A statistical methodology to detect low-dimensionality

Antonio Cuevas
Departamento de Matemáticas
Universidad Autónoma de Madrid, Spain

19th Workshop on Stochastic Geometry, Stereology and Image Analysis (SGSIA)
Marseille, May 19, 2017

Co-authors

- **Catherine Aaron** (Université Blaise Pascal, Clermont-Ferrand II, France)
- **Alejandro Cholaquidis** (Centro de Matemática, Universidad de la República, Uruguay)

Source:
Aaron, C., Cholaquidis, A. and Cuevas, A. (2017). Stochastic detection of some topological and geometric features. Submitted. `https://arxiv.org/abs/1702.05193`.

The general setup

This talk is concerned with the following general problem:

How much can we learn about a (**nice enough**) compact set $\mathcal{M} \subset \mathbb{R}^d$ **from a random sample** of points $X_1, \ldots, X_n$?

The general setup

This talk is concerned with the following general problem:

How much can we learn about a (**nice enough**) compact set $\mathcal{M} \subset \mathbb{R}^d$ **from a random sample** of points $X_1, \ldots, X_n$?

In this talk we will consider the following aspects of this general problem:

- **To check whether $\mathcal{M}$ has an empty interior, $\mathring{\mathcal{M}} = \emptyset$**: under regularity conditions this amounts to saying that $dim_H(\mathcal{M}) < d$
- **Estimation of $\mathcal{M}$** when the sample is "noisy (around $\mathcal{M}$)
- **Estimation of the measure of $\mathcal{M}$**: more specifically, if d' is the dimension of $\mathcal{M}$, we are interested on the d'-dimensional Minkowski content of $\mathcal{M}$.

The tools we use

- Statistical tools: definition of different sample models, methods for analyzing stochastic convergences and convergence rates, set estimation methodologies.
- Geometrically motivated conditions for sets: standardness, positive reach, rolling conditions,...
- Some basic tools of differential geometry,
- Some results of stochastic geometry borrowed from Penrose (1999, J. London Math. Soc), Walther (1997, Ann. Statist.) among others

Hausdorff measure and Hausdorff dimension

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

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

Checking empty interior (low dimensionality) under the noiseless model

Let $\aleph_n = \{X_1, \ldots, X_n\}$ be a *iid* sample of points, drawn from a distribution P with support $\mathcal{M}$ in $\mathbb{R}^d$.

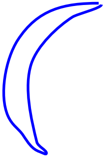
We want to detect whether or not $\mathring{\mathcal{M}} = \emptyset$.
First note that, if $\mathcal{M} \subset \mathbb{R}^d$ is "regular enough", $\dim_H(\mathcal{M}) < d$ is in fact equivalent to $\mathring{\mathcal{M}} = \emptyset$.
Indeed, in general $\dim_H(\mathcal{M}) < d$ implies $\mathring{\mathcal{M}} = \emptyset$. The converse implication is not always true, even for sets fulfilling the property $\mathcal{H}^d(\partial\mathcal{M}) = 0$. However it holds if $\mathcal{M}$ **has positive reach** (*), since in this case $\mathcal{H}^{d-1}(\partial\mathcal{M}) < \infty$ (see Ambrosio et al. (2008)).

Checking empty interior (low dimensionality) under the noiseless model

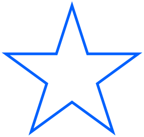
Let $\aleph_n = \{X_1, \ldots, X_n\}$ be a *iid* sample of points, drawn from a distribution P with support $\mathcal{M}$ in $\mathbb{R}^d$.

We want to detect whether or not $\mathring{\mathcal{M}} = \emptyset$.
First 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 $\mathcal{H}^d(\partial\mathcal{M}) = 0$. However it holds if $\mathcal{M}$ **has positive reach** (*), since in this case $\mathcal{H}^{d-1}(\partial\mathcal{M}) < \infty$ (see Ambrosio et al. (2008)).

(*) reach$(\mathcal{M}) = R > 0$ iff R is the supremum of those values $r > 0$ such that every point x with $d(x, \mathcal{M}) < r$ has only one metric projection on $\mathcal{M}$
This regularity condition **rules out the presence of sharp inward peaks** in $\mathcal{M}$.



reach>0



reach=0

A simple tool: the offset estimator

The r-offset estimator (Grenander (1981), Devroye & Wise (1980),...) based on the sample $\aleph_n$ is defined by

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

$B(X_i, r)$ is a **boundary ball** of $\hat{S}_n(r)$ if there exists a point $y \in \partial B(X_i, r)$ 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 result of removing from $\hat{S}_n(r)$ all the boundary balls.

