ESAIM: PS 17 (2013) 359–369
DOI: 10.1051/ps/2011160

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

TOWARDS A UNIVERSALLY CONSISTENT ESTIMATOR OF THE MINKOWSKI CONTENT*

Antonio Cuevas[1], Ricardo Fraiman[2] and László Györfi[3]

Abstract. We deal with a subject in the interplay between nonparametric statistics and geometric measure theory. The measure $L_0(G)$ of the boundary of a set $G \subset \mathbb{R}^d$ (with $d \geq 2$) can be formally defined, *via* a simple limit, by the so-called Minkowski content. We study the estimation of $L_0(G)$ from a sample of random points inside and outside G. The sample design assumes that, for each sample point, we know (without error) whether or not that point belongs to G. Under this design we suggest a simple nonparametric estimator and investigate its consistency properties. The main emphasis in this paper is on generality. So we are especially concerned with proving the consistency of our estimator under minimal assumptions on the set G. In particular, we establish a mild shape condition on G under which the proposed estimator is consistent in L_2. Roughly speaking, such condition establishes that the set of "very spiky" points at the boundary of G must be "small". This is formalized in terms of the Minkowski content of such set. Several examples are discussed.

Mathematics Subject Classification. 62G05, 62G99.

Received March 28, 2011.

1. Introduction

The problem under study. The sampling model. The statistical estimation of the boundary measure of a subset G of $[0,1]^d$, $d \geq 2$, has been usually tackled from the point of view of stereology (see, *e.g.*, Baddeley and Vedel-Jensen [3]), using information provided by lower dimensional sections drawn on this set.

We consider here a different approach, using just a random sample of points. Our sample data will be given by $(Z_i, \mathbb{I}_G(Z_i))$, $i = 1, \ldots, n$, where the Z_i are i.i.d. random vectors uniformly distributed on $[0,1]^d$ and $\mathbb{I}_G$ denotes the indicator function of G.

We are interested in estimating from the sample data the boundary measure $L_0(G)$ of G, as given by the Minkowski content (to be defined below).

Keywords and phrases. Minkowski content, nonparametric set estimation, boundary estimation.

* *The authors are grateful to Elena Villa for some interesting remarks and for pointing out some inaccuracies. The useful and constructive reports from two referees and an associate editor are also gratefully acknowledged. The work of A. Cuevas and R. Fraiman has been partially supported by Spanish grants MTM2010-17366 and CCG10-UAM/ESP-5494 (A. Cuevas). The work of L. Györfi was supported in part by the Computer and Automation Research Institute of the Hungarian Academy of Sciences and by the PASCAL2 Network of Excellence under EC Grant No. 216886.*

[1] Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain. antonio.cuevas@uam.es

[2] Departamento de Matemáticas y Ciencias, Universidad de San Andrés, Argentina and CMAT, Universidad de la República, Uruguay

[3] Department of Computer Science and Information Theory, Budapest University of Technology and Economics, Hungary.

Article published by EDP Sciences

© EDP Sciences, SMAI 2013

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

We will use some typical tools of nonparametrics (as the smoothing parameters) but some key concepts, as well as the main target of our study, come from the field of geometric measure theory.

The problem of estimating the boundary measure using nonparametric methods has been previously considered by Cuevas *et al.* [5], who obtain consistency results and convergence rates (of order $n^{-1/(2d)}$) for the estimation of $L_0(G)$ under some conditions on G which exclude the existence of sharp peaks or inlands or impose some kind of smoothness on the boundary. Faster rates (of order $n^{-1/(d+1)}$ up to a logarithmic factor) have been obtained by Pateiro−López and Rodríguez−Casal [14] under the additional assumption of r-convexity (which is used in the definition of the corresponding estimator). Armendáriz *et al.* [2] consider also the problem under a slightly different sampling design (based on two independent samples with different size orders from G and $[0,1]^d \setminus G$), obtaining asymptotic normality for a simple estimator under shape conditions similar to those in Cuevas *et al.* [5]. When the surface measure is defined in terms of a surface integral (rather than by the Minkowski content), Jiménez and Yukich [10] provide a consistent estimator for the surface measure based on the use of Delaunay triangulations.

In this work the emphasis is mainly on generality. We propose a simple, computationally feasible, nonparametric estimator of the boundary Minkowski content $L_0(G)$ which is "nearly universal" in the sense that it provides a consistent estimation of $L_0(G)$ for a broad class of sets G, defined by very general conditions which apply even in the case of complicated and "unsmooth" G's. The possible existence of a fully universally consistent estimator remains as an open problem.

Some notation. In what follows μ will denote the uniform distribution on $[0,1]^d$ (*i.e.*, the Lebesgue measure μ_L restricted to the unit square), $B(x,\epsilon)$ will denote the closed ball of center x and radius ϵ. The Euclidean norm in $\mathbb{R}^d$ will be denoted by $\|.\|$. $B(A,\epsilon)$ will stand for the ϵ-parallel set of a set $A \subset \mathbb{R}^d$, $B(A,\epsilon) = \{x : D(x,A) \leq \epsilon\}$, where $D(x,A) = \inf_{y\in A} \|x-y\|$.

For any Borelian set $A \subset \mathbb{R}^d$, $\mathrm{diam}(A)$ will represent the diameter of A, $\mathrm{diam}(A) = \sup\{\|x-y\| : \ x,y \in A\}$.

Throughout this paper G^c will denote the complementary set of G relative to $[0,1]^d$, that is $G^c = [0,1]^d \setminus G$ and ∂G will represent the topological boundary of G. Unless otherwise stated, $\to$ will denote convergence as $n \to \infty$.

General assumptions on G. Denote by ∂G_μ the μ-boundary of G,

$$\partial G_\mu = \{x : \text{for all } \epsilon > 0,\ \mu(B(x,\epsilon) \cap G) > 0 \text{ and } \mu(B(x,\epsilon) \cap G^c) > 0\}. \tag{1.1}$$

Obviously, $\partial G_\mu \subset \partial G$ In what follows we will assume that $\partial G_\mu = \partial G$.

