Detection of low dimensionality and data denoising via set estimation techniques

Catherine Aaron, Alejandro Cholaquidis, Antonio Cuevas

Stochastic detection of some topological and geometric features

Catherine Aaron[a], Alejandro Cholaquidis[b] and Antonio Cuevas[c]
[a] Université Blaise-Pascal Clermont II, France
[b] Centro de Matemática, Universidad de la República, Uruguay
[c] Departamento de Matemáticas, Universidad Autónoma de Madrid

Abstract

This work is closely related to the theories of set estimation and manifold estimation. Our object of interest is a, possibly lower-dimensional, compact set $S \subset \mathbb{R}^d$. The general aim is to identify (via stochastic procedures) some qualitative or quantitative features of S, of geometric or topological character. The available information is just a random sample of points drawn on S. The term "to identify" means here to achieve a correct answer almost surely (a.s.) when the sample size tends to infinity. More specifically the paper aims at giving some partial answers to the following questions:
1. Is S full dimensional?
2. If S is full dimensional, is it "close to a lower dimensional set" $\mathcal{M}$?
3. If S is "close to a lower dimensional $\mathcal{M}$", can we
   a) estimate $\mathcal{M}$?
   b) estimate some functionals defined on $\mathcal{M}$ (in particular, the Minkowski content of $\mathcal{M}$)?

The theoretical results are complemented with some simulations and graphical illustrations.

1 Introduction

The general setup. Some related literature. The emerging statistical field currently known as *manifold estimation* (or, sometimes, *statistics on manifolds*, or *manifold learning*) is the result of the confluence of, at least, three classical theories: (a) the analysis of directional (or circular) data Mardia and Jupp (2000), Bhattacharya and Patrangenaru (2008) where the aims are similar to those of the classical statistics but the data are supposed to be drawn on the sphere or, more generally, on a lower-dimensional manifold; (b) the study of non-linear methods of dimension reduction, Delicado (2001), Hastie and Stuetzle (1989), aiming at recovering a lower-dimensional structure from random points taken around it, and (c) some techniques of stochastic geometry Chazal and Lieutier (2005) and set estimation Cuevas and Fraiman (2010), Cholaquidis et al. (2014), Cuevas et al. (2007) whose purpose is to estimate some relevant quantities of a set (or the set itself) from the information provided by a random sample whose distribution is closely related to the set.

There are also strong connections with the theories of persistent homology and computational topology, Carlsson (2009), Niyogi, Smale and Weinberger (2011), Fasy et al. (2014).

In all these studies, from different points of view, the general aim is similar: one wants to get information (very often of geometric or topological type) on a set from a sample of points. To be more specific, let us mention some recent references on these topics, roughly grouped according the subject (the list is largely non-exhaustive):

Manifold recovery from a sample of points, Genovese et al. (2012b); Genovese et al (2012c).

Inference on dimension, Fefferman et al. (2016), Brito et al. (2013).

Estimation of measures (perimeter, surface area, curvatures), Cuevas et al. (2007), Jiménez and Yukich (2011), Berrendero et al. (2014).

Estimation of some other relevant quantities in a manifold, Niyogi, Smale and Weinberger (2008), Chen and Müller (2012).

Dimensionality reduction, Genovese et al. (2012a), Tenebaum et al. (2000).

The problems under study. The contents of the paper. Let $X_1, \ldots, X_n$ be random sample points drawn on an unknown compact set $S \subset \mathbb{R}^d$. We consider two different models:

> *The noiseless model*: the data $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ are taken from a distribution whose support is S itself; Aamari and Levrard (2015), Amenta et al. (2002), Cholaquidis et al. (2014), Cuevas and Fraiman (1997).
>
> *The parallel (noisy) model*: The data $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ have a distribution whose support is the parallel set S of points within a distance to $\mathcal{M}$ smaller than R_1, for some $R_1 > 0$, where $\mathcal{M}$ is a d'-dimensional set with $d' \leq d$; Berrendero et al. (2014). Note that other different models "with noise" are considered in Genovese et al. (2012a), Genovese et al. (2012b) and Genovese et al (2012c).

Our general aim is to identify, eventually almost surely (a.s.), some geometric or topological properties of $\mathcal{M}$ or S. Note that with an eventual a.s. identification procedure, no statistical test is needed (asymptotically) since eventually the property (or the lack of it) is identified with no error. Moreover, the identification methods are "algorithmic" in the sense that they are based on automatic procedures to perform them with arbitrary precision. This will require to impose some restrictions on $\mathcal{M}$ or S. Section 2 includes all the relevant definitions, notations and basic geometric concepts we will need.

In Section 3 we first develop, under the noiseless model, an algorithmic procedure to identify, eventually, a.s., whether or not S has an empty interior; this is achieved in Theorems 1 and 2 below. A positive answer would essentially entail (under some conditions, see the beginning of Section 3) that we are in fact in the noiseless model and $\mathcal{M}$ has a dimension smaller than that of the ambient space.

Then, assuming the noisy model and $\mathring{\mathcal{M}} = \emptyset$, Theorems 3 (i) and 4 (i) provide two methods for the estimation of the maximum level of noise R_1, giving also the corre-

sponding convergence rates. If R_1 is known in advance, the results in Theorems 3 and 4 allow us also to decide whether or not the "inside set" $\mathcal{M}$ has an empty interior or not.

In Section 4 we consider again the noisy, model where the data are drawn on the R_1 -parallel set around a lower dimensional set $\mathcal{M}$. We propose a method to "denoise" the sample, which essentially amounts to estimate $\mathcal{M}$ from sample data drawn around the parallel set S around $\mathcal{M}$.

In Section 5 we consider the problem of estimating the d'-dimensional Minkowski measure of $\mathcal{M}$ under both the noiseless and the noisy model. We assume throughout the section that the dimension d' (in Hausdorff sense, see below) of the set $\mathcal{M}$ is known.

Finally, in Section 6 we present some simulations and numerical illustrations.

2 Some geometric background

This section is devoted to make explicit the notations, and basic concepts and definitions (mostly of geometric character) we will need in the rest of the paper.

Some notation. Given a set $S \subset \mathbb{R}^d$, we will denote by $\mathring{S}$, $\overline{S}$ and ∂S the interior, closure and boundary of S, respectively with respect to the usual topology of $\mathbb{R}^d$. We will also denote $\rho(S) = \sup_{x \in S} d(x, \partial S)$. Notice that $\rho(S) > 0$ is equivalent to $\mathring{S} \neq \emptyset$.

The parallel set of S of radii ε will be denoted as $B(S, \varepsilon)$, that is $B(S, \varepsilon) = \{y \in \mathbb{R}^d : \inf_{x \in S} \|y - x\| \leq \varepsilon\}$. If $A \subset \mathbb{R}^d$ is a Borel set, then $\mu_d(A)$ (sometimes just $\mu(A)$) will denote its Lebesgue measure. We will denote by $\mathcal{B}(x, \varepsilon)$ (or $\mathcal{B}_d(x, \varepsilon)$, when necessary) the closed ball in $\mathbb{R}^d$, of radius ε, centred at x, and $\omega_d = \mu_d(\mathcal{B}_d(x, 1))$. Given two compact non-empty sets $A, B \subset \mathbb{R}^d$, the *Hausdorff distance* or *Hausdorff-Pompei distance* between A and C is defined by

$$d_H(A, C) = \inf\{\varepsilon > 0 : \text{such that } A \subset B(C, \varepsilon) \text{ and } C \subset B(A, \varepsilon)\}. \tag{1}$$

Some geometric regularity conditions for sets. The following conditions have been used many times in set estimation topics see, e.g., Niyogi, Smale and Weinberger (2008), Genovese et al. (2012b), Cuevas and Fraiman (2010) and references therein.

Definition 1. *Let $S \subset \mathbb{R}^d$ be a closed set. The set S is said to satisfy the outside r-rolling condition if for each boundary point $s \in \partial S$ there exists some $x \in S^c$ such that $\mathcal{B}(x, r) \cap \partial S = \{s\}$. A compact set S is said to satisfy the inside r-rolling condition if $\overline{S^c}$ satisfies the outside r-rolling condition at all boundary points.*

Definition 2. *A set $S \subset \mathbb{R}^d$ is said to be r-convex, for $r > 0$, if $S = C_r(S)$, where*

$$C_r(S) = \bigcap_{\left\{\mathring{\mathcal{B}}(x,r):\ \mathring{\mathcal{B}}(x,r) \cap S = \emptyset\right\}} \left(\mathring{\mathcal{B}}(x, r)\right)^c, \tag{2}$$

is the r-convex hull of S. When S is r-convex, a natural estimator of S from a random sample $\mathcal{X}_n$ of points (drawn on a distribution with support S), is $C_r(\mathcal{X}_n)$.

Following the notation in Federer (1959), let $\mathrm{Unp}(S)$ be the set of points $x \in \mathbb{R}^d$ with a unique projection on S.

Definition 3. *For $x \in S$, let* $\mathrm{reach}(S,x) = \sup\{r > 0 : \mathring{\mathcal{B}}(x,r) \subset \mathrm{Unp}(S)\}$. *The reach of S is defined by* $\mathrm{reach}(S) = \inf\{\mathrm{reach}(S,x) : x \in S\}$, *and S is said to be of positive reach if* $\mathrm{reach}(S) > 0$.

The study of sets with positive reach was started by Federer (1959); see Thäle (2008) for a survey. This is now a major topic in different problems of manifold learning or topological data analysis. See, e.g., Adler et al. (2016) for a recent reference.

The conditions established in Definitions 1, 2 and 3 have an obvious mutual affinity. In fact, they are collectively referred to as "rolling properties" in Cuevas, Fraiman and Pateiro-López (2012). However, they are not equivalent: if the reach of S is r then S is r-convex, which in turn implies the (outer) r-rolling condition. The converse implications are not true in general; see Cuevas, Fraiman and Pateiro-López (2012) for details.

Definition 4. *A set $S \subset \mathbb{R}^d$ is said to be standard with respect to a Borel measure ν in a point x if there exists $\lambda > 0$ and $\delta > 0$ such that*

$$\nu(\mathcal{B}(x,\varepsilon) \cap S) \geq \delta \mu_d(\mathcal{B}(x,\varepsilon)), \quad 0 < \varepsilon \leq \lambda. \tag{3}$$

A set $S \subset \mathbb{R}^d$ is said to be standard if (3) *hold for all $x \in S$.*

The following result will be useful below. It establishes a simple connection between standardness and the inside r-rolling condition.

Proposition 1. *Let $S \subset \mathbb{R}^d$ the support of a Borel measure ν, whose density f with respect to the Lebesgue measure is bounded from below by f_0, if S satisfies the inside r-rolling condition for all $x \in \partial S$ then it is standard with respect to ν, for any $\delta \leq f_0/3$ and $\lambda = r$.*

Proof. Let $0 < \varepsilon \leq r$ and $x \in S$, if $d(x, \partial S) \geq r$ the result is obvious. Let $x \in S$ such that $d(x,\partial S) < r$. $\mathrm{reach}(\overline{S^c}) \geq r$ implies that there exists $z \in \mathbb{R}^d$ such that $x \in \mathcal{B}(z,r) \subset S$. Then, for all $\varepsilon \leq r$

$$\nu(\mathcal{B}(x,\varepsilon) \cap S) \geq \nu(\mathcal{B}(x,\varepsilon) \cap \mathcal{B}(z,r)) \geq f_0 \mu_d(\mathcal{B}(x,\varepsilon) \cap \mathcal{B}(z,r)) \geq \frac{f_0}{3} \mu_d(\mathcal{B}(x,\varepsilon))$$

□

Some basic definitions on manifolds.