We are going to explore the following natural idea

$\mathring{\mathcal{M}} \neq \emptyset$ iff $\mathrm{peel}(\hat{S}_n(r_n)) \neq \emptyset$, eventually a.s. for suitably chosen r_n

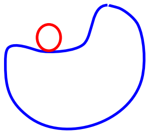
An identification result in terms of $\mathrm{peel}(\hat{S}_n(r_n))$

Theorem 1 (Identification of empty interior, noiseless case)

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact non-empty set. We have,
(a) if $\mathring{\mathcal{M}} = \emptyset$, and $\mathcal{M}$ fulfills the **outside rolling condition** *(**) for some $r > 0$, then $\mathrm{peel}(\hat{S}_n(r')) = \emptyset$ for any $r' < r$.*

(b) If $\mathring{\mathcal{M}} \neq \emptyset$, 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. Then $\mathrm{peel}(\hat{S}_n(r_n)) \neq \emptyset$ 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 $\kappa > (\delta\omega_d)^{-1}$.*

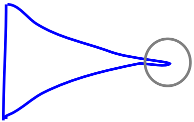
(**) The set S is said to satisfy the outside r-rolling condition if for $s \in \partial S$ there exists some $x \in S^c$ such that $\mathcal{B}(x, r) \cap \partial S = \{s\}$.



(***)

$\exists\ \lambda, \delta > 0$ such that $, \forall x \in S$ and $0 < \varepsilon \leq \lambda$,

$P_X(\mathcal{B}(x, \varepsilon) \cap S) \geq \delta \mu_d(\mathcal{B}(x, \varepsilon))$



Why $\left(\frac{\log n}{n}\right)^{1/d}$?

- This is the appropriate order to use some basic Borel-Cantelli arguments in the proof.
- Note this is also the order of the **maximal multivariate spacing**, i.e., the radius of the ball containing no sample point, Janson (1987, Ann. Prob.)

When $\mathcal{M}$ is a manifold

Theorem 2 (Identification of empty interior, the manifold case)

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 $x \in \mathcal{M}$. 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 $\mathrm{peel}(\hat{S}_n(r_n)) \neq \emptyset$ eventually, a.s.

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

The proof is based upon Th. 1, using results by Walther (1997, Ann. Stat.) and Penrose (1999, J. London Mat. Soc.)

"Noisy model": sample data on the parallel set $B(M, R_1)$

Theorem 3 (Estimation of the noise level)

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact set such that 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 us denote $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$, and $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$ where $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$.

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

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

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

$$|\hat{R}_n - R_1| > C \quad \textit{for } n \textit{ large enough}. \tag{3}$$

The proof relies on Federer (1959) and C. & R.-Casal (2004, AAP)

An index of closeness to lower dimensionality

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

Identifying the boundary balls

Proposition 1

Let $\mathcal{X}_n = \{X_1, \dots, X_n\}$ 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 $i = 1, \dots, n$ and all $r > 0$, $\sup\{\|z - X_i\|, z \in Vor(X_i)\} \geq r$ if and only if $\mathcal{B}(X_i, r)$ is a boundary ball for the Devroye-Wise estimator $\cup_i \mathcal{B}(X_i, r)$.

An algorithm to partially "de-noise" the sample (I)

Let $\mathcal{M} \subset \mathbb{R}^d$ with $\mathrm{reach}(\mathcal{M}) = R_0 > 0$. Let $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ be an iid sample on $S = B(\mathcal{M}, R_1)$ for some $0 < R_1 < R_0$, with density bounded from below.

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. *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,

$$Z_i = \pi_{\partial \hat{S}_n}(Y_i^\lambda) + \hat{R}_n \frac{Y_i^\lambda - \pi_{\partial \hat{S}_n}(Y_i^\lambda)}{\| Y_i^\lambda - \pi_{\partial \hat{S}_n}(Y_i^\lambda)\|}, \qquad (4)$$

being $\pi_{\partial \hat{S}_n}(Y_i^\lambda)$ the metric projection of Y_i^λ on $\partial \hat{S}_n$.