In intuitive terms (*e.g.*, in the case $d = 2$) this entails that ∂G does not include "superfluous" subsets, such as line segments of null measure, irradiating from the central core of G. These line segments would contribute to the topological boundary measure but they are "invisible" with our statistical model since they are made of points which do not have simultaneously a substantial part of G and G^c around them. It can be easily seen that, if $\partial G_\mu = \partial G$ is not fulfilled, we cannot have any consistent sequence of estimators T_n for the boundary measure of a closed $G \subsetneq [0,1]^d$ based on a sample (Z_i, Y_i), $i = 1,\ldots,n$ with $Y_i = \mathbb{I}_G(Z_i)$ and Z_i i.i.d. uniform on $[0,1]^d$.

We will assume throughout that $\mu(G) > 0$, with $\bar{G} = G \cup \partial G \subset (0,1)^d$ and that the boundary measure of G, $L_0(G)$ (to be defined below), is finite.

2. Some geometric concepts. Definition of the estimator

This section is devoted to some geometric and statistical preliminaries. First, we briefly recall some well-known concepts in geometric measure theory. Then in Section 2.2 we introduce and discuss the shape assumption imposed on G in order to get a consistent estimation of its boundary measure. Finally, in Section 2.3 we define and motivate the estimator whose properties are analyzed in the rest of the paper.

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

2.1. Some basic ideas on lower-dimensional measures

The measurement of boundary surfaces is a particular case of an especially delicate and elusive problem: the study of lower dimensional measures in the Euclidean space. There is no unique natural way of deciding what surfaces can be measured in the d-dimensional space and how to assign them the corresponding lower-dimensional measure. Several, not equivalent, proposals have been made. Of course, all of them coincide when applied to "easy" sets (for example, convex polyhedra) but these are just a tiny part of the immense family of really complicated sets which can appear in different mathematical problems, in particular in the fractal sets theory. The area of mathematics dealing with these subjects is called geometric measure theory. Its beginning as a structured mathematical discipline goes back at least fifty years. The books by Federer [9] and Mattila [12] are two classical references. We briefly review here some relevant concepts of geometric measure theory which will be used or mentioned below.

The Hausdorff measure of a surface. The best known concept of k-dimensional measure on $\mathbb{R}^d$, for $k \leq d$ is the so-called Hausdorff measure defined by

$$\mathcal{H}^k(A) = \lim_{\delta \downarrow 0} \inf \left\{ \sum_i \operatorname{diam}(E_i)^k : \ E_i \text{ Borelian}, \ A \subset \bigcup E_i, \ \operatorname{diam}(E_i) \leq \delta \right\}. \tag{2.1}$$

We are specially interested in the case $k = d-1$, $A = \partial G$, where $\mathcal{H}^{d-1}(\partial G)$ provides a standard notion for the measure of the boundary of $G \subset \mathbb{R}^d$.

It can be seen (*e.g.*, Mattila [12], Chap. 4) that $\mathcal{H}^k$ is a true measure which for $k = 0$ coincides with the "counting measure" (that is $\mathcal{H}^0(A)$ is the cardinality of A) and for $k = d$ coincides, when multiplied by the normalizing factor $2^{-d}\mu_L(B(0,1))$, with the Lebesgue measure on $\mathbb{R}^d$.

The Minkowski content. While Hausdorff measure is the most popular notion of low-dimensional measure in the Euclidean space, we will rather use here another simpler concept usually called the *Minkowski content*. If we focus in our target of measuring the boundary of G we may define the ($(d-1)$-dimensional) boundary content of G by

$$L_0(G) = \lim_{\epsilon \to 0} \frac{\mu(B(\partial G, \epsilon))}{2\epsilon}, \tag{2.2}$$

provided that this limit exists. The recent papers by Ambrosio *et al.* [1] and Villa [15] provide deep studies of the Minkowski content and some other closely related notions.

Rectifiability. It is intuitively clear that $L_0(G)$ provides a definition, alternative to $\mathcal{H}^{d-1}(\partial G)$, for the "surface area" of G. In general, the Minkowski content is not a true measure as it fails to be σ-subadditive. However, it coincides with $\mathcal{H}^{d-1}(\partial G)$ if G is "regular enough". For example, this is the case if G is compact and ∂G is $(d-1)$-*rectifiable*, (∂G is said to be k-rectifiable if there is a compact set $C \subset \mathbb{R}^k$ and a Lipschitz function $f : \mathbb{R}^k \to \mathbb{R}^d$ such that $\partial G = f(C)$; see, *e.g.*, Federer [9] p. 251, 275).

This concept is closely related with the more usual notion of rectifiable curve, defined by a continuous function $\gamma : I \to \mathbb{R}^d$, where $I = [a,b]$ is a compact interval of $\mathbb{R}$. The curve γ is said to be *rectifiable* if

$$\operatorname{Length}(\gamma) = \sup \left\{ \sum_{i=1}^{n} \|\gamma(t_i) - \gamma(t_{i-1})\| : \ n \in \mathbb{N}, \ a = t_0 < \ldots < t_n = b \right\} < \infty.$$

It is clear that if γ is 1-rectifiable, then it is rectifiable in the above sense. A curve $\gamma : [a,b] \to \mathbb{R}^d$ is called *simple* if the function γ is injective when restricted to the open interval (a,b) (that is we allow $\gamma(a) = \gamma(b)$). It can be seen that the value of the one-dimensional Hausdorff measure of a simple rectifiable curve coincides with its length.