The following basic concepts are stated here for the sake of completeness and notational clarity. More complete information on these topics can be found, for example, in the classical textbooks Boothby (1975) and Do Carmo (1992). See also the nice book Galbis and Maestre (2010) and the summary (Zhang, 2011, chapter 3).

Definition 5. *A topological manifold $\mathcal{M}$ of dimension k in $\mathbb{R}^d$ is a subset of $\mathbb{R}^d$ with $k \leq d$ such that every point in $\mathcal{M}$ has a neighbourhood homeomorphic to an open set in $\mathbb{R}^k$. We will say that $\mathcal{M}$ is a regular k-surface, or a differentiable k-manifold of class $p \geq 1$, is there is a family (often called atlas) $\mathcal{V} = \{(V_\alpha, x_\alpha)\}$ of pairs (V_α, x_α) (often called parametrizations, coordinate systems or charts) such that the V_α are open sets in $\mathbb{R}^k$ and the $x_\alpha : V_\alpha \to \mathcal{M}$ are functions of class p satisfying: (i) $\mathcal{M} = \cup_\alpha x_\alpha(V_\alpha)$, (ii) every x_α is a homeomorphism between V_α and $x_\alpha(V_\alpha)$ and (iii) for every $v \in V_\alpha$ the differential $dx_\alpha(u) : \mathbb{R}^k \to \mathbb{R}^d$ is injective.*

The notion of manifold with boundary is defined in a similar way by replacing $\mathbb{R}^k$ with $\mathbb{R}^k_+ = \{x \in \mathbb{R}^k : \ x_k \geq 0\}$.

A manifold $\mathcal{M}$ is said to be *compact* when it is compact as a topological space. As a direct consequence of the definition of compactness, any compact manifold has a finite atlas. Typically, in most relevant cases the required atlas for a manifold has, at most, a denumerable set of charts.

An equivalent definition of the notion of manifold (see Do Carmo (1992, Def 2.1, p. 2)) can be stated in terms of *parametrizations or coordinate systems* of type $(U_\alpha, \varphi_\alpha)$ with $\varphi_\alpha : V_\alpha \subset \mathcal{M} \to \mathbb{R}^k$. The conditions would be completely similar to the previous ones, except that the φ_α are defined in a reverse way to that of Definition 5.

In the simplest case, just one chart $x : V \to \mathcal{M}$ is needed. The structures defined in this way are sometimes called *planar manifolds*.

Some background on geometric measure theory. The important problem of defining lower-dimensional measures (surface measure, perimeter, etc.) has been tackled in different ways. The book by Mattila (1995) is a classical reference. We first recall the so-called Hausdorff measure. It is defined for any separable metric space $(\mathcal{M}, \rho)$. Given $\delta, r > 0$ and $E \subset \mathcal{M}$, let

$$\mathcal{H}^r_\delta(E) = \inf \left\{ \sum_{j=1}^{\infty} (\mathrm{diam}(B_j))^r : \ E \subset \cup_{j=1}^{\infty} B_j, \ \mathrm{diam}(B_j) \leq \delta \right\},$$

where $\mathrm{diam}(B) = \sup\{\rho(x, y) : x, y \in B\}$, $\inf \emptyset = \infty$. Now, define $\mathcal{H}^r(E) = \lim_{\delta \to 0} \mathcal{H}^r_\delta(E)$.

The set function $\mathcal{H}^r$ is an outer measure. If we restrict $\mathcal{H}^r$ to the measurable sets (according to standard Caratheodory's definition) we get the r-dimensional Hausdorff measure on $\mathcal{M}$.

The Hausdorff dimension of a set E is defined by

$$\dim_H(E) = \inf\{r : \mathcal{H}^r(E) = 0\} = \sup\{r : \mathcal{H}^r(E) = \infty\}. \tag{4}$$

It can be proved hat, when $\mathcal{M}$ is a k-dimensional smooth manifold, $\dim_H(\mathcal{M}) = k$.

Another popular notion to define lower-dimensional measures for the case $\mathcal{M} \subset \mathbb{R}^d$ is the *Minkowski content*. For an integer $d' < d$, denote $\omega_{d-d'}$ the volume of the unit ball

in $\mathbb{R}^{d-d'}$. Now, define the d'-dimensional Minkowski content of a set $\mathcal{M}$ by

$$L_0^{d'}(\mathcal{M}) = \lim_{\varepsilon \to 0} \frac{\mu_d\left(\mathcal{B}(\mathcal{M}, \varepsilon)\right)}{\omega_{d-d'} \varepsilon^{d-d'}}, \tag{5}$$

provided that this limit does exist.

In what follows we will often denote $L_0^{d'}(\mathcal{M}) = L_0(\mathcal{M})$, when the value of d' is understood. The term "content" is used here as a surrogate for "measure", as the expression (5) does not generally leads to a true (sigma-additive) measure.

A compact set $\mathcal{M} \subset \mathbb{R}^d$ is said to be *d'-rectifiable* if there exists a compact set $K \subset \mathbb{R}^{d'}$ and a Lipschitz function $f : \mathbb{R}^{d'} \to \mathbb{R}^d$ such that $\mathcal{M} = f(K)$. In (Federer, 1969, Th. 3.2.39) it is proved that for a compact d'-rectifiable set $\mathcal{M}$, $L_0^{d'}(\mathcal{M}) = \mathcal{H}^{d'}(\mathcal{M})$. More details on the relations between the rectifiability property and the structure of manifold can be found in (Federer, 1969, Th. 3.2.29).

3 Checking closeness to lower dimensionality

We consider here the problem of identifying whether or not the set $\mathcal{M} \subset \mathbb{R}^d$ (not necessarily a manifold) has an empty interior.

Note that, if $\mathcal{M} \subset \mathbb{R}^d$ is "regular enough", $\dim_H(\mathcal{M}) < d$ is in fact equivalent to $\mathring{\mathcal{M}} = \emptyset$. Indeed, in general $\dim_H(\mathcal{M}) < d$ implies $\mathring{\mathcal{M}} = \emptyset$. The converse implication is not always true, even for sets fulfilling the property $\mathcal{H}^d(\partial\mathcal{M}) = 0$ (see ((2007))). However it holds if $\mathcal{M}$ has positive reach, since in this case $\mathcal{H}^{d-1}(\partial\mathcal{M}) < \infty$ (see the comments after Th. 7 and inequality (27) in Ambrosio, Colesanti and Villa (2008)).

Also, clearly, in the case where $\mathcal{M}$ is a manifold, the fact that $\mathcal{M}$ has an empty interior *amounts to say that its dimension is smaller than that of the ambient space.*

3.1 The noiseless model

We first consider the case where the sample information follows the noiseless model explained in the Introduction, that is, the data $X_1, \ldots, X_n$ are assumed to be an *iid* sample of points drawn from an unknown distribution P_X with support $\mathcal{M} \subset \mathbb{R}^d$. When $\mathcal{M}$ is a lower-dimensional set, this model can be considered as an extension of the classical theory of directional (or spherical) data, in which the sample data are assumed to follow a distribution whose support is the unit sphere in $\mathbb{R}^d$. See, e.g.,Mardia and Jupp (2000).

Our main tool here will be the simple *offset* or *Devroye-Wise estimator* (see Devroye and Wise (1980)) given by

$$\hat{S}_n(r) = \bigcup_{i=1}^{n} \mathcal{B}(X_i, r). \tag{6}$$

More specifically, we are especially interested in the "boundary balls" of $\hat{S}_n(r)$.

Definition 6. *Given $r > 0$ let $\hat{S}_n(r)$ be a set estimator of type (6) based on the data $x_1, \ldots, x_n$. We will say that $\mathcal{B}(x_i, r)$ is a boundary ball of $\hat{S}_n(r)$ if there exists a point $y \in \partial\mathcal{B}(x_i, r)$ such that $y \in \partial\hat{S}_n(r)$. The "peeling" of $\hat{S}_n(r)$, denoted by* $\mathrm{peel}(\hat{S}_n(r))$*, is*

the union of all non-boundary balls of $\hat{S}_n(r)$. *In other words,* $\text{peel}(\hat{S}_n(r))$ *is the result of removing from* $\hat{S}_n(r)$ *all the boundary balls.*

The following theorem is the main result of this section. It relates, in statistical terms, the emptiness of $\mathring{\mathcal{M}}$ with $\text{peel}(\hat{S}_n)$.

Theorem 1. *Let* $\mathcal{M} \subset \mathbb{R}^d$ *be a compact non-empty set. Then under the model and notations stated in the two previous paragraphs we have,*

(a) if $\mathring{\mathcal{M}} = \emptyset$, *and* $\mathcal{M}$ *fulfils the outside rolling condition for some* $r > 0$, *then* $\text{peel}(\hat{S}_n(r')) = \emptyset$ *for any set* $\hat{S}_n(r')$ *of type (6) with* $r' < r$.

(b) In the case $\mathring{\mathcal{M}} \neq \emptyset$, *assume that there exists a ball* $\mathcal{B}(x_0, \rho_0) \subset \mathring{\mathcal{M}}$ *such that* $\mathcal{B}(x_0, \rho_0)$ *is standard w.r.t to* P_X, *with constants* δ *and* λ *(see Definition (4)). Then* $\text{peel}(\hat{S}_n(r_n)) \neq \emptyset$ *eventually, a.s., where* r_n *is a radius sequence such that:* $(\kappa \frac{\log(n)}{n})^{1/d} \leq r_n \leq \min\{\rho_0/2, \lambda\}$ *for a given* $\kappa > (\delta\omega_d)^{-1}$.

Proof. (a) Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an iid sample of X with distribution P_X. To prove that $\text{peel}(\hat{S}_n(r')) = \emptyset$ for all $r' < r$ it is enough to prove that for all $r' < r$ and for all $i = 1, \ldots, n$ there exists a point $y_i \in \partial\mathcal{B}(X_i, r')$ such that $y \notin \mathcal{B}(X_j, r')$ for all $X_j \neq X_i$. Since $\mathcal{M}$ is closed and $\mathring{\mathcal{M}} = \emptyset$, $\partial\mathcal{M} = \mathcal{M}$. The outside rolling ball property implies that for all $X_i \in \mathcal{M}$ exists $z_i \in \mathcal{M}^c$ such that $\mathcal{B}(z_i, r) \cap \mathcal{M} = \{X_i\}$. Let us denote $u_i = \frac{z_i - X_i}{r}$ and consider $y_i = X_i + r'u_i$. Clearly $y_i \in \partial\mathcal{B}(X_i, r')$. From $\mathcal{B}(y_i, r') \subset \mathcal{B}(z_i, r)$ and the outside rolling ball property we get that $\{X_i\} \subset \mathcal{B}(y_i, r') \cap \mathcal{X}_n \subset \mathcal{B}(z_i, r) \cap \mathcal{M} \subset \{X_i\}$ so that, for all $X_j \neq X_i$, $X_j \notin \mathcal{B}(y_i, r')$ and thus, $y_i \notin \mathcal{B}(X_j, r')$.