An algorithm to partially "de-noise" the sample (II)

Theorem 4 (Estimation by denoising)

If $\mathrm{reach}(\mathcal{M}) = R_0 > 0$ *and* $\mathcal{Y}_n = \{Y_1, \ldots, Y_n\}$ *is an iid sample on* $S = B(\mathcal{M}, R_1)$ *with density* $f > f_0 > 0$, *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, the de-noised sub-sample* $\mathcal{Z}_m$ *satisfies*

$$d_H(\mathcal{Z}_m, \mathcal{M}) \leq b_n$$

where $\varepsilon_n = c(\log(n)/n)^{1/d}$ *with* $c > (4/f_0\omega_d)^{1/d}$.

Note that, when $\mathring{\mathcal{M}} = \emptyset$, the result simplifies since, according to Theorem 3 we can take $e_n = 2\epsilon_n$ and, according to C. & R.-Casal (2004, Adv. Appl. Prob.) $a_n = (\log n/n)^{1/d}$. Therefore, in this case $b_n = \mathcal{O}\left((\log n/n)^{1/3d}\right)$. In particular, the result is true for $b_n = n^{-q}$ for any $q < \frac{1}{3d}$.

Estimation of the Minkowski content

The target now is to estimate the d'-dimensional Minkowski content, as defined by

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

This is just (alongside with Hausdorff measure, among others) one of the possible ways to measure lower-dimensional sets.
The following result shows that the denosing process allows us to estimate the Minkowski content, even under the noisy model.

Theorem 5 (Boundary measure estimation in the noisy case)

With the hypothesis and notation of the previous denoising theorem, assume that $L_0(\mathcal{M}) < \infty$. Now, take r_n such that $b_n/r_n \to 0$. Then,

$$\lim_{n \to \infty} \frac{\mu_d(B(\mathcal{Z}_m, r_n))}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) \quad a.s. \tag{6}$$

For $d' < d$ the result holds with $r_n = n^{-q}$ for any $q < \frac{1}{3d}$.

Toy examples: **identifying low dimensionality**

200 samples of sizes $n = 50$, 100, 200, 300, 400, 500, 1000, 2000, 5000, 10000 on the set $\mathcal{B}(0, 1 + A) \setminus \mathring{\mathcal{B}}(0, 1 - A)$. 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.

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: Minimum sample sizes required to correct detection (i.e. in 190 out of 200 cases) for different values of d and A.

Denoising (Lamé curve $|x|^3 + |y|^3 = 1$)

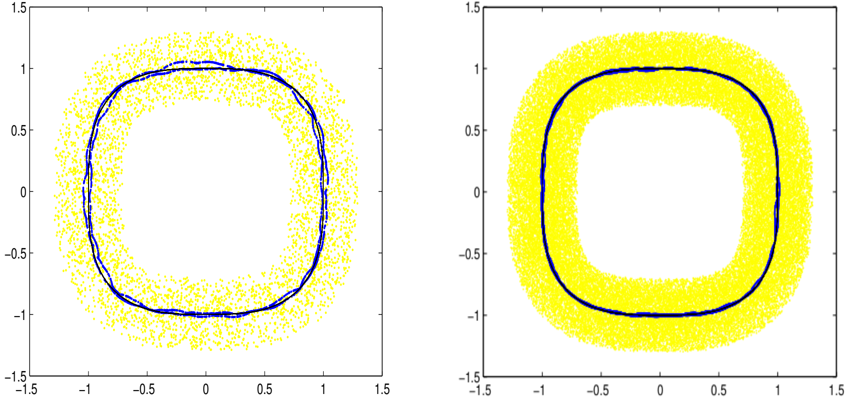


Figure: The yellow background is made of 5000 points (left) and 50000 points (right) drawn on on $\mathcal{B}(S_{L_3}, 0.3)$, with $S_{L_3} = \{(x, y), |x|^3 + |y|^3 = 1\}$. The blue points are the result of the denoising process. The black line corresponds to the original set S_{L_3}

Denoising (Trefoil knot)

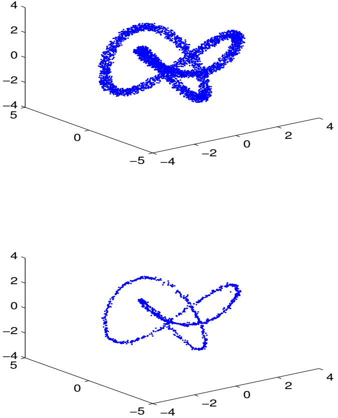


Figure: The upper panel shows 5000 noisy points (left) and 50000 noisy points (right) drawn on $\mathcal{B}(T, 0.3)$. The lower panel shows the result of the corresponding denoising process.

THANKS!