Stochastic detection of some topological and geometric features.

Alejandro Cholaquidis

CMAT-Facultad de Ciencias, UdelaR
Montevideo Uruguay

A joint work with: C. Aaron and A. Cuevas

Seminario de Probabilidad y Estadística

CMAT - Facultad de Ciencias.

Motivation

To identify features of $S \subset \mathbb{R}^d$, of geometric or topological character from a random sample of points drawn on S.

"To identify" means to answer correctly, almost surely (a.s.) when the sample size tends to infinity.

Motivation

To identify features of $S \subset \mathbb{R}^d$, of geometric or topological character from a random sample of points drawn on S.

"To identify" means to answer correctly, almost surely (a.s.) when the sample size tends to infinity.

More specifically aims at giving some partial answers to the following questions:

- Is S full dimensional?

Motivation

To identify features of $S \subset \mathbb{R}^d$, of geometric or topological character from a random sample of points drawn on S.

"To identify" means to answer correctly, almost surely (a.s.) when the sample size tends to infinity.

More specifically aims at giving some partial answers to the following questions:

- Is S full dimensional?
- If S is full dimensional, is it "close to a lower dimensional set" $\mathcal{M}$?

Motivation

To identify features of $S \subset \mathbb{R}^d$, of geometric or topological character from a random sample of points drawn on S.

"To identify" means to answer correctly, almost surely (a.s.) when the sample size tends to infinity.

More specifically aims at giving some partial answers to the following questions:

- Is S full dimensional?
- If S is full dimensional, is it "close to a lower dimensional set" $\mathcal{M}$?
- If S is "close to a lower dimensional $\mathcal{M}$", can we
  - a) estimate $\mathcal{M}$?

Motivation

To identify features of $S \subset \mathbb{R}^d$, of geometric or topological character from a random sample of points drawn on S.

"To identify" means to answer correctly, almost surely (a.s.) when the sample size tends to infinity.

More specifically aims at giving some partial answers to the following questions:

- Is S full dimensional?
- If S is full dimensional, is it "close to a lower dimensional set" $\mathcal{M}$?
- 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 $\mathcal{M}$)?

Example, denoising samples

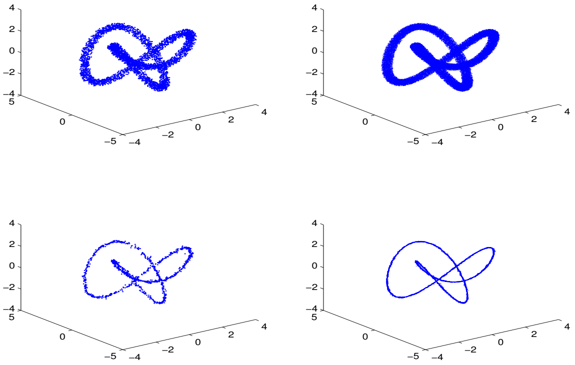

Figure: The upper panel shows 5000 noisy points (left) and 50000 noisy points (right) drawn on $\mathcal{B}(T, 0.3)$. The lower panel shows the result of the corresponding denoising process.

The models

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be random sample points drawn on an unknown compact set $S \subset \mathbb{R}^d$. We consider two different models:

- ***The noiseless model*: ** $\mathcal{X}_n$ is 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 models

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be random sample points drawn on an unknown compact set $S \subset \mathbb{R}^d$. We consider two different models:

- ***The noiseless model*:** $\mathcal{X}_n$ is 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*:** $\mathcal{X}_n$ supported on $S = B(\mathcal{M}, R_1)$, $R_1 > 0$, where $\mathcal{M}$ is a d'-dimensional set with $d' \leq d$; Berrendero et al. (2014). Other different models "with noise" are considered in Genovese et al. (2012a), Genovese et al. (2012b) and Genovese et al (2012c).

Hausdorff dimension

Hausdorff Measure

Given $(\mathcal{M},\rho)$ metric space, $\delta, r > 0$ and $E \subset \mathcal{M}$, let

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

where $\mathrm{diam}(B) = \sup\{\rho(x,y) : x, y \in B\}$, $\inf \emptyset = \infty$.

Hausdorff dimension

Hausdorff Measure

Given $(\mathcal{M}, \rho)$ metric space, $\delta, r > 0$ and $E \subset \mathcal{M}$, let

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

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

Hausdorff dimension

$$\dim_H(E) = \inf\{r : \mathcal{H}^r(E) = 0\} = \sup\{r : \mathcal{H}^r(E) = \infty\}. \tag{1}$$

Hausdorff dimension II

- When $\mathcal{M}$ is a k-dimensional smooth manifold, $\dim_H(\mathcal{M}) = k$.

Hausdorff dimension II

- When $\mathcal{M}$ is a k-dimensional smooth manifold, $\dim_H(\mathcal{M}) = k$.
- In general: $\dim_H(\mathcal{M}) < d \Rightarrow \mathring{\mathcal{M}} = \emptyset$.
The converse implication not always true, even if $\mathcal{H}^d(\partial\mathcal{M}) = 0$, see Avila and Lyubich (2007).
It holds if $\text{reach}(\mathcal{M}) > 0$ since $\mathcal{H}^{d-1}(\partial\mathcal{M}) < \infty$ (Ambrosio, Colesanti and Villa (2008)).