(b) First we are going to prove that

$$\text{if } \Big(\frac{C\log(n)}{\delta\omega_d n}\Big)^{1/d} \leq r_n \leq \min\{\rho_0/2, \lambda\} \text{ for a given } C > 1 \text{ then:}$$
$$\text{eventually a.s. for all } y \in \mathcal{B}(x_0, 2r_n) \text{ we have } \mathring{\mathcal{B}}(y, r_n) \cap \mathcal{X}_n \neq \emptyset. \quad (7)$$

Consider only $n \geq 3$ and let $\varepsilon_n = (\log(n))^{-1}$, there is a positive constant τ_d, such that we can cover $\mathcal{B}(x_0, 2r_n)$ with $\nu_n = \tau_d \varepsilon_n^{-d}$ balls of radius $r_n\varepsilon_n$ centred in $\{t_1, \ldots, t_{\nu_n}\}$. Let es define

$$p_n = P_X\Big(\exists y \in \mathcal{B}(x_0, 2r_n), \mathring{\mathcal{B}}(y, r_n) \cap \mathcal{X}_n = \emptyset\Big),$$

then,

$$p_n \leq \sum_{i=1}^{\nu_n} P_X\Big(\mathcal{B}(t_i, r_n(1-\varepsilon_n)) \cap \mathcal{X}_n = \emptyset\Big). \quad (8)$$

Notice that for any given i,

$$P_X\Big(\mathcal{B}\big(t_i, r_n(1-\varepsilon_n)\big) \cap \mathcal{X}_n = \emptyset\Big) = \Big(1 - P_X\big(\mathcal{B}(t_i, r_n(1-\varepsilon_n))\big)\Big)^n.$$

Since $r_n \leq \rho_0/2$, $t_i \in \mathcal{B}(x_0, \rho_0)$, then using that $\mathcal{B}(x_0, \rho_0)$ is standard with the same δ and λ,

$$P_X\Big(\mathcal{B}\big(t_i, r_n(1-\varepsilon_n)\big) \cap \mathcal{X}_n = \emptyset\Big) \leq \Big(1 - \omega_d \delta r_n^d (1-\varepsilon_n)^d\Big)^n$$
$$\leq \Big(1 - C\frac{\log(n)}{n}\big(1-\varepsilon_n\big)^d\Big)^n.$$

Which, according to (8) provides:

$$p_n \leq \tau_d \varepsilon_n^{-d}\Big(1 - C\frac{\log(n)}{n}\big(1-\varepsilon_n\big)^d\Big)^n \leq \tau_d \varepsilon_n^{-d} n^{-C(1-\varepsilon_n)^d},$$

where we have used that $(1-x)^n \leq \exp(-nx)$. Since $C > 1$, we can choose $\beta > 1$ such that $p_n/n^{-\beta} \to 0$, then, $\sum p_n < \infty$. Finally (7) follows as a direct application of Borel Cantelli Lemma. Observe that (7) implies that $x_0 \in \hat{S}_n(r_n)$ eventually a.s., so there exists X_i such that $x_0 \in \mathcal{B}(X_i, r_n)$ eventually a.s. Again by (7) we get that, eventually a.s. for all $z \in \partial\mathcal{B}(X_i, r_n)$ there exists X_j such that $z \in \mathring{\mathcal{B}}(X_j, r_n)$ and so $z \notin \partial\hat{S}_n(r_n)$, which implies that, eventually a.s., $\mathcal{B}(X_i, r_n)$ is not removed by the peeling process and so $\text{peel}(\hat{S}_n(r_n)) \neq \emptyset$ eventually, a.s..

□

Remark 1. *Observe that the standardness conditions required in Theorem 1 b) is fulfilled if* (3) *holds for $\nu = \mu_d$ and if P_X has a density f bounded from below by a positive constant.*

The following result can be seen as an application of Theorem 1 for differentiable manifolds, with a specific, data driven, choice of r_n.

Theorem 2. *Let $\mathcal{M}$ be a d'-dimensional compact manifold in $\mathbb{R}^d$. Suppose that the sample points $X_1, \ldots, X_n$ are drawn from a probability measure P_X with support $\mathcal{M}$ which has a continuous density f with respect the d'-dimensional Hausdorff measure on $\mathcal{M}$, and $f(x) > f_0$ for all $x \in \mathcal{M}$. Let us define, for any $\beta > 6^{1/d}$, $r_n = \beta \max_i \min_{j\neq i} \|X_j - X_i\|$. Then,*

i) if $d' = d$ and $\partial\mathcal{M}$ is a $\mathcal{C}^2$ manifold then $\text{peel}(\hat{S}_n(r_n)) \neq \emptyset$ *eventually, a.s..*

ii) if $d' < d$ and $\mathcal{M}$ is a $\mathcal{C}^2$ manifold without boundary, then $\text{peel}(\hat{S}_n(r_n)) = \emptyset$ *eventually, a.s..*

Proof. *i)* As $d' = d$ then $\partial\mathcal{M}$ is a $\mathcal{C}^2$ a compact $(d-1)$-manifold thus, by Theorem 1 in Walther, G. (1999) $\mathcal{M}$ fulfils both the inside and outside rolling ball property for a small enough radius $r > 0$; note that such result can be applied since the $\mathcal{C}^2$ assumption on the compact hypersurface $\partial\mathcal{M}$ implies the Lipschitz condition for the outward normal vector and the interior of every path-connected component of $\mathcal{M}$ is guaranteed from the fact that $\mathcal{M}$ is the support of an absolutely continuous distribution. By Lemma 2.3 in Pateiro-López and Rodríguez-Casal (2012) and Proposition 1, $\mathcal{M}$ satisfies the standardness condition established in Definition 4 with $\nu = P_X$, $\delta = f_0/3$ and $\lambda < r$.

In order to prove that r_n fulfils the conditions in 1 b) we will use Theorem 1.1 in Penrose (1999). First observe that in the full-dimensional case $d' = d$ the intrinsic volume in $\mathcal{M}$ coincides with the restricted Lebesgue measure; see (Taylor, 2006, Prop. 12.6). As a consequence, f is equal to the density of P_X w.r.t. the Lebesgue measure restricted to $\mathcal{M}$. Let us denote $f_1 = \min_{x\in\partial\mathcal{M}} f(x)$, then with probability one,

$$\frac{nr_n^d\omega_d}{\log(n)\beta^d} \to \max\left\{\frac{1}{f_0}, \frac{2(d-1)}{df_1}\right\} \geq \frac{1}{f_0}.$$

Then for n large enough,

$$r_n \geq \Big(\frac{\log(n)}{n}\frac{\beta^d}{\omega_d 2 f_0}\Big)^{1/d},$$

now if we denote $\kappa = \beta^d/(\omega_d 2 f_0)$, it fulfils that $\kappa > (\delta\omega_d)^{-1}$, so we are in the hypotheses of Theorem 1 *b*) and then we can conclude $\mathrm{peel}(\hat{S}_n(r_n)) \neq \emptyset$ eventually, with probability 1.

ii) Notice that we can use Theorem 1 *a*) indeed, as $\mathcal{M}$ is a $\mathcal{C}^2$ compact manifold of $\mathbb{R}^d$ by (Thäle, 2008, Prop. 14) it has a positive reach and, thus, it satisfies the outside rolling ball condition (for some radius $r > 0$). Then it remains to be proved that $r_n \leq r$ for n large enough. Let us endow $\mathcal{M}$ with the standard Riemannian structure, where a local metric is defined on every tangent space just by restricting on it the standard inner product on $\mathbb{R}^d$. Under or smoothness assumptions, the Riemannian measure induced by such a metric on the manifold $\mathcal{M}$ agrees with the d'-dimensional Hausdorff measure on $\mathcal{M}$ (this is just a particular case of the Area Formula; see (Federer, 1969, 3.2.46)). So we may use Theorem 5.1 in Penrose (1999). As a consequence of that result

$$\max_i \min_{j\neq i} \gamma(X_i, X_j) = \mathcal{O}\left(\left(\frac{\log n}{n}\right)^{1/d}\right), \text{ a.s.,} \tag{9}$$

where γ denotes the geodesic distance on $\mathcal{M}$ associated with the Riemannian structure. Now, since the Euclidean distance is smaller than the geodesic distance, we have for all i, j, $\|X_j - X_i\| \leq \gamma(X_i, X_j)$ and $\min_j \gamma(X_i, X_j) = \gamma(X_i, X_{i'}) \geq \|X_i - X_{i'}\| \geq \min_j \|X_i - X_j\|$ and finally $\max_i \min_{j\neq i} \gamma(X_i, X_j) \geq \max_i \min_{j\neq i} \|X_i - X_j\|$. Finally from (9) we have $\max_i \min_{j\neq i} \|X_j - X_i\| \xrightarrow{a.s.} 0$, which concludes the proof. □

3.2 The case of noisy data: the "parallel" model

The following two theorems are meaningful in at least two ways. On the one hand, if we know the amount of noise (R_1 in the notation introduced before), these results can be used to detect whether or not the support $\mathcal{M}$ of the original sample is full dimensional (see (11) and (15)).

On the other hand, in the lower dimensional setting, they give an easy-to-implement way to estimate R_1 (see (10) and (14)).

Observe that when $\mathring{\mathcal{M}} = \emptyset$, then $R_1 = \max_{x \in S} d(x, \partial S)$. If $\widehat{\partial S}_n$ denotes a consistent estimator of $\partial B(\mathcal{M}, R_1)$, a natural plug-in estimator for R_1 is $\max_{Y_i \in \mathcal{Y}_n} d(\mathcal{Y}_n, \widehat{\partial S}_n)$.

In Theorem 3 $\widehat{\partial S}_n$ this estimator is constructed in terms of the set of the centers of the boundary balls, while in Theorem 4 we use the boundary of the r-convex hull. The second theorem is stronger than the first one in several aspects: the parameter choice is easier and the convergence rate is better (and does not depend on the parameter). The price to pay is computational since the corresponding statistic is much more difficult to implement; see Section 6.

Theorem 3. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set such that* $\text{reach}(\mathcal{M}) = R_0 > 0$. *Let $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ be an iid sample of a distribution P_Y with support $S = B(\mathcal{M}, R_1)$ with $0 < R_1 < R_0$, absolutely continuous with respect to the Lebesgue measure, whose density f, is bounded from below by $f_0 > 0$. Let $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$, let us denote $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$ where $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$.*

i) if $\mathring{\mathcal{M}} = \emptyset$, then, with probability one,

$$\left| \hat{R}_n - R_1 \right| \leq 2\varepsilon_n \quad \textit{for } n \textit{ large enough,} \tag{10}$$

ii) if $\mathring{\mathcal{M}} \neq \emptyset$, then there exists $C > 0$ such that, with probability one

$$|\hat{R}_n - R_1| > C \quad \textit{for } n \textit{ large enough.} \tag{11}$$

Proof. *i)* Observe, that, since $\mathring{\mathcal{M}} = \emptyset$, $R_1 = \max_{x \in S} d(x, \partial S)$.

From Corollary 4.9 in Federer (1959), $\text{reach}(S) \geq R_0 - R_1 > 0$ and $\text{reach}(\overline{S^c}) \geq R_1$. A first consequence of the positive reach of S is that it has a Lebesgue null boundary and thus, with probability one for all i, $Y_i \in \mathring{S}$ and then, with probability one

$$\hat{S}_n(\varepsilon_n) \subset B(\mathring{S}, \varepsilon_n). \tag{12}$$

Since $\text{reach}(\overline{S^c}) \geq R_1$, by Proposition 1 S is standard with respect to P_X for any constant $\delta < f_0/3$ (see Definition 4).