It should be clear, from the above definitions and the discussion below, that there are some reasons for using the Minkowski content (instead of other mathematical notions) for defining the surface area. The expression (2.2)

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

of the Minkowski content suggests, in a quite natural way, several estimators for $L_0(G)$. Also, the Minkowski content (2.2) is particularly suitable for statistical purposes as it is defined just in terms of measure, with no resort to analytic concepts such as rectifiability or differentiability notions. This allows for an easier handling of statistical properties such as convergence rates, see Cuevas *et al.* [5], or asymptotic normality, see Armendáriz *et al.* [2].

2.2. Our main assumption and its geometric interpretation

Let us recall that our target is to provide an estimator for the Minkowski content of a set G which is consistent under very general conditions.

In Cuevas *et al.* [5] the consistency of a plug-in type estimator, see (2.6) below, is established under the following *double standardness assumption*: the set G is said to be *doubly standard* if there exist two constants $\delta_0 > 0$ and $\epsilon_0 > 0$ such that, for all $\epsilon \in (0, \epsilon_0)$ and $x \in \partial G$,

$$\mu(B(x,\epsilon) \cap G^c) \geq \delta_0 \epsilon^d \text{ and } \mu(B(x,\epsilon) \cap G) \geq \delta_0 \epsilon^d. \tag{2.3}$$

In fact the standardness assumption appears under similar, slightly different forms in the set estimation literature. See Cuevas and Fraiman [4], for further details and references. In intuitive terms, this condition rules out the existence of too sharp peaks or inlands in the boundary of G.

A new type of standardness: our main assumption on the shape of G.

For $x \in \partial G$, let us introduce the notations

$$r_{G^c}(x) = \inf_{0<\epsilon\leq 1} \frac{\mu(B(x,\epsilon) \cap G^c)}{\epsilon^d} \text{ and } r_G(x) = \inf_{0<\epsilon\leq 1} \frac{\mu(B(x,\epsilon) \cap G)}{\epsilon^d}.$$

In order to have measurable functions r_{G^c} and r_G, we take the infimum for rational $0 < \epsilon \leq 1$.

These functions are closely related with the notion of *lower d-density* (see, *e.g.*, Mattila [12], p. 89). Throughout the discussion below it will be clear that the bound 1 for ϵ in the definition of r_G and r_{G^c} is not relevant and can be replaced with any other positive value ϵ_0.

These definitions imply that for each $0 < \epsilon \leq 1$,

$$\mu(B(x,\epsilon) \cap G^c) \geq r_{G^c}(x)\epsilon^d \text{ and } \mu(B(x,\epsilon) \cap G) \geq r_G(x)\epsilon^d.$$

Let us now define for each $z \in \mathbb{R}^d$ the projection $P(z)$ of z on ∂G, that is $P(z) = \operatorname{argmin}_{g\in\partial G}\|z - g\|$ is the point which realizes the minimum distance from z to ∂G, $\|z - P(z)\| = D(z, \partial G)$. It is known that $P(z)$ is uniquely defined almost everywhere. More precisely, let us define $\mathbf{P}(z) = \{g \in \partial G : \|g - z\| = D(z, \partial G)\}$ and $C = \{z : \mathbf{P}(z) \text{ has a unique point}\}$. As shown in Erdős [7] Remark 3, $\mu(C^c) = 0$. Moreover, it can be proved that the projection function P is continuous when restricted to C; see Federer [8], Theorem 4.8. In particular, P is measurable as it coincides with a continuous function except for a (Lebesgue) null set.

Then we are ready to establish our main assumption. We will say that G is *(doubly) standard almost everywhere (a.e.)* if for any numerical sequence $\gamma_n \to 0$.

$$\mu(\{z : r_{G^c}(P(z)) \leq \gamma_n\} \cap B(\partial G, \epsilon_n)) = o(\epsilon_n), \text{ and } \mu(\{z : r_G(P(z)) \leq \gamma_n\} \cap B(\partial G, \epsilon_n)) = o(\epsilon_n), \tag{2.4}$$

as $\epsilon_n \to 0$, where the notation $R_n = o(\epsilon_n)$ stands for $\limsup_{n\to\infty} R_n/\epsilon_n = 0$. Without loss of generality we may assume $\gamma_n \downarrow 0$ in (2.4).

It can be seen that assumption (2.4) is much more general than (2.3) since, under (2.3), the sets appearing in condition (2.4) are eventually empty.

Interpretation in terms of the Minkowski content. The rest of this section is devoted to analyze and explain the real extent of our assumption (2.4).

Let us call

$$T_0 = T_0^- \cup T_0^+ := \{x \in \partial G : r_{G^c}(x) = 0\} \cup \{x \in \partial G : r_G(x) = 0\}$$

the *set of non-standard points* of G.

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

The following result provides an interpretation of condition (2.4) in terms of the Minkowski measure of T_0 thus accounting for the name "double standardness almost everywhere".

Proposition 2.1. *Let γ_n be a numerical sequence $\gamma_n \downarrow 0$. Denote $T_n^- = \{z \in \partial G : r_{G^c}(x) \leq \gamma_n\}$, $T_n^+ = \{z \in \partial G : r_G(x) \leq \gamma_n\}$, $T_n = T_n^- \cup T_n^+$. Then,*

(a) $\mu(T_n) = o(\epsilon_n)$ is a sufficient condition for the double standardness a.e. (2.4).
(b) If $L_0(T_0) = 0$ then G also fulfills the double standardness a.e. condition (2.4).

Proof.

(a) Note that $T_n^- \downarrow \{x \in \partial G : \ r_{G^c}(x) = 0\} = T_0^-$. Moreover,

$$\{z : r_{G^c}(P(z)) \leq \gamma_n\} \cap B(\partial G, \epsilon_n) = \bigcup_{x \in \partial G,\, r_{G^c}(x) \leq \gamma_n} P^{-1}(x) \cap B(\partial G, \epsilon_n)$$
$$= \bigcup_{x \in T_n^-} P^{-1}(x) \cap B(\partial G, \epsilon_n) \subset B(T_n^-, \epsilon_n).$$

A similar inclusion holds for T_n^+. Now, since $T_n = T_n^+ \cup T_n^-$, the result follows.