Hausdorff dimension II

- When $\mathcal{M}$ is a k-dimensional smooth manifold, $\dim_H(\mathcal{M}) = k$.
- In general: $\dim_H(\mathcal{M}) < d \Rightarrow \mathring{\mathcal{M}} = \emptyset$.
  The converse implication not always true, even if $\mathcal{H}^d(\partial\mathcal{M}) = 0$, see Avila and Lyubich (2007).
  It holds if $\text{reach}(\mathcal{M}) > 0$ since $\mathcal{H}^{d-1}(\partial\mathcal{M}) < \infty$ (Ambrosio, Colesanti and Villa (2008)).
- If $\mathcal{M} \subset \mathbb{R}^d$ is a manifold, $\mathring{\mathcal{M}} = \emptyset \Leftrightarrow \dim_H(\mathcal{M}) < d$.

Boundary Balls

Devroye-Wise estimator

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

Boundary Balls

$\mathcal{B}(x_i, r)$ is a boundary ball of $\hat{S}_n(r)$ if $\exists y \in \partial\mathcal{B}(x_i, r)$ such that $y \in \partial\hat{S}_n(r)$.
$\mathrm{peel}(\hat{S}_n(r))$ is the union of all non-boundary balls of $\hat{S}_n(r)$.

Boundary Balls

Devroye-Wise estimator

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

Boundary Balls

$\mathcal{B}(x_i, r)$ is a boundary ball of $\hat{S}_n(r)$ if $\exists y \in \partial\mathcal{B}(x_i, r)$ such that $y \in \partial\hat{S}_n(r)$.
peel$(\hat{S}_n(r))$ is the union of all non-boundary balls of $\hat{S}_n(r)$.

Proposition

$\mathcal{X}_n = \{X_1, \dots, X_n\}$ *iid* of $P_X \ll \mu$, being μ the Lebesgue measure. Then, with probability one, for all $i = 1, \dots, n$ and all $r > 0$,

$$\sup\{\|z - X_i\|, z \in \text{Vor}(X_i)\} \geq r \Leftrightarrow \mathcal{B}(X_i, r) \text{ is a b.b}$$

Boundary Balls and empty interior of general sets

Theorem

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact non-empty set.

- 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 $\hat{S}_n(r')$ with $r' < r$.

Boundary Balls and empty interior of general sets

Theorem

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact non-empty set.

- 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 $\hat{S}_n(r')$ with $r' < r$.
- if $\mathring{\mathcal{M}} \neq \emptyset$ and 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.

Boundary Balls and empty interior of general sets

Theorem

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact non-empty set.

- 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 $\hat{S}_n(r')$ with $r' < r$.
- if $\mathring{\mathcal{M}} \neq \emptyset$ and 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.
  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 $\kappa > (\delta \omega_d)^{-1}$.

Boundary Balls and empty interior of manifolds

Theorem

Let $\mathcal{M}$ be a d'-dimensional compact manifold in $\mathbb{R}^d$ and $X_1, \ldots, X_n$ from P_X with support $\mathcal{M}$ with 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 $\mathcal{C}^2$ then $\text{peel}(\hat{S}_n(r_n)) \neq \emptyset$ eventually, a.s..

Boundary Balls and empty interior of manifolds

Theorem

Let $\mathcal{M}$ be a d'-dimensional compact manifold in $\mathbb{R}^d$ and $X_1, \ldots, X_n$ from P_X with support $\mathcal{M}$ with 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 $\mathcal{C}^2$ 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..

Some simulations

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;

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 to correctly decide on, at least 190 out of 200, that the support is lower dimensional (in the case $A = 0$) or that it is full dimensional (cases with $A > 0$).

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume $\text{reach}(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume reach$(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Theorem

$P_Y \ll \mu$, μ the the Lebesgue measure. The density f fulfils $f > f_0$. Let

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

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume $\text{reach}(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Theorem

$P_Y \ll \mu$, μ the the Lebesgue measure. The density f fulfils $f > f_0$. Let

- $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$,
- $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n) \text{ is a boundary ball}\}$.

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume reach$(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Theorem

$P_Y \ll \mu$, μ the the Lebesgue measure. The density f fulfils $f > f_0$. Let

- $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$,
- $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$.
- $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume reach$(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Theorem

$P_Y \ll \mu$, μ the the Lebesgue measure. The density f fulfils $f > f_0$. Let

- $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$,
- $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$.
- $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$

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

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

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume reach$(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Noisy model

$\mathcal{Y}_n$ supported on $S = B(\mathcal{M}, R_1)$, assume $\text{reach}(\mathcal{M}) = R_0$ and $0 < R_1 < R_0$.

Goal

- if we know R_1, decide if $\mathcal{M}$ is full dimensional or not.
- if $\dim_H(\mathcal{M}) < d$, estimate R_1.