Then according to Proposition 1 and Theorem 4 in Cuevas and Rodriguez-Casal (2004) to conclude that for large enough n, with probability one,

$$S \subset \hat{S}_n(\varepsilon_n) \tag{13}$$

Now, for all $x \in S$ let us consider $z \in \partial S$ a point such that $||x - z|| = d(x, \partial S)$ and $t = z + \varepsilon_n \eta$ where $\eta = \eta(z)$ is a normal vector to ∂S at z that points outside S (η can be defined according to Definition 4.4 and Theorem 4.8 (12) in Federer (1959)). Notice that the metric projection of t on S is y thus $d(t, S) = \varepsilon_n$ so, according to (12), with probability one $t \notin \hat{S}_n(\varepsilon_n)$. The point z belongs to S so,

by (13), with probability one for n large enough $z \in \hat{S}_n(\varepsilon_n)$. We thus conclude $[t,z] \cap \partial\hat{S}_n(\varepsilon_n) \neq \emptyset$, with probability one, for n large enough. Let then consider $y \in [t,z] \cap \partial\hat{S}_n(\varepsilon_n)$, as $y \in \partial\hat{S}_n(\varepsilon_n)$ there exists $i \in I_{bb}$ such that $y \in \partial\mathcal{B}(Y_i, \varepsilon_n)$ and, as $y \in [t,z]$, $||y-z|| \leq \varepsilon_n$ thus $\|x-Y_i\| \leq \|x-z\|+\|z-y\|+\|y-Y_i\| \leq d(x,\partial S)+2\varepsilon_n$. To sumurize we just have proved that: for all $x \in S$ there exits $i \in I_{bb}$ such that $\|x - Y_i\| \leq d(x,\partial S) + 2\varepsilon_n$ thus for all $x \in S : \min_{i\in I_{bb}} \|x - Y_i\| \leq d(x,\partial S) + 2\varepsilon_n$. To conclude $\max_j \min_{i\in I_{bb}} \|Y_j - Y_i\| \leq \max_j d(Y_j,\partial S) + 2\varepsilon_n \leq \max_{x\in S}(d(x,\partial S) + 2\varepsilon_n = R_1 + 2\varepsilon_n$ (with probability one for n large enough).

The reverse inequality is easier to prove, let us consider $x_0 \in S$ such that $d(x_0,\partial S) = R_1$, notice that, by (13) (with probability one for n large enough) there exists i_0 such that $\|x_0 - Y_{i_0}\| \leq \varepsilon_n$. By triangular inequality $\mathcal{B}(Y_{i_0}, R_1 - \varepsilon_n) \subset S$ and by (13) we also have $\mathcal{B}(Y_{i_0}, R_1 - \varepsilon_n) \subset \hat{S}_n(\varepsilon_n)$ thus $\min_{i\in I_{bb}}\{\|Y_{i_0} - Y_i\|\} \geq R_1 - 2\varepsilon_n$. Then we have proved $\max_j \min_{i\in I_{bb}}\{\|Y_i - Y_j\|\} \geq R_1 - 2\varepsilon_n$. This concludes the proof of (10).

ii) Observe that to prove *i)* we proved that $|\hat{R}_n - \max_{x\in S} d(x,\partial S)| < 2\varepsilon_n$. Then, with probability one, for n large enough, $|\hat{R}_n - R_1| > |c_1 - R_1|/2 = C > 0$, where $c_1 = \max_{x\in\partial S} d(x,\partial S)$.

□

Theorem 4. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set such that* $\mathrm{reach}(\mathcal{M}) = R_0 > 0$. *Suppose that the sample $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ has a distribution with support $S = B(\mathcal{M}, R_1)$ for some $R_1 < R_0$ with a density bounded from below by a constant $f_0 > 0$. Let us denote $\tilde{R}_n = \max_i d(Y_i, \partial C_r(\mathcal{Y}_n))$ where $C_r(\mathcal{Y}_n)$ denotes the r-convex hull of the sample, as defined in* (2) *for $r \leq \min(R_1, R_0 - R_1)$.*

i) If $\mathring{\mathcal{M}} = \emptyset$ and, for some $d' < d$, $\mathcal{M}$ has a finite, strictly positive d'-dimensional Minkowski content, then, with probability one,

$$\left|\tilde{R}_n - R_1\right| = \mathcal{O}(\log(n)/n)^{\min(1/(d-d'),2/(d+1))}, \tag{14}$$

ii) if $\mathring{\mathcal{M}} \neq \emptyset$, then there exists $C > 0$ such that, with probability one

$$|\tilde{R}_n - R_1| > C \quad \textit{for } n \textit{ large enough.} \tag{15}$$

Proof. Again, as shown in the proof of Theorem 3, $\mathrm{reach}(B(\mathcal{M}, R_1)) \geq \mathrm{reach}(\mathcal{M}) - R_1 = R_0 - R_1$; also $\mathrm{reach}(\overline{B(\mathcal{M}, R_1)^c}) \geq R_1$. Hence, according to Proposition 1 in Cuevas and Rodriguez-Casal (2004), $B(\mathcal{M}, R_1)$ and $\overline{B(\mathcal{M}, R_1)^c}$ are both r-convex for $r = \min(R_1, R_0 - R_1) > 0$. Note, in addition, that by construction of $S = B(\mathcal{M}, R_1)$ we have that $\mathring{S}_i \neq \emptyset$ for every path-connected component $S_i \subset S$. So, we can use Theorem 3 in Rodríguez-Casal (2007) to conclude

$$d_H\big(\partial C_r(\mathcal{Y}_n), \partial S\big) = \mathcal{O}\big((\log(n)/n)^{2/(d+1)}\big), \text{ a.s.} \tag{16}$$

Let us prove that, with probability one, for n large enough,

$$B\big(\mathcal{M}, R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S)\big) \subset C_r(\mathcal{Y}_n). \tag{17}$$

Proceeding by contradiction, let $x \in B\big(\mathcal{M}, R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S)\big)$ such that $x \notin C_r(\mathcal{Y}_n)$, let y be the projection of x onto $\mathcal{M}$. It is easy to see that, for n large enough, with probability one, $\mathcal{M} \subset C_r(\mathcal{Y}_n)$ then $y \in C_r(\mathcal{Y}_n)$. Observe that, by Corollary 4.9 in Federer (1959)

$$B\big(\partial S, d_H(\partial C_r(\mathcal{Y}_n), \partial S)\big) = B\big(\mathcal{M}, R_1 + d_H(\partial C_r(\mathcal{Y}_n), \partial S)\big) \setminus B\big(\mathcal{M}, R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S)\big),$$

then, the segment there exists $z \in \partial C_r(\mathcal{Y}_n) \cap (x, y)$, (x, y) being the open segment joining x and y, but then by (23), $d(z, \partial S) > d_H(\partial C_r(\mathcal{Y}_n), \partial S)$ which is a contradiction, that concludes the proof of (17).

First we prove *i*). Suppose now that $\mathring{\mathcal{M}} = \emptyset$. Then $R_1 = \max_{x \in S} d(x, \partial S) = \max_{x \in \mathcal{M}} d(x, \partial S) = d_H(\mathcal{M}, \partial S)$. Also, as $C_r(\mathcal{Y}_n) \subset S$ thus

$$\tilde{R}_n \leq R_1. \tag{18}$$

Now for all observation Y_i let m_i denotes its projection on $\mathcal{M}$ by (17) we have $d(m_i, \partial C_r(\mathcal{Y}_n)) \geq R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S)$ so that, with triangular inequality $d(m_i, Y_i) + d(Y_i, \partial C_r(\mathcal{Y}_n)) \geq R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S)$. Thus

$$\tilde{R}_n \geq R_1 - d_H(\partial C_r(\mathcal{Y}_n), \partial S) - \min_i d(Y_i, \mathcal{M}) \tag{19}$$

From the assumption of finiteness of the Minkowski content of $\mathcal{M}$, given a constant $A > 0$ there exists a constant $c_{\mathcal{M}}$ such that for n large enough,

$$\mu_d\left(\mathcal{B}(\mathcal{M}, (A\log(n)/n)^{1/(d-d')})\right) \geq c_{\mathcal{M}} A(\log(n)/n).$$

Thus,

$$P_X\big(\forall i, d(Y_i, \mathcal{M}) \geq (A\log(n)/n)^{1/(d-d')}\big) \leq \big(1 - f_0 c_{\mathcal{M}} A(\log(n)/n)\big)^n \leq n^{-f_0 c_{\mathcal{M}} A}$$

If we take $A > 1/(f_0 c_{\mathcal{M}})$ we obtain, from Borel-Cantelli lemma,

$$\min_i d(Y_i, \mathcal{M}) = \mathcal{O}\left((\log(n)/n)^{1/(d-d')}\right), \text{ a.s.} \tag{20}$$

Finally, (14) is a direct consequence of (16), (18), (19) and (20).

The proof of *ii*) is obtained as in Theorem 3 part *ii*)

□

Remark 2. *The assumption imposed on $\mathcal{M}$ in part (i) can be seen as an statement of d'-dimensionality. For example if we assume that $\mathcal{M}$ is rectifiable then, from Theorem 3.2.39 in Federer (1969), the d'-dimensional Hausdorff measure of $\mathcal{M}$, $\mathcal{H}^{d'}(\mathcal{M})$ coincides with the corresponding Minkowski content. Hence $0 < \mathcal{H}^{d'}(\mathcal{M}) < \infty$ and, according to expression (4), this entails $dim_H(\mathcal{M}) = d'$.*

3.3 An index of closeness to lower dimensionality

According to Theorem 3 in the case $R_1 = 0$, the value $2\hat{R}_n/\widehat{\text{diam}}(\mathcal{M})$ (where $\widehat{\text{diam}}(\mathcal{M}) = \max_{i\neq j} \|X_i - X_j\|$) can be seen as an index of departure from low-dimensionality. Observe that if $\mathcal{M} = \overline{\mathring{\mathcal{M}}}$ we get $2\hat{R}_n/\widehat{\text{diam}}(\mathcal{M}) \to 1$, a.s. and if $\mathcal{M}$ has empty interior, $2\hat{R}_n/\widehat{\text{diam}}(\mathcal{M}) \to 0$ a.s.

4 An algorithm to partially de-noise the sample

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set with $\text{reach}(\mathcal{M}) = R_0 > 0$. Let $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ be an iid sample of Y, with support $S = B(\mathcal{M}, R_1)$ for some $0 < R_1 < R_0$, and distribution P_Y, absolutely continuous with respect to the Lebesgue measure, whose density f, is bounded from below. We now propose an algorithm to get from $\mathcal{Y}_n$, a "partially de-noised" sample of points that allow us to estimate the target set $\mathcal{M}$.

The procedure works as follows:

1. *Take suitable auxiliary estimators for S and R_1.* Let $\hat{S}_n$ be an estimator of S (based on $\mathcal{Y}_n$) such that $d_H(\partial\hat{S}_n, \partial S) < a_n$ eventually a.s., for some $a_n \to 0$. Let $\hat{R}_n$ be an estimator of R_1 such that $|\hat{R}_n - R_1| \leq e_n$ eventually a.s. for some $e_n \to 0$.

2. *Select a λ-subsample far from the estimated boundary of S.* Take $\lambda \in (0,1)$ and define $\mathcal{Y}_m^\lambda = \{Y_1^\lambda, \ldots, Y_m^\lambda\} \subset \mathcal{Y}_n$ where $Y_i^\lambda \in \mathcal{Y}_m^\lambda$ if and only if $d(Y_i^\lambda, \partial\hat{S}_n) > \lambda\hat{R}_n$.

3. *The projection + translation stage.* For every $Y_i^\lambda \in \mathcal{Y}_m^\lambda$, we define $\{Z_1, \ldots, Z_m\} = \mathcal{Z}_m$ as follows,
$$Z_i = \pi_{\partial\hat{S}_n}(Y_i^\lambda) + \hat{R}_n \frac{Y_i^\lambda - \pi_{\partial\hat{S}_n}(Y_i^\lambda)}{\|Y_i^\lambda - \pi_{\partial\hat{S}_n}(Y_i^\lambda)\|}, \tag{21}$$
being $\pi_{\partial\hat{S}_n}(Y_i^\lambda)$ the metric projection of Y_i^λ on $\partial\hat{S}_n$.

The following result shows that the above de-noising procedure allows us to recover the "inner set" $\mathcal{M}$.

Theorem 5. *With the notation introduced before, there exists $b_n = \mathcal{O}\left(\max(a_n^{1/3}, e_n, \varepsilon_n)\right)$ such that, with probability one, for n large enough,*
$$d_H(\mathcal{Z}_m, \mathcal{M}) \leq b_n$$
where $\varepsilon_n = c(\log(n)/n)^{1/d}$ with $c > (4/\omega_d)^{1/d}$.