(b) According to the previous result it suffices to prove that $L_0(T_0) = 0$ implies $\mu(T_n) = o(\epsilon_n)$, as $\epsilon_n \to 0$. Indeed, for all $m, n \in \mathbb{N}$ define $c_{m,n} = \mu(B(T_m^-, \epsilon_n))/\epsilon_n$. We have

$$\limsup_n \frac{\mu(B(T_n^-, \epsilon_n))}{\epsilon_n} \leq \lim_m \limsup_n c_{m,n} = \inf_m \limsup_n c_{m,n}.$$

Denote $c = \inf_m \limsup_n c_{m,n}$. Let us prove $c = 0$. By contradiction, assume that $c > 0$. In this case we would have for all m, $\limsup_n c_{m,n} \geq c$ so that we could choose a subsequence $\{n_k\}$ such that $\mu(B(T_m^-, \epsilon_{n_k})) \geq c\epsilon_{n_k}/2$ (note that c is finite since $L_0(G) < \infty$). Therefore,

$$\inf_m \mu(B(T_m^-, \epsilon_{n_k})) \geq c\epsilon_{n_k}/2 > 0.$$

Then, since $B(T_m^-, \epsilon_{n_k}) \downarrow B(T_0^-, \epsilon_{n_k})$, as $m \to \infty$, the above result entails

$$\inf_m \mu(B(T_m^-, \epsilon_{n_k})) = \mu(B(T_0^-, \epsilon_{n_k})) \geq c\epsilon_{n_k}/2,$$

which leads to a contradiction with the assumption $L_0(T_0) = 0$; indeed note that, since $T_0 \subset \partial G$, $B(T_0, \epsilon) = B(\partial T_0, \epsilon)$ so that $L_0(T_0)$ coincides in fact with the $(d-1)$-dimensional Minkowski content of T_0. The reasoning for T_0^+ is identical. Then we conclude that $\mu(T_n) = o(\epsilon_n)$ as this property holds for both $\mu(T_n^-)$ and $\mu(T_n^+)$. □

As a consequence of Proposition 2.1, we could say that G is (doubly) standard (a.e.) when the set of very-spiky non-standard points T_0 in the boundary of G is "small", *i.e.* $(d-1)$-Minkowski-null (see, *e.g.*, Mattila [12], p. 79). Thus, as we will discuss below, the class of sets fulfilling (2.4) is extremely general.

2.3. The estimator

Before defining our estimate, let us recall (for comparison purposes) another estimator which has been previously considered in Cuevas *et al.* [5].

A plug-in type estimator. Given $z \in [0,1]^d$ and $\epsilon \geq 0$, denote

$$G_{n,z}(\epsilon) = \sum_{i=1}^n \mathbb{I}_{\{Z_i \in G, \|Z_i - z\| \leq \epsilon\}}, \quad R_{n,z}(\epsilon) = \sum_{i=1}^n \mathbb{I}_{\{Z_i \notin G, \|Z_i - z\| \leq \epsilon\}}. \tag{2.5}$$

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

Now we may define

$$L_{0n} = \frac{\mu(B_n)}{2\epsilon_n}, \tag{2.6}$$

where $\{\epsilon_n\}$ is a deterministic sequence with $\epsilon_n \downarrow 0$, and B_n is an estimator of the "dilated boundary" $B(\partial G, \epsilon)$ defined by

$$B_n = \left\{ z \in [0,1]^d : R_{n,z}(\epsilon_n) \geq 1 \text{ and } G_{n,z}(\epsilon_n) \geq 1 \right\}. \tag{2.7}$$

Of course, the exact evaluation of $\mu(B_n)$ could be difficult in practice but this quantity can be approximated with an arbitrary precision by Monte Carlo sampling. That is, we could draw a (large) artificial sample $X_1 \ldots, X_N$ from the uniform distribution on $[0,1]^d$ and estimate $\mu(B_n)$ as the proportion of X_i's observations belonging to B_n.

Another estimator based on a smoothed empirical average. In the present work we will consider a slightly different approach by estimating (2.2) through a sort of smoothed empirical average which does not require any Monte Carlo approximation. To motivate our estimator let us start with a pseudo-estimate

$$\begin{aligned} L_{1n} &= \sum_{i=1}^{n} \frac{\mathbb{I}_{\{Z_i \in B(\partial G, \epsilon_n)\}}}{2\epsilon_n n} \\ &= \sum_{i=1}^{n} \frac{\mathbb{I}_{\{Z_i \in B(\partial G, \epsilon_n), Z_i \in G\}}}{2\epsilon_n n} + \sum_{i=1}^{n} \frac{\mathbb{I}_{\{Z_i \in B(\partial G, \epsilon_n), Z_i \in G^c\}}}{2\epsilon_n n} \end{aligned} \tag{2.8}$$

where again $\{\epsilon_n\}$ is a sequence of smoothing parameters with $\epsilon_n \downarrow 0$. Obviously, L_{1n} is not a true estimator since it is based on the unrealistic assumption that ∂G is known. Nevertheless, the expression (2.8) suggests the following simple empirical approximation for L_{1n} which can be calculated from the sample under the assumptions of our sampling model,

$$L_n = \frac{1}{2\epsilon_n n} \sum_{i=1}^{n} \left(\mathbb{I}_{\{\|Z_i - Z_{n,G^c}(Z_i)\| \leq \epsilon_n, Z_i \in G\}} + \mathbb{I}_{\{\|Z_i - Z_{n,G}(Z_i)\| \leq \epsilon_n, Z_i \in G^c\}} \right), \tag{2.9}$$

where for $z \in G$, $Z_{n,G^c}(z)$ denotes the first nearest neighbor of z among the points $Z_1, \ldots, Z_n$ which are in G^c, and, for $z \in G^c$, $Z_{n,G}(z)$ is the first nearest neighbor of z among the points $Z_1, \ldots, Z_n$ which are in G. Note that the definition (2.9) of L_n requires to have sample observations in both G and G^c, which happens eventually with probability one. If there is no data on G or in G^c, L_n can be (arbitrarily) defined to be 0.

