_H(S, C_{\widetilde{r}}(\mathcal{X}_n)) \leq A \left(\frac{\log n}{n}\right)^{2/(d+1)}.$$

This completes the proof of the first statement of Theorem 4.3. Similarly, it is possible to prove the result for the other error criteria considered in Theorem 4.3. □

Acknowledgements. We thank the referee's remarks that have led to an improved version of the paper. This work has been supported by Project MTM2013-41383P of the Spanish Ministry of Economy and Competitiveness including support from the European Regional Development Fund and the IAP network StUDyS (Developing crucial Statistical methods for Understanding major complex Dynamic Systems in natural, biomedical and social sciences) of Belgian Science Policy.

References

[1] C. Aaron, A. Cholaquidis and R. Fraiman, On the maximal multivariate spacing extension and convexity tests. Preprint arXiv:1411.2482 (2014).
[2] E. Arias-Castro and A. Rodríguez-Casal, On estimating the perimeter using the alpha-shape. Preprint arXiv:1507.00065 (2015).
[3] A. Baíllo and A. Cuevas, On the estimation of a star-shaped set. *Adv. Appl. Probab.* **33** (2001) 717–726.
[4] A. Baíllo and A. Cuevas, Parametric versus nonparametric tolerance regions in detection problems. *Comput. Stat.* **21** (2006) 527–536.
[5] A. Baíllo, A. Cuevas and A. Justel, Set estimation and nonparametric detection. *Can. J. Stat.* **28** (2000) 765–782.
[6] J.R. Berrendero, A. Cuevas and B. Pateiro-López, A multivariate uniformity test for the case of unknown support. *Stat. Comput.* **22** (2012) 259–271.

[7] J. Chevalier, Estimation du support et du contour du support d'une loi de probabilité. *Ann. Inst. Henri Poincaré, Probab. Stat.* **12** (1976) 339–364.
[8] A. Cholaquidis, A. Cuevas and R. Fraiman, On Poincaré cone property. *Ann. Stat.* **42** (2014) 255–284.
[9] A. Cuevas, On pattern analysis in the non-convex case. *Kybernetes* **19** (1990) 26–33.
[10] A. Cuevas, R. Fraiman and B. Pateiro-López, On statistical properties of sets fulfilling rolling-type conditions. *Adv. Appl. Probab.* **44** (2012) 311–329.
[11] A. Cuevas and A. Rodríguez-Casal, On boundary estimation. *Adv. Appl. Probab.* **36** (2004) 340–354.
[12] L. Devroye and G.L. Wise, Detection of abnormal behavior *via* nonparametric estimation of the support. *SIAM J. Appl. Math.* **38** (1980) 480–488.
[13] L. Dümbgen and G. Walther, Rates of convergence for random approximations of convex sets. *Adv. Appl. Probab.* **28** (1996) 384–393.
[14] H. Edelsbrunner, Alpha shapes – a survey. To apprear in *Tessellations in the Sciences*. Springer (2016).
[15] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli and L. Wasserman, The geometry of nonparametric filament estimation. *J. Amer. Statist. Assoc.* **107** (2012) 788–799.
[16] U. Grenander, *Abstract Inference*. Wiley, New York (1981).
[17] S. Janson, Maximal spacings in several dimensions. *Ann. Probab.* **15** (1987) 274–280.
[18] A.P. Korostelëv and A.B. Tsybakov, *Minimax Theory of Image Reconstruction*. Springer (1993).
[19] D.P. Mandal and C.A. Murthy, Selection of alpha for alpha-hull in $\mathbb{R}^2$. *Pattern Recogn.* **30** (1997) 1759–1767.
[20] B. Pateiro-López and A. Rodríguez-Casal, Generalizing the convex hull of a sample: the R package alphahull. *J. Stat. Softw.* **34** (2010) 1–28.
[21] B. Pateiro-López and A. Rodríguez-Casal, Recovering the shape of a point cloud in the plane. *TEST* **22** (2013) 19–45.
[22] M. Reitzner, Random polytopes and the efron'stein jackknife inequality. *Ann. Probab.* **31** (2003) 2136–2166.
[23] B.D. Ripley and J.P. Rasson, Finding the edge of a poisson forest. *J. Appl. Probab.* **14** (1977) 483–491.
[24] A. Rodríguez-Casal, Set estimation under convexity type assumptions. *Ann. Inst. Henri Poincaré, Probab. Stat.* **43** (2007) 763–774.
[25] R. Schneider, Random approximation of convex sets. *J. Microsc.* **151** (1988) 211–227.
[26] R. Schneider, *Convex Bodies: the Brunn–Minkowski Theory*. Cambridge University Press (1993).
[27] J. Serra, *Image Analysis and Mathematical Morphology*. Academic Press, London (1982).
[28] G. Walther, Granulometric smoothing. *Ann. Stat.* **25** (1997) 2273–2299.
[29] G. Walther, On a generalization of blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. Sci.* **22** (1999) 301–316.