Proof. First let us observe that $d_H(\mathcal{Y}_n, S) \leq \varepsilon_n$ eventually a.s.. Let us fix $Y_i^\lambda \in \mathcal{Y}_m^\lambda$.

Let us denote $l = \|Y_i^\lambda - \pi_{\partial S}(Y_i^\lambda)\|$ and $\eta_i = (Y_i^\lambda - \pi_{\partial S}(Y_i^\lambda))/l$, let us introduce two estimators $\hat{l} = \|Y_i^\lambda - \pi_{\partial\hat{S}_n}(Y_i^\lambda)\|$ and $\hat{\eta}_i = (Y_i^\lambda - \pi_{\partial\hat{S}_n}(Y_i^\lambda))/\hat{l}$. With this notation $Z_i = \pi_{\partial\hat{S}_n}(Y_i^\lambda) + \hat{R}_n\hat{\eta}_i$. Recall that since $\text{reach}(\mathcal{M}) > R_1$ we have (by Corollary 4.9 in Federer (1959)) that $\pi_{\mathcal{M}}(Y_i^\lambda) = \pi_{\partial S}(Y_i^\lambda) + R_1\eta_i$,

For all Y_i^λ there exists a point $x \in \partial\hat{S}_n$ with $||x - \pi_{\partial S}(Y_i^\lambda)|| \leq a_n$ so that, by triangular inequality: $d(Y_i^\lambda, \partial\hat{S}_n) \leq l + a_n$ that is,

$$\pi_{\partial\hat{S}_n}(Y_i^\lambda) \in \mathcal{B}(Y_i^\lambda, l + a_n) \tag{22}$$

Now let us prove that

$$\pi_{\partial\hat{S}_n}(Y_i^\lambda) \in \mathcal{B}(Y_i^\lambda, l - a_n)^c \tag{23}$$

Suppose by contradiction that $\pi_{\partial\hat{S}_n}(Y_i^\lambda) \in \mathcal{B}(Y_i^\lambda, l - a_n)$, since $d_H(\partial S_n, \partial S) < a_n$ there exists $t \in \partial S$ such that $\|t - \pi_{\partial\hat{S}_n}(Y_i^\lambda)\| < a_n$, but then $l = d(Y_i^\lambda, \partial S) \leq \|Y_i^\lambda - \pi_{\partial\hat{S}_n}(Y_i^\lambda)\| + \|\pi_{\partial\hat{S}_n}(Y_i^\lambda) - t\| < l$. That concludes the proof of (23).

By (22) and (23) we have:

$$l - a_n \leq \hat{l} \leq l + a_n \tag{24}$$

In the same way it can be proved that

$$\pi_{\partial\hat{S}_n}(Y_i^\lambda) \in \mathcal{B}(\pi_{\mathcal{M}}(Y_i^\lambda), R_1 - a_n)^c. \tag{25}$$

Let us prove that there exists $C_0 > 0$ such that

$$\text{for all } Y_i^\lambda \in \mathcal{Y}_m^\lambda \text{ , } \|Z_i - \pi_{\mathcal{M}}(Y_i^\lambda)\| \leq C_0\sqrt{a_n^{2/3} + a_n^{1/3}e_n + e_n^2} \tag{26}$$

First consider the case $0 \leq R_1 - l \leq a_n^{1/3}$, which implies that $\|Y_i^\lambda - \pi_{\mathcal{M}}(Y_i^\lambda)\| \leq a_n^{1/3}$. Notice that, by (24), $\|Y_i^\lambda - Z_i\| = |\hat{R}_n - \hat{l}| \leq a_n^{1/3} + e_n + a_n$, finally we get

$$\|Z_i - \pi_{\mathcal{M}}(Y_i^\lambda)\| \leq 2a_n^{1/3} + a_n + e_n. \tag{27}$$

Now we consider the case $R_1 - l \geq a_n^{1/3}$, recall that by (22) and (25) we have.

$$\pi_{\partial\hat{S}_n}(Y_i^\lambda) \in \mathcal{B}(Y_i^\lambda, l + a_n) \setminus \mathcal{B}(\pi_{\mathcal{M}}(Y_1^\lambda), R_1 - a_n). \tag{28}$$

In Figure 1 it is represented the case for which $\|\pi_{\partial\hat{S}_n}(Y_i^\lambda) - \pi_{\partial S}(Y_i^\lambda)\|$ takes its largest possible value.

To find an upper bound for such value, let us first note that $\pi_{\partial\hat{S}_n}(Y_i^\lambda) + R_1\hat{\eta}_i$, $\pi_{\partial S}(Y_i^\lambda)$, $\pi_{\mathcal{M}}(Y_i^\lambda)$, Y_i^λ and $\pi_{\partial\hat{S}_n}(Y_i^\lambda)$ are in the same plane Π. Let us now apply a translation T in order to get, $T(\pi_{\partial S}(Y_1^\lambda)) = 0$. Let us consider in Π a coordinate system (x, y) such that $\pi_{\mathcal{M}}(Y_i^\lambda) = (0, -R_1)$.

Let (x_1, y_1) be the coordinates of the point $\pi_{\partial\hat{S}_n}(Y_i^\lambda)$. From (28) we get

$$x_1^2 + (y_1 + l)^2 \leq (l + a_n)^2 \tag{29}$$

$$x_1^2 + (y_1 + R_1)^2 \geq (R_1 - a_n)^2 \tag{30}$$

If we multiply (30) by $-l$, we get $-l(x_1^2 + y_1^2) - 2y_1lR_1 \leq -la_n^2 + 2a_nlR_1$ and if we multiply (29) by R_1 we get $R_1(x_1^2 + y_1^2) + 2y_1lR_1 \leq 2R_1a_nl + a_n^2R_1$. Then, if sum this two equations we get,

$$x_1^2 + y_1^2 \leq \frac{4lR_1}{R_1 - l}a_n + a_n^2 \leq 4R_1^2a_n^{2/3} + a_n^2 \tag{31}$$

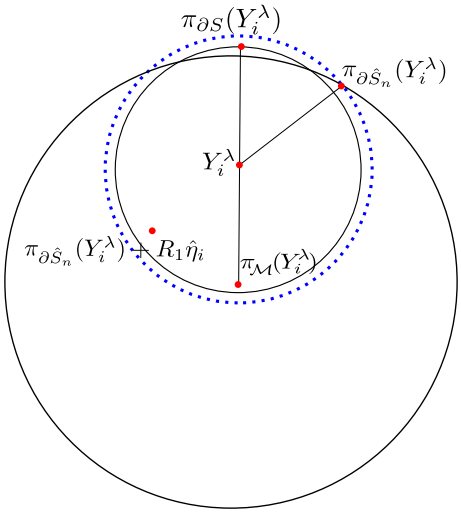

Figure 1: In solid black line $\mathcal{B}(Y_i^\lambda, l)$ and $\mathcal{B}(\pi_{\mathcal{M}}(Y_1^\lambda), R_1 - a_n)$, in dashed line $\mathcal{B}(Y_i^\lambda, \hat{l})$

Notice that $Z_i \in \Pi$, let us denote (x, y) the coordinates of Z_i in Π, then

$$x = x_1 - \hat{R}_n \frac{x_1}{\|Y_i^\lambda - \pi_{\partial \hat{S}_n}(Y_i^\lambda)\|} = x_1 - \hat{R}_n \frac{x_1}{\hat{l}}$$

and

$$y = y_1 - \hat{R}_n \frac{l + y_1}{\|Y_i^\lambda - \pi_{\partial \hat{S}_n}(Y_i^\lambda)\|} = y_1 - \hat{R}_n \frac{l + y_1}{\hat{l}}.$$

Since the coordinates of $\pi_{\mathcal{M}}(Y_i^\lambda)$ are $(0, -R_1)$ we get that

$$\|Z_i - \pi_{\mathcal{M}}(Y_i^\lambda)\|^2 = \|(x, y + R_1)\|^2 = (x_1^2 + y_1^2)\left(\frac{\hat{l} - \hat{R}_n}{\hat{l}}\right)^2 + \left(R_1 - \hat{R}_n \frac{l}{\hat{l}}\right)^2 + 2y_1 \left|\frac{\hat{l} - \hat{R}_n}{\hat{l}}\right| \left(R_1 - \hat{R}_n \frac{l}{\hat{l}}\right). \quad (32)$$

Observe that $|l - \hat{l}| \leq a_n$ and $|R_1 - \hat{R}_n| < e_n$. For n large enough we can bound $|\frac{\hat{l} - \hat{R}_1}{\hat{l}}| \leq 2$ and $\hat{l} \geq \lambda R_1/2$, then

$$\left|R_1 - \hat{R}_n \frac{l}{\hat{l}}\right| = \left|\frac{R_1(\hat{l} - l) - l(\hat{R}_n - R_1)}{\hat{l}}\right| \leq 2\frac{a_n + e_n}{\lambda}. \quad (33)$$

Finally by equations (31),(32) and (33), if $R_1 - l \geq a_n^{1/3}$, there exists C_0 such that

$$\|Z_i - \pi_{\mathcal{M}}(Y_i^\lambda)\| \leq C_0 \sqrt{a_n^{2/3} + a_n^{1/3} e_n + e_n^2} \quad (34)$$

That concludes the proof of (26).

To conclude the proof of the theorem let us prove that $\mathcal{M} \subset B(\mathcal{Z}_m, a_n + e_n + 2\varepsilon_n)$ eventually, a.s. Recall that $d_H(\mathcal{Y}_n, S) \leq \varepsilon_n$ eventually, a.s. thus for all $x \in \mathcal{M}$, there exists $Y_i \in \mathcal{Y}_n$ such that $\|x - Y_i\| \leq \varepsilon_n$. For n large enough we have $Y_i \in \mathcal{Y}_m^{\lambda}$. Following the same ideas used to prove (27) we obtain $\|Z_i - Y_i\| \leq \varepsilon_n + a_n + e_n$. By triangular inequality we get

$$\mathcal{M} \subset B(\mathcal{Z}_m, a_n + e_n + 2\varepsilon_n) \text{ eventually, a.s.} \tag{35}$$

Combining (27), (34) and (35) we obtain:

$$d_H(\mathcal{Z}_m, \mathcal{M}) = \mathcal{O}\left(\max(a_n^{1/3}, e_n, \sqrt{a_n^{2/3} + a_n^{1/3} e_n + e_n^2}, \varepsilon_n)\right) = \mathcal{O}\left(\max(a_n^{1/3}, e_n, \varepsilon_n)\right).$$

□

The two following corollaries give the exact convergence rate for the denoising process introduced before, using the centres of the boundary balls (Corollary 1), and the boundary of the r-convex hull (Corollary 2), as estimators of the boundary of the support.

Corollary 1. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set such that* $\operatorname{reach}(\mathcal{M}) = R_0 > 0$. *Let $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ be an iid sample of a distribution P_Y with support $B(\mathcal{M}, R_1)$ for some $0 < R_1 < R_0$. Assume that P_Y is absolutely continuous with respect to the Lebesgue measure and the density f, is bounded from below by a constant $f_0 > 0$. Let $\varepsilon_n = c(\log(n)/n)^{1/d}$ and $c > (4/(f_0\omega_d))^{1/d}$.*

Given $\lambda \in (0,1)$, let $\mathcal{Z}_n$ be the points obtained after the denoising process using $\hat{R}_n$ to estimate R_1 and $\{Y_i, i \in I_{bb}\}$ as an estimator of ∂S where $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$. Then,

$$d_H(\mathcal{Z}_m, \mathcal{M}) = \mathcal{O}\big((\log(n)/n)^{1/(3d)}\big), \text{ a.s.}$$

Using the assumption of r-convexity for $\mathcal{M}$ (see Definitions 2 and 3 and the subsequent comments) in the construction of the set estimator, we can replace $\hat{R}_n$ with $\tilde{R}_n$ (see Theorem 4). Then, at the cost of some additional complexity in the numerical implementation, a faster convergence rate can be obtained. This is made explicit in the following result.