This paper is devoted to the study of the consistency properties of L_n.

3. Consistency

We provide here our main result, concerning L_2-consistency, for the estimator L_n defined in (2.9). This result holds under the very mild shape assumption (2.4) introduced in the previous section.

In order to compare L_n with the pseudo-estimate L_{1n} let us note that

$$\mathbb{E}(L_{1n}) = \frac{\mu(B(\partial G, \epsilon_n))}{2\epsilon_n} \to L_0(G), \text{ as } n \to \infty. \tag{3.1}$$

Note also that if $\|Z_i - Z_{n,G^c}(Z_i)\| \leq \epsilon_n$ and $Z_i \in G$ then $\mu(B(Z_i, \epsilon_n) \cap G^c) > 0$, therefore $Z_i \in B(\partial G, \epsilon_n)$. Similarly $\|Z_i - Z_{n,G}(Z_i)\| \leq \epsilon_n$ and $Z_i \in G^c$ imply $Z_i \in B(\partial G, \epsilon_n)$. Thus, L_n is a lower approximation of L_{1n},

$$L_n \leq L_{1n}.$$

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

Theorem 3.1. *Let us consider the estimator L_n of $L_0(G)$ defined in (2.9). Assume that*

$$\epsilon_n \downarrow 0 \text{ and } n\epsilon_n^d \to \infty, \text{ as } n \to \infty. \tag{3.2}$$

Assume further that G fulfils the (double) a.e. standardness condition (2.4). Then

$$\lim_{n\to\infty} \mathbb{E}\left((L_n - L_0(G))^2\right) = 0.$$

Proof. Since $L_n \le L_{1n}$,

$$\begin{aligned}\mathbb{E}\left((L_n - L_0(G))^2\right) &= \mathbb{E}\left((L_n - \mathbb{E}(L_n))^2\right) + \left(\mathbb{E}(L_n) - L_0(G)\right)^2 \\ &\le \mathbb{E}(L_{1n}^2) - \mathbb{E}^2(L_n) + \left(\mathbb{E}(L_n) - L_0(G)\right)^2. \end{aligned} \tag{3.3}$$

As $\mathbb{E}(L_{1n}) = \mu\left(B(\partial G, \epsilon)\right)/2\epsilon \to L_0(G)$, the proof will be complete if we show that $\mathrm{Var}(L_{1n}) \to 0$ and $\mathbb{E}(L_{1n}) - \mathbb{E}(L_n) \to 0$.

Concerning $\mathrm{Var}(L_{1n})$, introduce the notation $L(\epsilon) = \mu\left(B(\partial G, \epsilon)\right)/2\epsilon$, then

$$\mathrm{Var}(L_{1n}) = \frac{\mathrm{Var}(\mathbb{I}_{\{Z_1 \in B(\partial G, \epsilon_n)\}})}{4\epsilon_n^2 n} \le \frac{\mu\left(B(\partial G, \epsilon_n)\right)}{4\epsilon_n^2 n} = \frac{L(\epsilon_n)}{2\epsilon_n n} \to 0,$$

since $L(\epsilon_n) \to L_0(G)$ and $n\epsilon_n \to \infty$.

One can upper bound $(\mathbb{E}(L_{1n}) - \mathbb{E}(L_n))$ as follows

$$\begin{aligned} 0 \le 2\epsilon_n \left(\mathbb{E}(L_{1n}) - \mathbb{E}(L_n)\right) &= \mathbb{P}\{Z_1 \in G, Z_1 \in B(\partial G, \epsilon_n), \|Z_1 - Z_{n,G^c}(Z_1)\| > \epsilon_n\} \\ &\quad + \mathbb{P}\{Z_1 \in G^c, Z_1 \in B(\partial G, \epsilon_n), \|Z_1 - Z_{n,G}(Z_1)\| > \epsilon_n\}, \end{aligned} \tag{3.4}$$

therefore

$$\begin{aligned} 0 \le \mathbb{E}(L_{1n}) - \mathbb{E}(L_n) = &\frac{1}{2\epsilon_n} \int_{G \cap B(\partial G, \epsilon_n)} \left(1 - \mu(B(z, \epsilon_n) \cap G^c)\right)^{n-1} \mu(\mathrm{d}z) \\ &+ \frac{1}{2\epsilon_n} \int_{G^c \cap B(\partial G, \epsilon_n)} \left(1 - \mu(B(z, \epsilon_n) \cap G)\right)^{n-1} \mu(\mathrm{d}z). \end{aligned}$$

Now, to handle the first term in the right-hand side we use the inequality $1 - x \le \mathrm{e}^{-x}$, for $x \in [0, 1]$, together with assumption (3.2). Thus, given any $C > 0$, we have that for n large enough,

$$\begin{aligned} &\frac{1}{2\epsilon_n} \int_{G \cap B(\partial G, \epsilon_n)} \left(1 - \mu(B(z, \epsilon_n) \cap G^c)\right)^{n-1} \mu(\mathrm{d}z) \le \frac{1}{2\epsilon_n} \int_{B(\partial G, \epsilon_n)} \mathrm{e}^{-(n-1)\mu(B(z,\epsilon_n) \cap G^c)} \mu(\mathrm{d}z) \\ &\le \frac{\mu\left(B(\partial G, \epsilon_n)\right)}{2\epsilon_n} \mathrm{e}^{-C} + \frac{\mu\left(\{z : (n-1)\mu(B(z, \epsilon_n) \cap G^c) \le C\} \cap B(\partial G, \epsilon_n)\right)}{2\epsilon_n}. \end{aligned}$$

