ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Polynomial volume estimation and its applications

Antonio Cuevas [a,*], Beatriz Pateiro-López [b]

[a] Departamento de Matemáticas, Universidad Autónoma de Madrid, Spain
[b] Departamento de Estadística, Análisis Matemático y Optimización, Universidad de Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 7 December 2016
Received in revised form 23 November 2017
Accepted 25 November 2017
Available online xxxx

Keywords:
Set estimation
Volume estimation
Boundary length estimation

ABSTRACT

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

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

1. Introduction

The general background. Set estimation

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

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

Some notations and basic definitions

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

* Corresponding author.
E-mail address: antonio.cuevas@uam.es (A. Cuevas).

https://doi.org/10.1016/j.jspi.2017.11.005
0378-3758/

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

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

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

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

Statement of the problem

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

A brief overview of boundary measure and volume estimation

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

(i) the assumed model to generate the sample observations,

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

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

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

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

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

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

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

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

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

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

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

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

Please cite this article in press as: Cuevas A., Pateiro-López B., Polynomial volume estimation and its applications. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.11.005.