Corollary 2. *Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact d'-dimensional set (in the sense of Theorem 4, i)) such that* $\operatorname{reach}(\mathcal{M}) = R_0 > 0$. *Let $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ be an iid sample of a distribution P_Y with support $B(\mathcal{M}, R_1)$ for some $0 < R_1 < R_0$. Assume that P_Y is absolutely continuous with respect to the Lebesgue measure and the density f, is bounded from below by a constant $f_0 > 0$.*

For a given $\lambda \in (0,1)$, let $\mathcal{Z}_n$ be the set of the points obtained after the denoising process, based on the estimator $\partial C_r(\mathcal{Y}_n)$ of ∂S (for some r with $0 < r < \min(R_0 - R_1, R_1)$) and the estimator $\tilde{R}_n$ of R_1.

Then,

$$d_H(\mathcal{Z}_m, \mathcal{M}) = \mathcal{O}\big((\log(n)/n)^{2/(3(d+1))}\big).$$

5 Estimation of lower-dimensional measures

5.1 Noiseless model

In this section, we go back to the noiseless model, that is, we assume that the sample points $X_1, \ldots, X_n$ are drawn according to a distribution whose support is $\mathcal{M}$. The target is to estimate the d'-dimensional Minkowski content of $\mathcal{M}$, as given by

$$\lim_{\epsilon \to 0} \frac{\mu_d(B(\mathcal{M}, \epsilon))}{\omega_{d-d'} \epsilon^{d-d'}} = L_0(\mathcal{M}) < \infty. \tag{36}$$

This is just (alongside with Hausdorff measure, among others) one of the possible ways to measure lower-dimensional sets; see Mattila (1995) for background.

In recent years, the problem of estimating the d'-dimensional measures of a compact set from a random sample has received some attention in the literature. The simplest situation corresponds to the full-dimensional case $d' = d$. Any estimator $\mathcal{M}_n$ of $\mathcal{M}$ consistent with respect to the *distance in measure*, that is $\mu_d(\mathcal{M}_n \Delta \mathcal{M}) \to 0$ (in prob. or a.s., where Δ stands for the symmetric difference), will provide a consistent estimator for $\mu_d(\mathcal{M})$. In fact, as a consequence of Th. 1 in Devroye and Wise (1980) (recall that S is compact here) this will the always the case (in probability) when $\mathcal{M}_n$ is the *offset* estimator (6), provided that μ_d is absolutely continuous (on $\mathcal{M}$) with respect to P_X together with $r_n \to 0$ and $nr_n^d \to \infty$.

Other more specific estimators of $\mu_d(\mathcal{M})$ can be obtained by imposing some shape assumptions on $\mathcal{M}$, such as convexity or r-convexity, which are incorporated to the estimator $\mathcal{M}_n$; see Arias-Castro et al. (2016), Baldin and Reiss (2015), Pardon (2011).

Regarding the estimation of lower-dimensional measures, with $d' < d$, the available literature mostly concerns the problem of estimating $L_0(\mathcal{M})$, $\mathcal{M}$ being the boundary of some compact support S. The sample model is also a bit different, as it is assumed that we have sample points *inside and outside* S. Here, typically, $d' = d-1$; see, Armendáriz et. al (2009), Cuevas et al. (2007), Cuevas *et al.* (2013), Jiménez and Yukich (2011).

Again, in the case $\mathcal{M} = \partial S$ with $d = 2$, under the extra assumption of r-convexity for S, the consistency of the plug-in estimator $L_0(\partial C_r(\mathcal{X}_n))$ of $L_0(\partial S)$ is proved in Cuevas, Fraiman and Pateiro-López (2012) under the usual *inside* model (points taken on S). Finally, in Berrendero et al. (2014), assuming an *outside* model (points drawn in $B(S, R) \setminus S$), estimators of $\mu_d(S)$ and $L_0(\partial S)$ are proposed, under the condition of *polynomial volume* for S

From the perspective of the above references, our contribution here (Th. 6 below) could be seen as a sort of lower-dimensional extension of the mentioned results of type $\mu_d(\mathcal{M}_n) \to \mu_d(\mathcal{M})$ regarding volume estimation. But, obviously, in this case the Lebesgue measure μ_d must be replaced with a lower-dimensional counterpart, such as the Minkowski content (36). We will also need the following lower-dimensional version of the standardness property given in Definition 3.

Definition 7. *A Borel probability measure P_X defined on a d'-dimensional set $\mathcal{M} \subset \mathbb{R}^d$ (considered with the topology induced by $\mathbb{R}^d$) is said to be standard with respect to the*

d'-dimensional Lebesgue measure $\mu_{d'}$ if there exist λ and δ such that, for all $x \in \mathcal{M}$,

$$P_X(\mathcal{B}(x,r)) = \mathbb{P}(X \in \mathcal{B}(x,r) \cap \mathcal{M}) \geq \delta \mu_{d'}(\mathcal{B}(x,r)), \text{ for } 0 \leq r \leq \lambda.$$

Remark 3. *Observe that, by Lemma 5.3 in (Niyogi, Smale and Weinberger (2008)) this condition is fulfilled if P_X has a density f bounded from below and $\mathcal{M}$ is a manifold with positive condition number (also known as positive reach). Standardness of the distribution has also been used in Chazal et al. (2013), Aamari and Levrard (2015).*

Theorem 6. *Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be an iid sample drawn according to a distribution P_X on a set $\mathcal{M} \subset \mathbb{R}^d$. Let us assume that the distribution P_X is standard with respect to the d'-dimensional Lebesgue measure (see 7) and that there exists the d' Minkowski content $L_0(\mathcal{M})$ of $\mathcal{M}$, given by (36). Let us take r_n such that $r_n \to 0$ and $(\log(n)/n)^{1/d'} = o(r_n)$, then*

(a)

$$\lim_{n\to\infty} \frac{\mu_d\big(B(\mathcal{X}_n, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) \quad a.s.. \tag{37}$$

(b) If $\operatorname{reach}(M) = R_0 > 0$*, then*

$$\frac{\mu\big(B(\mathcal{X}_n, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) = \mathcal{O}\Big(\frac{\beta_n}{r_n} + r_n\Big),$$

where $\beta_n := d_H(\mathcal{X}_n, \mathcal{M}) = \mathcal{O}\big(\log(n)/n\big)^{1/d'}$.

Proof. (a) First we will see that, following the same ideas as in Theorem 3 in Cuevas and Rodriguez-Casal (2004) it can be readily proved that

$$d_H(\mathcal{X}_n, \mathcal{M}) = \mathcal{O}\big(\log(n)/n\big)^{1/d'}. \tag{38}$$

In order to see (38), let us consider M_Δ a minimal covering of $\mathcal{M}$, with balls of radius Δ centred in N_Δ points belonging to $\mathcal{M}$. Let us prove that $N_\Delta = \mathcal{O}(\Delta^{-d'})$. Indeed, since M_Δ is a minimal covering it is clear that $\mu_d(B(\mathcal{M}, \Delta)) \geq N_\Delta \omega_d (\Delta/2)^d$, and then

$$\frac{\mu_d(B(\mathcal{M}, \Delta))}{\omega_{d-d'}\Delta^{d-d'}} \geq \frac{N_\Delta \omega_d (\Delta/2)^d}{\omega_{d-d'}\Delta^{d-d'}} = c_1 N_\Delta \Delta^{d'},$$

being c_1 a positive constant. Since there exists $L_0(\mathcal{M})$ it follows that $N_\Delta = \mathcal{O}(\Delta^{-d'})$. Then the proof of (38) follows easily from the standardness of P_X and $N_\Delta = \mathcal{O}(\Delta^{-d'})$, so we will omit it.

Now, in order to prove (37), let us first prove that, if we take $\alpha_n = 1 - \beta_n/r_n$,

$$B(\mathcal{M}, \alpha_n r_n) \subset B(\mathcal{X}_n, r_n) \subset B(\mathcal{M}, r_n) \quad a.s.. \tag{39}$$

To prove this, consider $x_n \in B(\mathcal{M}, \alpha_n r_n)$, then there exists $t_n \in \mathcal{M}$ such that $x_n \in \mathcal{B}(t_n, \alpha_n r_n)$. Since $\beta_n = d_H(\mathcal{X}_n, \mathcal{M})$ there exists $y_n \in \mathcal{B}(t_n, \beta_n)$, $y_n \in \mathcal{X}_n$. It is enough to prove that $x_n \in \mathcal{B}(y_n, r_n)$. But this follows from the fact that, eventually a.s.,

$$\|y_n - x_n\| \leq \|x_n - t_n\| + \|t_n - y_n\| \leq \alpha_n r_n + \beta_n = r_n.$$

Then, from (39)

$$\alpha_n^{d-d'} \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{w_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) \leq \frac{\mu\big(B(\mathcal{X}_n, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) \leq \frac{\mu\big(B(\mathcal{M}, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}). \quad (40)$$

Since there exists $L_0(\mathcal{M})$, the right hand side of (40) goes to zero. To prove that the left hand side of (40) goes to zero, let us observe that, as $\alpha_n = 1 - \beta_n/r_n$, and $\alpha_n^{d-d'} = 1 - \mathcal{O}(\beta_n/r_n)$, then

$$\alpha_n^{d-d'} \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{\omega_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) = \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{\omega_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - \mathcal{O}(\beta_n/r_n) - L_0(\mathcal{M}), \quad (41)$$

since $\alpha_n \to 1$ and $\beta_n/r_n \to 0$ we get

$$\lim_{n\to\infty} \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) \quad a.s..$$

(b) Since $\text{reach}(\mathcal{M}) = R_0$ we get that $\mu\big(B(\mathcal{M}, r_n)\big) = P_d(r_n)$ being $P_d(x)$ a polynomial of degree at most d for $0 < x < R_0$ whose coefficient to the $d - d'$ term is $\omega_{d-d'} L_0(\mathcal{M})$, then

$$\frac{\mu(B(\mathcal{M}, r_n))}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) + r_n A(\mathcal{M}) + o(r_n),$$

for some constant $A(\mathcal{M})$. Now the proof follows from (40) and (41). □

Remark 4. *In the case of sets with positive reach, part (b) suggests to take* $r_n = \sqrt{\max_i \min_{j\neq i} \|X_i - X_j\|}$ *since we know by Theorem 1 in Penrose (1999) that* $r_n^2 = \mathcal{O}\big((\log(n)/n)^{1/d}\big)$ *that gives the optimal convergence rate.*

5.2 Noisy Model

The estimation of the Minkowski content in the noisy model has been tackled in Berrendero et al. (2014), where the random sample is assumed to have uniform distribution in the parallel set U. In this section we will see that even if the sample is not uniformly distributed on $B(\mathcal{M}, R_1)$ for some $0 < R_1 < R_0 = \text{reach}(\mathcal{M})$, it is still possible, by applying first the de-noising algorithm introduced in Section 4, to estimate $L_0(\mathcal{M})$. Following the notation in Section 4, let $\mathcal{Y}_n$ be an iid sample of a random variable Y with support $B(\mathcal{M}, R_1)$, let us denote $\mathcal{Z}_m$ the de-noised sample defined by 21. The estimator is defined as in 37 but replacing $\mathcal{X}_n$ with $\mathcal{Z}_m$. Although the subset $\mathcal{Z}_m$ is not an iid sample (since the random variables Z_i are not independent), the consistency is based on the fact that $\mathcal{Z}_m$ converge in Hausdorff distance to $\mathcal{M}$, as we will prove in the following theorem.