Obviously, $z \in B(\partial G, \epsilon_n)$ implies that $B\left(P(z), \epsilon_n - \|z - P(z)\|\right) \subset B(z, \epsilon_n)$. Therefore

$$\begin{aligned} &\frac{1}{2\epsilon_n} \int_{G \cap B(\partial G, \epsilon_n)} \left(1 - \mu(B(z, \epsilon_n) \cap G^c)\right)^{n-1} \mu(\mathrm{d}z) \\ &\le L(\epsilon_n)\mathrm{e}^{-C} + \frac{\mu\left(\{z : (n-1)\mu(B(P(z), \epsilon_n - \|z - P(z)\|) \cap G^c) \le C\} \cap B(\partial G, \epsilon_n)\right)}{2\epsilon_n}. \end{aligned}$$

For an arbitrary $\alpha \in (0, 1)$, we have that

$$\begin{aligned} &\{z : (n-1)\mu\left(B(P(z), \epsilon_n - \|z - P(z)\|) \cap G^c\right) \le C\} \cap B(\partial G, \epsilon_n) \\ &\subset \{z : (n-1)\mu\left(B(P(z), \epsilon_n - \|z - P(z)\|) \cap G^c\right) \le C\} \cap B(\partial G, \alpha\epsilon_n) \cup \left(B(\partial G, \epsilon_n) \setminus B\left(\partial G, \alpha\epsilon_n\right)\right), \end{aligned}$$

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

so

$$\begin{aligned}
&\frac{1}{2\epsilon_n}\int_{G\cap B(\partial G,\epsilon_n)}(1-\mu(B(z,\epsilon_n)\cap G^c))^{n-1}\,\mu(\mathrm{d}z)\\
&\le L(\epsilon_n)\mathrm{e}^{-C}+\frac{\mu(\{z:(n-1)\mu(B(P(z),\epsilon_n-\alpha\epsilon_n)\cap G^c\le C\}\cap B(\partial G,\alpha\epsilon_n))}{2\epsilon_n}\\
&\quad +L(\epsilon_n)-\alpha L(\alpha\epsilon_n)\\
&\le L(\epsilon_n)\mathrm{e}^{-C}+\frac{\mu(\{z:(n-1)r_{G^c}(P(z))(1-\alpha)^d\epsilon_n^d\le C\}\cap B(\partial G,\alpha\epsilon_n))}{2\epsilon_n}\\
&\quad +L(\epsilon_n)-\alpha L(\alpha\epsilon_n)\\
&= L(\epsilon_n)\mathrm{e}^{-C}+\frac{\mu(\{z:r_{G^c}(P(z))\le C/((n-1)(1-\alpha)^d\epsilon_n^d)\}\cap B(\partial G,\alpha\epsilon_n))}{2\epsilon_n}\\
&\quad +L(\epsilon_n)-\alpha L(\alpha\epsilon_n)
\end{aligned}$$

This inequality together with (3.2) and (2.4) imply that

$$\limsup_{n\to 0}\frac{\mathbb{P}\{Z_1\in G,Z_1\in B(\partial G,\epsilon_n),\|Z_1-Z_{n,G^c}(Z_1)\|>\epsilon_n\}}{2\epsilon_n}\le L_0(G)\mathrm{e}^{-C}+(1-\alpha)L_0(G).$$

Since $\alpha<1$ and $C>0$ are arbitrary, we conclude

$$\lim_{n\to 0}\frac{\mathbb{P}\{Z_1\in G,Z_1\in B(\partial G,\epsilon_n),\|Z_1-Z_{n,G^c}(Z_1)\|>\epsilon_n\}}{2\epsilon_n}=0.$$

The term $\mathbb{P}\{Z_1\in G^c,Z_1\in B(\partial G,\epsilon_n),\|Z_1-Z_{n,G}(Z_1)\|>\epsilon_n\}$ in (3.4) can be handled in a similar way. □

3.1. Some examples

The purpose of this subsection is to gain some insight on the true meaning of the crucial condition (2.4) and, secondarily, of the assumption $L_0(G)<\infty$. As indicated above, hypothesis (2.4) is in fact a generalization of the more intuitive double standardness assumption (2.3) imposed, among others, by Cuevas *et al.* [5]. From the results of the above subsection we see that, roughly speaking, (2.4) relaxes the assumption (2.3) by allowing the presence of sharp inlands or peaks in a "small portion" of the boundary.

We next provide three examples, with an increasing level of complexity, of sets G where (2.4) is fulfilled and $L_0(G)$ is well-defined and finite so, according to Theorem 3.1, $L_0(G)$ can be consistently estimated. These examples show that our consistency result applies in fact to some complicated sets where the double standardness (2.4) does not hold. In order to see the intuitive meaning of $L_0(G)<\infty$ we also finally give a "negative example" where $L_0(G)=\infty$. When required, the general assumption $\bar{G}\subset(0,1)^d$ should be replaced with $\bar{G}\subset(a,b)^d$ for another suitable "frame box" $[a,b]^d$.

The pagoda. The standardness condition (2.3) is fulfilled in the "simple house" $G_1=\{(x,y):\ x\in[0.2,0.8],\ y\le f(x)=1-|x-1/2|\}$ on the left of Figure 1, but it fails in the example $G_2=\{(x,y):\ x\in[0.2,0.8],\ y\le f(x)=\exp(-|x-1/2|^{1/3})$ on the right, where the "linear peak" in the roof of the house is replaced by an exponential, pagoda-type peak. It is clear, from Proposition 2.1 (b), that our assumption (2.4) holds for the pagoda G_2, since the only element in T_0 is the upper vertex. Also $L_0(G_2)$ is well-defined and finite. Therefore, $L_0(G_2)$ can be estimated consistently with our estimator (2.9).