Theorem 7. *With the hypothesis and notation of Theorem 5, if* $\max(a_n^{1/3}, e_n, \varepsilon_n) = o(r_n)$ *where* $\varepsilon_n = c(\log(n)/n)^{1/d}$ *with* $c > (4/\omega_d)^{1/d}$. *Then,*

$$\lim_{n\to\infty} \frac{\mu_d(B(\mathcal{Z}_m, r_n))}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) \quad a.s.. \tag{42}$$

Proof. The proof is analogous to the one in Theorem 6. Observe that in Theorem 5 we proved that $d_H(\mathcal{Z}_m, \mathcal{M}) \leq b_n$, for some $b_n = \mathcal{O}(\max(a_n^{1/3}, e_n, \varepsilon_n))$, then $b_n/r_n \to 0$. As we did Theorem 6 if we take $\alpha_n = 1 - b_n/r_n$, then, with probability one,

$$B(\mathcal{M}, \alpha_n r_n) \subset B(\mathcal{Z}_m, r_n) \subset B(\mathcal{M}, r_n),$$

then we get

$$\alpha_n^{d-d'} \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{w_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) \leq \frac{\mu\big(B(\mathcal{Z}_n, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) \leq \frac{\mu\big(B(\mathcal{M}, r_n)\big)}{\omega_{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}),$$

from where it follows

$$\alpha_n^{d-d'} \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{\omega_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - L_0(\mathcal{M}) = \frac{\mu\big(B(\mathcal{M}, \alpha_n r_n)\big)}{\omega_{d-d'} \alpha_n^{d-d'} r_n^{d-d'}} - \mathcal{O}(b_n/r_n) - L_0(\mathcal{M}).$$

Since $\alpha_n \to 1$ and $b_n/r_n \to 0$ we get 42. □

6 Computational aspects and simulations

We discuss here some theoretical and practical aspects regarding the implementation of the algorithms. We present also some simulations and numerical examples.

6.1 Identifying the boundary balls

The cornerstone of the practical use of Theorem 1 is the effective identification of the boundary balls. The following proposition provides the basis for such identification, in terms of the Voronoi cells of the sample points. Recall that, given a finite set $\{x_1, \ldots, x_n\}$, the Voronoi cell associated with the point x_i is defined by $\mathrm{Vor}(x_i) = \{x : d(x, x_i) \leq d(x, x_j) \text{ for all } i \neq j\}$.

Proposition 2. *Let* $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ *be an iid sample of points, in* $\mathbb{R}^d$, *drawn according to a distribution* P_X, *absolutely continuous with respect to the Lebesgue measure. Then, with probability one, for all* $i = 1, \ldots, n$ *and all* $r > 0$, $\sup\{\|z - X_i\|, z \in \mathrm{Vor}(X_i)\} \geq r$ *if and only if* $\mathcal{B}(X_i, r)$ *is a boundary ball for the Devroye-Wise estimator (6).*

Proof. Let us take $r > 0$ and X_i such that there exists $z \in \partial\mathcal{B}(X_i, r) \cap \mathrm{Vor}(X_i) \neq \emptyset$, let us prove that $z \in \partial\hat{S}_n(r)$. Observe that since $z \in \mathrm{Vor}(X_i)$, $d(z, \mathcal{X}_n \setminus X_i) \geq r$ thus

$d(z, \mathcal{X}_n) = r$. Reasoning by contradiction suppose that $z \in \mathring{\hat{S}}_n$ then, with probability one, there exists j_0 such that $z \in \mathring{\mathcal{B}}(X_{j_0}, r)$ and so $\|z - X_{j_0}\| < r$ that is a contradiction.

Now to prove the converse implication let us assume that $\mathcal{B}(X_i, r)$ is a boundary ball, then there exists $z \in \partial\mathcal{B}(X_i, r)$ such that $z \in \partial\hat{S}_n(r)$. Let us prove that $d(z, \mathcal{X}_n \setminus X_i) \geq r$ (from where it follows that $z \in \mathrm{Vor}(X_i)$). Suppose that $d(z, \mathcal{X}_n \setminus X_i) < r$, then there exists $X_j \neq X_i$ such that $d(z, X_j) < r$ and then $\mathcal{B}\big(z, r - d(z, X_j)\big) \subset \mathring{\hat{S}}_n(r)$. □

6.2 An algorithm to detect empty interior in the noiseless case using Theorem 1

In order to use in practice Theorem 1 to detect lower-dimensionality in the noiseless case, we need to fix a sequence $r_n \downarrow 0$ under the conditions indicated in Theorem 1 (b) and (b). Note that this requires to assume lower bounds for the "thickness" constant $\rho(\mathcal{M}) = \sup d(x, \partial\mathcal{M})$ and the standardness constant δ (which quantifies the sharpness order of $\mathcal{M}$) as well as an upper bound for the radius of the outer rolling ball.

Now, according to Theorem 1, and Proposition 2, we will use the following algorithm.

1) For $i = 1, \ldots, n$, let $V^i = \{V_1^i, \ldots V_{k_i}^i\}$ the vertices of $\mathrm{Vor}(X_i)$,

2) Let $r_i = \sup\{\|z - X_i\|, z \in \mathrm{Vor}(X_i)\} = \max\{\|X_i - V_k^i\|, 1 \leq k \leq k_i\}$, since $\mathrm{Vor}(X_i)$ is a convex polyhedron. Define $r_0 = \min_i r_i$,

3) Decide $\mathring{\mathcal{M}} \neq \emptyset$ if and only if $r_0 \geq r_n$.

6.3 On the estimation of the maximum distance to the boundary

Theorems 3 and 4, involve the calculation of quantities such as $d(x, \partial\hat{S}_n(\epsilon_n))$ and $d(x, \partial C_r(\mathcal{Y}_n))$, where $\hat{S}_n(\epsilon_n)$ is a Devroye-Wise estimator of type (6) and $C_r(\mathcal{Y}_n)$ is the r-convex hull (2) of $\mathcal{Y}_n$.

It is somewhat surprising to note that, in spite of the much simpler structure of $\hat{S}_n(\epsilon_n)$ when compared to $C_r(\mathcal{Y}_n)$, the distance to the boundary $d(x, \partial C_r(\mathcal{Y}_n))$ can be calculated in a simpler, more accurate way than the analogous quantity $d(x, \partial\hat{S}_n(\epsilon_n))$ for the Devroye-Wise estimator $\hat{S}_n(\epsilon_n)$).

Indeed note that $d(x, \partial C_r(\mathcal{Y}_n))$ is relatively simple to calculate; this is done in Berrendero, Cuevas and Pateiro-López (2012) in the two-dimensional case although can be in fact used in any dimension. Observe first that $\partial C_r(\mathcal{Y}_n))$ is included in a finite union of spheres of radius r, with centres in $Z = \{z_1, \ldots, z_m\}$. Then $d(x, \partial C_r(\mathcal{Y}_n)) = \min_{z_i \in Z} \|x - z_i\| - r$. In order to find Z we need to compute the Delaunay triangulation. Recall that the Delaunay triangulation, $\mathrm{Del}(\mathcal{Y}_n)$, is defined as follows. Let $\tau \subset \mathcal{Y}_n$,

$$\tau \in \mathrm{Del}(\mathcal{Y}_n) \quad \text{if and only if} \quad \bigcap_{Y_i \in \tau} \mathrm{Vor}(Y_i) \neq \emptyset.$$

Observe finally, for any dimension, $\bigcap_{Y_i \in \tau} \mathrm{Vor}(Y_i) \neq \emptyset$ is a segment or a half line. If τ_i is the d-dimensional simplex with vertices $\{Y_{i_1}, \ldots, Y_{i_d}\} \subset \partial\mathcal{B}(z_i, r)$, the point z_i can be obtained as $\bigcap_{Y_j^i \in \tau_i} \mathrm{Vor}(Y_i) \cap \mathcal{B}(Y_1^i, r)$.

6.4 Experiments

The general aim of these experiments is not to make an extensive, systematic empirical study. We are just trying to show that the methods and algorithm proposed here can be implemented in practice.

Detection of full dimensionality

As a simple numerical illustration we consider here the noisy model of Subsection 3.2. In each case, we draw 200 samples of sizes $n = 50$, 100, 200, 300, 400, 500, 1000, 2000, 5000, 10000 on the A-parallel set around the unit sphere; that is, the sample data are selected on $B(0, 1 + A) \setminus B(0, 1 - A)$. The width parameter A takes the values $A = 0, 0.01, 0.05, 0.1, \ldots, 0.05$. Table 1 provides the minimum sample sizes to "safely decide" the correct answer. This means to correctly decide on, at least 190 out of 200 considered samples, that the support is lower dimensional (in the case $A = 0$) or that it is full dimensional (cases with $A > 0$).

We have used the boundary balls procedure (here and in the denoising experiment below for $A = 0$) with $r = 2 \max_i(\min_{j \neq i} \|X_j - X_i\|)$.

The results look quite reasonable: the larger the dimension d and the smaller the width parameter A, the harder the detection problem.

A	$d = 2$	$d = 3$	$d = 4$
0	≤ 50	≤ 50	≤ 50
0.01	$[51, 100]$	$[1001, 2000]$	> 10000
0.05	≤ 50	$[201, 300]$	$[1001, 2000]$
0.1	≤ 50	$[51, 100]$	$[101, 200]$
0.2	≤ 50	≤ 50	$[51, 100]$
0.3	≤ 50	≤ 50	$[51, 100]$
0.4	≤ 50	≤ 50	≤ 50
0.5	≤ 50	≤ 50	≤ 50

Table 1: Minimum sample sizes required to detect lower dimensionality for different values of the dimension d and the width parameter A.

Denoising

We draw points on $\mathcal{B}(0, 1.3) \setminus \mathcal{B}(0, 0.7)$ in $\mathbb{R}^2$ and $\mathbb{R}^3$.

In order to evaluate the effectiveness of the denoising procedure we define the random variable $e = \|Y\| - 1$ from the denoised data Y and also from the original data. Note that the "perfect" denoising would correspond to $e = 0$. The Figure 2 shows the kernel estimators of both densities of e for the case $d = 2$ (left panel) and for $d = 3$ (right panel). These estimators for the denoised case are based on $m = 100$ values of e extracted from samples of sizes $n =$100, 1000, 10000. The density estimators for the initial distribution are based on samples of size 100. Clearly, when the denoised sample of size $m = 100$ is based on a very large sample, with $n = 10000$, the denoising process is better, as

suggested by the fact that the corresponding density estimators are strongly concentrated around 0. The slight asymmetry in the three dimensional case, accounts for the fact that the "external" volume $\mathcal{B}(0, 1.3) \setminus \mathcal{B}(0, 1)$ is larger than the "internal" one $\mathcal{B}(0, 1) \setminus \mathcal{B}(0, 0.7)$.

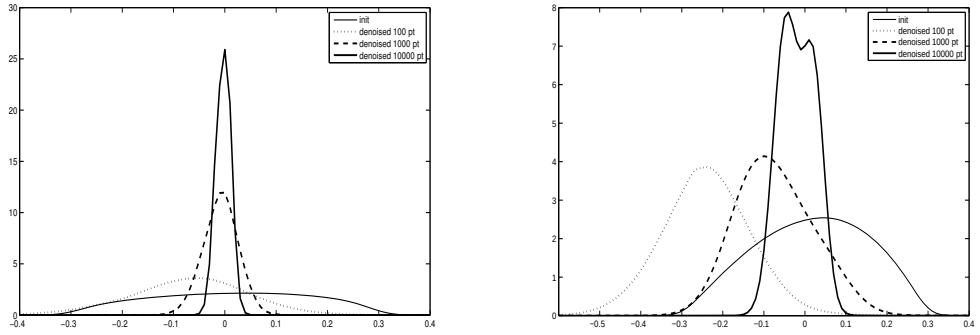

Figure 2: Estimated density functions of the random variable $e = \|Y\| - 1$ for $d = 2$ (left) and $d = 3$ right, with and without denoising.