The many-many-boxes. Let us consider the set $B=\bigcup_{n=1}^{\infty}B_n$, where $\{B_n\}$ is a sequence of disjoint closed square boxes, strictly contained in the unit square, chosen in such a way that $\sum_n L_0(B_n)<\infty$ We will also assume that the sequence of boxes accumulates towards a unique point $v\in[0,1]^2$. Let us now define $G=\{v\}\cup B$.

It is clear that G does not fulfill the double standardness assumption (2.3) imposed in Cuevas *et al.* [5] to estimate consistently the boundary length of G: it suffices to study this condition at the point v (note that

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

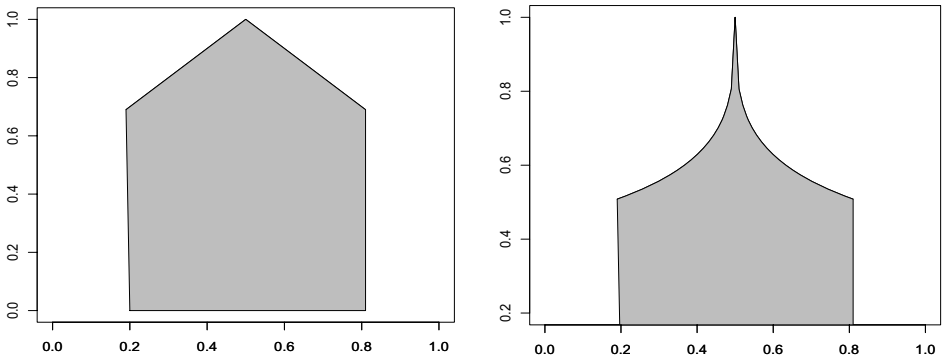


FIGURE 1. The house fulfills standardness assumption (2.3), the pagoda meets the broader condition (2.4).

$\mu(B_n) = o(1/n^2)$ since $L_0(B_n) = o(1/n)$). However, $L_0(G)$ can be estimated consistently using our estimator. This follows again from Proposition 2.1 (b) and Theorem 3.1.

The Cantor hypograph. Let X be a random variable taking values in $[0, 1]$ such that in its binary expansion the components are i.i.d. and the probability of 1 is denoted by p. If $p = 1/2$ then X has (uniform) density. For all other values of p, X has singular Cantor-type continuous distribution, and so its distribution function $F(x)$ is continuous but not absolutely continuous. In intuitive terms this means that F manages to "climb" from 0 to 1 increasing only in a Lebesgue null set, just as the discrete distributions do, but with no jumps. We will see that for all p (except for the "continuous" case $p = 1/2$) the length of the graph of F is 2. The appearance of the graph of F depends critically on the value of p. While for small values of p this graph does not look very far from that of a typical discrete distribution, the graphs corresponding to values of p close to $1/2$ would be almost undistinguishable from the linear uniform case $F_0(x) = x$, except for a few small inwards peaks. The striking fact is that such peaks must be important enough to account for the increase in length from $\sqrt{2}$ (corresponding to $p = 1/2$) to 2 (for any other p). Of course the point is that the peaks appear "everywhere" in the graph as suggested by the approximations in Figure 2 below.

Define G as the hypograph of F, that is $G = \{(x, y) : x \in [0, 1], 0 \le y \le F(x)\}$. Note that the μ-boundary of G, with respect to the unit square is $\mathrm{Gr}(F)$, the graph of F, that is $\mathrm{Gr}(F) = \{(x, F(x)) : x \in [0, 1]\}$.

We next show that Theorem 3.1 can be applied in this case so that the estimator (2.9) can be used to consistently estimate the length of $\mathrm{Gr}(F)$.

It can be proved that $L_0(\mathrm{Gr}(F)) = 2$. We will not fully develop all the details. Let us just outline some relevant points. First note that, since F is a bounded variation function, $\sup_{\mathcal{P}}\{\sum_i |F(x_i) - F(x_{i-1})|\} < \infty$ where the supremum is taken on the set $\mathcal{P}$ of all possible partitions $0 = x_0 < x_1 < \ldots < x_n = 1$ of the unit interval. This, together with the norm inequality $\|v\| \le \sum_i |v_i|$, for all $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$, $n \in \mathbb{N}$, entails that $\mathrm{Gr}(F)$ is a rectifiable curve. Second, as indicated in Section 2.1, the rectifiability of $\mathrm{Gr}(F)$ in turn implies that $\mathcal{H}^1(\mathrm{Gr}(F)) = \mathrm{Length}(\mathrm{Gr}(F))$. This follows from the definitions of $\mathcal{H}^1$ and the "rectifiable" length defined in Section 2. In the third place note that, as mentioned in Mattila [12], page 80, if Γ is a rectifiable curve, its Minkowski content $L_0(\Gamma)$ coincides with the corresponding one-dimensional Hausdorff measure; see also Federer [9], 3.2.37-44.

The true value of $L_0(\mathrm{Gr}(F))$ could be obtained as the limit of a polygonal approximation, see *e.g.* `http://en.wikipedia.org/wiki/Arc_length`. For example, we could construct an approximating sequence as follows: let $k \ge 1$ be an integer, and introduce the distribution function F_k, which is the piecewise linear and