Figures 3 and 4 provide a more visual idea on the result of the denoising algorithm. They correspond, respectively, to the set $\mathcal{B}(S_{L_3}, 0.3)$ (where $S_{L_3} = \{(x, y), |x|^3 + |y|^3 = 1\}$) and to $\mathcal{B}(T, 0.3)$, where T is the so-called *Trefoil Knot*, a well-known curve with interesting topological and geometric properties.

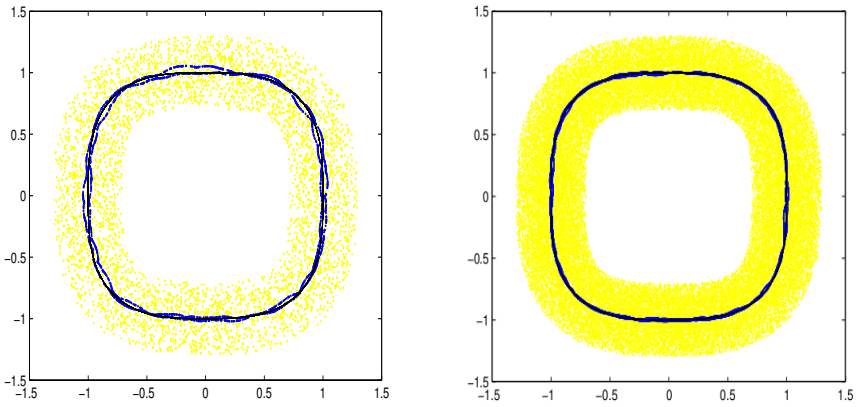

Figure 3: The yellow background is made of 5000 points (left) and 50000 points (right) drawn on on $\mathcal{B}(S_{L_3}, 0.3)$, with $S_{L_3} = \{(x, y), |x|^3 + |y|^3 = 1\}$. The blue points are the result of the denoising process. The black line corresponds to the original set S_{L_3}

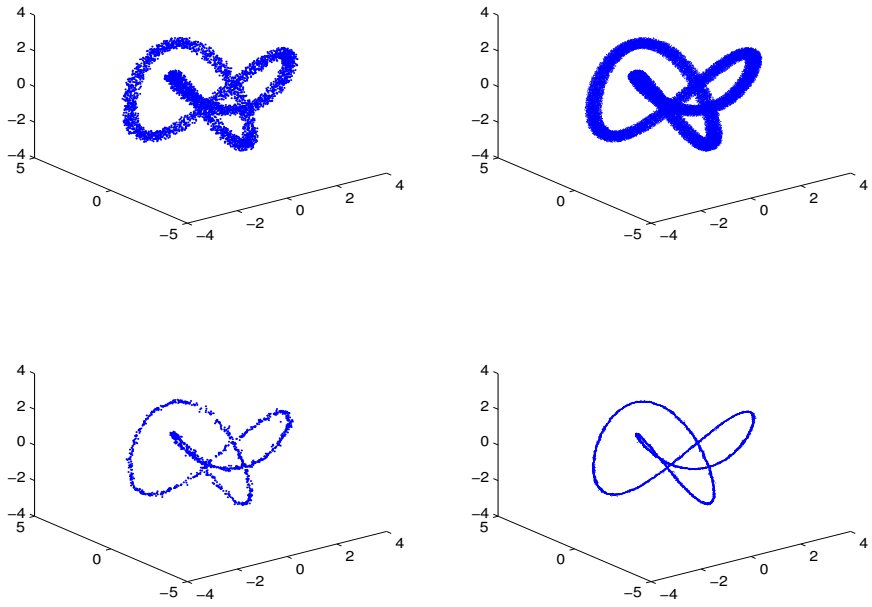

Figure 4: The upper panel shows 5000 noisy points (left) and 50000 noisy points (right) drawn on $\mathcal{B}(T, 0.3)$. The lower panel shows the result of the corresponding denoising process.

Minkowski contents estimation

Finally in Table 2 we show some results about the Minkowki contents estimation, again in the case of noisy points drawn around a sphere (with R_1=0,0.1,0.3) for different values for n and different dimensions. For every R_1, n, d we estimate the Minkowski contents (via Monte Carlo with 10^5 points) and simulated 100 random samples. Table 2 entries provide the average relative error (in percentage) in the estimation of the boundary Minkowski contents L. That is, the entries are $100 \cdot err(R_1, d)$ where $err(R_1, d) = \frac{1}{L}\sqrt{\sum_i (L_i(R_1, d) - L)^2/100}$.

In the Minkowski contents we used a radius $\rho = \sqrt{r}/2$, with $r = 2\max_i(\min_{j\neq i} \|X_j - X_i\|)$, when $\tilde{R}_n(\mathcal{X}_n) = 0$, where $\tilde{R}_n(\mathcal{X}_n) = \max_i d(X_i, \partial C_r(\mathcal{X}_n))$. In the case $\tilde{R}_n(\mathcal{X}_n) > 0$ we have used $\rho = \sqrt{r + \tilde{R}_n(\mathcal{Y}_n)}/2$, where $\mathcal{Y}_n$ stands for the denoised sample and $\tilde{R}_n(\mathcal{Y}_n) = \max_j d(Y_j, \partial C_r(\mathcal{Y}_n))$.

d	R_1	$n = 1000$	$n = 2000$	$n = 5000$	$n = 10000$
2	0	0.50	0.50	0.76	0.87
2	0.1	8.51	8.63	7.61	7.47
2	0.3	11.97	12.12	12.24	12.84
3	0	0.57	0.44	0.57	0.52
3	0.1	4.75	6.44	12.76	13.34
3	0.3	7.29	13.11	16.24	16.11
4	0	1.92	2.85	3.47	3.46
4	0.1	52.70	27.22	3.34	11.65
4	0.3	34.29	18.89	15.92	22.27

Table 2: Relative error for Minkowski contents estimation

Acknowledgements

This research has been partially supported by MATH-AmSud grant 16-MATH-05 SM-HCD-HDD (C. Aaron and A. Cholaquidis) and Spanish grant MTM2013-44045-P (A. Cuevas). We are grateful to Luis Guijarro and Jesús Gonzalo (Dept. Mathematics, UAM, Madrid) for useful conversations and advice.

References

Aamari, E. and Levrard, C. (2015). Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Manuscript arXiv:1512.02857v1.*

Aaron, C. and Cholaquidis, A.(2016). On boundary detection. *Manuscript.*

Adler, R.J., Krishnan, S.R., Taylor, J.E. and Weinberger, S. (2015). Convergence of the Reach for a Sequence of Gaussian-Embedded Manifolds. *arXiv preprint* arXiv:1503.01733.

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

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

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

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

Avila A. and Lybich, M. (2007). Hausdorff dimension and conformal measures of Feigenbaum Julia Sets. *Journal of the American Mathematical Society* 21(2), 305–363.

Baldin, N. and M. Reiss (2015). Unbiased estimation of the volume of a convex body. *Manuscript*, arXiv:1502.05510. To appear in *Stochastic Processes and their Applications*.

Berrendero, J.R., Cuevas, A.. and Pateiro-López, B. (2012). A multivariate uniformity test for the case of unknown support *Stat. Comput.* **22** 259–271.

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

Boothby, W.M. (1975). *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, New York.

Brito, M.R., Quiroz, A.J., Yukich, J.E. (2013). Intrinsic dimension identification via graph-theoretic methods. *J. Multivariate Anal.* 116, 263-277.

Bhattacharya, R. and Patrangenaru, V.(2014) Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications. *J. Statist. Plann. Inf.* 145, 1–22.

Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc. (N.S.)* 46, 255–308.

Chazal, F. and Lieutier, A. (2005). The "λ-medial Axis". *Graphical Models*, 67, 304–331.

Chen, D. and Müller, H. G. (2012). Nonlinear manifold representations for functional data. *Ann. Statist.*, 40(1), 1-29.

Chazal, F., Glisse, M., Labru'ere, C. and Michel, B. (2013) Optimal rates of convergence for persistence diagrams in Topological Data Analysis. *ArXiv e-prints, May 2013.*

Cholaquidis, A., Cuevas, A. and Fraiman, R. (2014) On Poincaré cone property. *Ann. Statist.*, **42**, 255–284.

Cuevas, A. and Fraiman, R. (1997). A plug-in approach to support estimation. *Ann. Statist.* **25,** 2300-2312.

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

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

Cuevas, A. and Fraiman, R. (2010). Set Estimation. In *New Perspectives on Stochastic Geometry*, W.S. Kendall and I. Molchanov, eds., pp. 374–397. Oxford University Press.

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

Cuevas, A., Fraiman, R. and Györfi, L. (2013). Towards a universally consistent estimator of the Minkowski content. *ESAIM: Probability and Statistics*, **17**, 359-369.

Delicado, P. (2001) Another look at principal curves and surfaces. *J. Multivariate Anal.* 77, 84116.

Do Carmo, M. (1992). *Riemannian Geometry.* Birkhäuser, Boston.

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

Evans, L. and Gariepy, R. (1992). Measure theory and fine properties of functions. *CRC Press, Inc.*

Fasy, B.T., Lecci, F., Rinaldo, R., Wasserman, L. Balakrishnan, S. and Singh, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42**, 2301–2339.

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

Federer, H. (1969). Geometric measure theory *Springer.*

Fefferman, C., Mitter, S. and Narayanan, H. Testing the manifold hypothesis To appear in *J. Amer. Math. Soc.*

Galbis, A. and Maestre, M. (2010). *Vector Analysis Versus vector Calculus.* Springer, New York.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012a). The geometry of nonparametric filament estimation. *J. Amer. Statist. Assoc.* 107, 788-799.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012b). Minimax Manifold Estimation. *Journal of Machine Learning Research* 13, 1263-1291.

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012c). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* 40, 941-963

Hastie, T. and Stuetzle, W. (1989). Principal curves. *J. Amer. Statist. Assoc.* 84, 502-516.

Guillemin, V. and Pollack, A. *Differential Topology.* Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Hirsch, M.W. Differential Topology. Springer-Verlag, New York.

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

Mardia, K.V. and Jupp, P.E. (2000) *Directional Statistics.* Wiley, Chichester.

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

Niyogi, P., Smale, S. and Weinberger, S.(2008) Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete Comput. Geom.* **39**. 419–441.

Niyogi, P., Smale, S. and Weinberger, S. (2011) A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* 40, no. 3, 646–663.

Pateiro-López, B. and Rodríguez-Casal, A. (2009) Surface area estimation under convexity type assumptions. *Journal of Nonparametric Statistics* **21**(6), 729–741

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

Pennec, X2006 Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements *Journal of Mathematical Imaging and Vision* **25**(1) pp 127–154

Penrose. M.D. (1999) A strong law for the largest nearest-neighbour link between random points. *J. London Math. Soc.* **60**(3), 951–960.

Ranneby, B. (1984). The maximal spacing method. An estimation method related to maximum likelihood method. *Scand. J. Statist.* **11** 93–112.

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

Taylor, M.E. (2006). *Measure Theory and Integration.* American Mathematical Society. Providence.

Tenenbaum, , J.B., de Silva, V. and Langford, J.C. (2000). A Global geometric framework for nonlinear dimensionality reduction. *Science* 290, 2319-2323.

Thäle, C. (2008). 50 years sets with positive reach. A survey. *Surveys in Mathematics and its Applications* 3, 123–165.

Walther, G. (1999) On a generalization of Blaschke's rolling theorem and the smoothing of surfaces, *Math. Meth. Appl. Sci.* **22** 301–316.

Zhang, Q.S. (2011). *Sobolev Inequalities, Heat Kernel under Ricci Flow and the Poincar Conjecture.* CRC Press, Boca Raton.