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

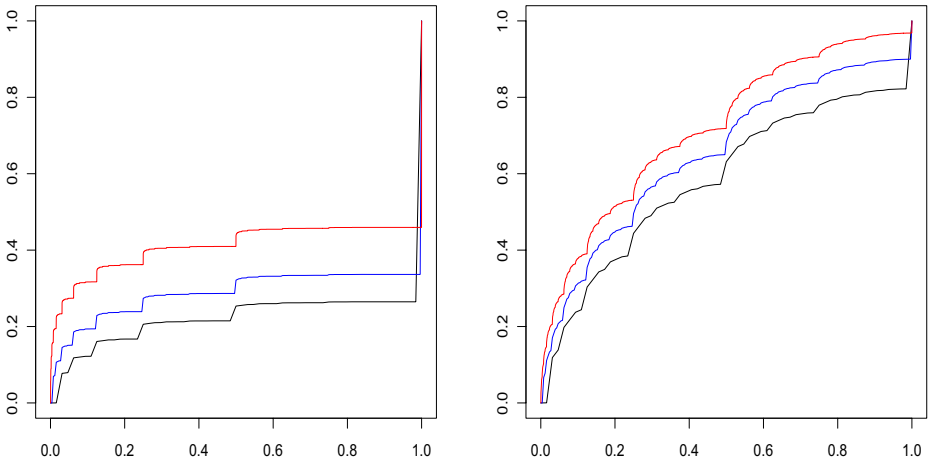


FIGURE 2. Approximating functions F_k for $p = 0.05$ (left) and $p = 0.25$ (right) and $k = 6$, 8, 12.

continuous interpolation of F such that for each integer $1 \leq i \leq 2^k$, $F_k(i/2^k) = F(i/2^k)$. Figure 2 shows the graphs of the approximating functions F_k for different values of k and p.

Finally, to prove that (2.4) holds we only need to prove that (2.3) is fulfilled since this is a stronger assumption. In this case the double standardness property (2.3) is valid since, as F is a monotone non decreasing function, inequalities in (2.3) hold for $\delta_0 = 1/4$. □

A pathological example. As indicated in O'Neill [13], see also `http://mcs.open.ac.uk/tcon2/encycloWeb.pdf`, "a standard example of a 1-rectifiable set in the plane is a countable union of circles whose centres are dense in the unit square and with radii having a finite sum". So consider a such union S of open balls included in the square $C = [1/2, 3/4]^2$ whose centres are dense in C and denote by T the union of the corresponding circumferences. We may construct S such that $\mu(S) = 1/32$. Define $G = C \setminus S$. As S is dense in C, we have that G coincides with ∂G. Hence ∂G has a positive area, $\mu(\partial G) = \mu(G) = 1/16 - 1/32 = 1/32$. From the definition of the Minkowski content this entails $L_0(G) = \infty$, which do not coincides with $\text{Length}(T) < \infty$. This example shows that $L_0(G) < \infty$ is itself an important shape restriction that excludes from consideration some pathological examples, as that just described, in which our approach would fail.

4. Open problems and final remarks

Universal estimation. In our view, the main challenge suggested by the results in this paper is to decide whether or not the estimator L_n defined in (2.9) is universally consistent in the sense that L_n converges (at least in probability) to $L_0(G)$ as $n \to \infty$, under suitable conditions on the smoothing parameters ϵ_n but with no additional assumption on G (except for $L_0(G) < \infty$ and $\partial G_\mu = \partial G$). Though our shape condition (2.4) is indeed very general, it is not clear to us whether it could be dropped using another method of proof. In that case, our estimate L_n would be indeed universally consistent.

Strong consistency. Another technical issue of some interest is to study whether under our assumption (2.4) the strong (almost sure) consistency holds under suitable conditions on ϵ_n. This remains as an open problem. The difficulty is clearly due to the definition of the estimate as an empirical measure where each sample point contributes conditionally to many others, thus complicated conditional events have to be dealt with. As a referee pointed out to us, this could be a place to use local empirical processes tools and U-statistics type techniques as well as some appropriate concentration inequality.

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press

Extension to further designs. We could also think of analyzing the estimation of $L_0(G)$ from different sampling designs. A relevant issue in this line would be to tackle the estimation using just *one inside sample*, instead of having points inside and outside G. This is clearly a more difficult problem which, typically, will require stronger assumptions on G; see. *e.g.*, Cuevas *et al.* [6]. Another interesting question is to consider a generalized version of our inside-outside design incorporating the possibility of "error in variables", in the sense that the membership, to either G or G^c, of the observations Z_i is correctly identified with a probability p (in our case $p = 1$); see Mammen and Tsybakov [11] for related ideas. While this design looks more realistic than that considered here, it seems to us that our results could be extended for it with some technical changes and our estimator will work as well under this design (at some unavoidable cost in efficiency).

Some practical issues. The comparison, *via* simulations, with the estimators considered in Cuevas *et al.* [5], Pateiro−López and Rodríguez−Casal [14] and Jiménez and Yukich [10] is another obvious pending task. Some real-data problems presented or commented in these papers could also be considered.

References

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

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

[3] A.J. Baddeley and E.B. Vedel-Jensen, *Stereology for Statisticians.* Chapman & Hall, London (2005).

[4] A. Cuevas, and R. Fraiman, Set estimation, in *New Perspectives on Stochastic Geometry*, edited by W.S. Kendall and I. Molchanov. Oxford University Press (2010) 374–397.

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

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

[7] P. Erdős, Some remarks on the measurability of certain sets. *Bull. Amer. Math. Soc.* **51** (1945) 728–731.

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

[9] H. Federer, Geometric Measure Theory. Springer, New York (1969).

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

[11] E. Mammen and A.B. Tsybakov, Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Stat.* **23** (1995) 502–524.

[12] P. Mattila, *Geometry of Sets and Measures in Euclidean Spaces: Fractals and rectifiability.* Cambridge University Press, Cambridge (1995).

[13] T.C. O'Neill, Geometric measure theory, in *Encyclopedia of Mathematics, Supplement III*. Kluwer Academic Publishers (2002).

[14] B. Pateiro-López and A. Rodríguez-Casal, Length and surface area estimation under convexity type restrictions. *Adv. Appl. Probab.* **40** (2008) 348–358.

[15] E. Villa, On the outer Minkowski content of sets. *Ann. Mat. Pura Appl.* **188** (2009) 619–630.

https://doi.org/10.1051/ps/2011160 Published online by Cambridge University Press