Theorem

$P_Y \ll \mu$, μ the the Lebesgue measure. The density f fulfils $f > f_0$. Let

- $\varepsilon_n = c(\log(n)/n)^{1/d}$, with $c > (4/(f_0\omega_d))^{1/d}$,
- $I_{bb} = \{j : \mathcal{B}(Y_j, \varepsilon_n)$ is a boundary ball$\}$.
- $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$

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

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

The algorithm

$\mathcal{M} \subset \mathbb{R}^d$ compact, reach$(\mathcal{M}) = R_0 > 0$, $\mathcal{Y}_n$ iid of Y, with support $S = B(\mathcal{M}, R_1)$, $0 < R_1 < R_0$. P_Y with density $f > f_0 > 0$.

The algorithm

$\mathcal{M} \subset \mathbb{R}^d$ compact, reach$(\mathcal{M}) = R_0 > 0$, $\mathcal{Y}_n$ iid of Y, with support $S = B(\mathcal{M}, R_1)$, $0 < R_1 < R_0$. P_Y with density $f > f_0 > 0$.

1. $\hat{S}_n$ 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 $e_n \to 0$.

The algorithm

$\mathcal{M} \subset \mathbb{R}^d$ compact, reach$(\mathcal{M}) = R_0 > 0$, $\mathcal{Y}_n$ iid of Y, with support $S = B(\mathcal{M}, R_1)$, $0 < R_1 < R_0$. P_Y with density $f > f_0 > 0$.

1. $\hat{S}_n$ 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. Fixed $\lambda \in (0, 1)$, 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 $d(Y_i^\lambda, \partial\hat{S}_n) > \lambda\hat{R}_n$.

The algorithm

$\mathcal{M} \subset \mathbb{R}^d$ compact, $\text{reach}(\mathcal{M}) = R_0 > 0$, $\mathcal{Y}_n$ iid of Y, with support $S = B(\mathcal{M}, R_1)$, $0 < R_1 < R_0$. P_Y with density $f > f_0 > 0$.

1. $\hat{S}_n$ 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. Fixed $\lambda \in (0,1)$, define $\mathcal{Y}_m^\lambda = \{Y_1^\lambda, \dots, Y_m^\lambda\} \subset \mathcal{Y}_n$ 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. For every $Y_i^\lambda \in \mathcal{Y}_m^\lambda$, define $\{Z_1, \dots, Z_m\} = \mathcal{Z}_m$ as follows,

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

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

Consistency

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

Consistency

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,

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

Consistency

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

Consistency

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,

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

Corollary

Given $\lambda \in (0,1)$, let $\mathcal{Z}_n$ be the points obtained using $\hat{R}_n = \max_{Y_i \in \mathcal{Y}_n} \min_{j \in I_{bb}} \|Y_i - Y_j\|$ to estimate R_1 and $\{Y_i, i \in I_{bb}\}$ as an estimator of ∂S. Then,

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

Simulations

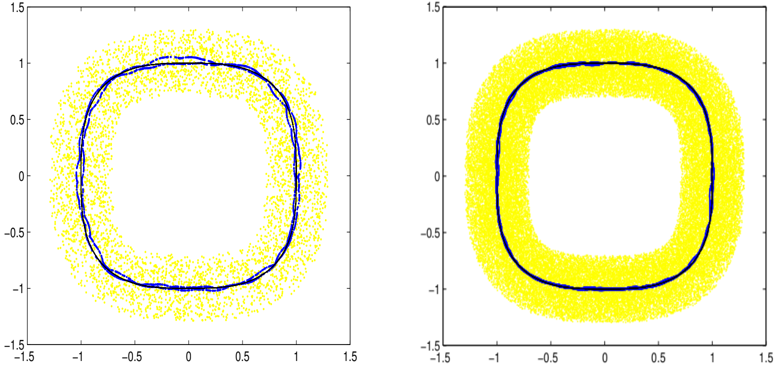

Figure: 5000 points (left) and 50000 points (right) drawn on $\mathcal{B}(S_{L_3}, 0.3)$, with $S_{L_3} = \{(x, y), |x|^3 + |y|^3 = 1\}$. The black line corresponds to the original set S_{L_3}

Minkowski content estimation

Definition

d'-dimensional Minkowski content of $\mathcal{M}$,

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

Minkowski content estimation

Definition

d'-dimensional Minkowski content of $\mathcal{M}$,

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

Noiseless model

$\mathcal{X}_n = \{X_1, \ldots, X_n\}$ iid of P_X on $\mathcal{M} \subset \mathbb{R}^d$, P_X is standard w.r.t the d'-dimensional Lebesgue measure, there exists $L_0(\mathcal{M})$. Let r_n such that $r_n \to 0$ and $(\log(n)/n)^{1/d'} = o(r_n)$, then

(a)

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

Minkowski content estimation

Definition

d'-dimensional Minkowski content of $\mathcal{M}$,

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

Minkowski content estimation

Definition

d'-dimensional Minkowski content of $\mathcal{M}$,

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

Noiseless model

$\mathcal{X}_n = \{X_1, \ldots, X_n\}$ iid of P_X on $\mathcal{M} \subset \mathbb{R}^d$, P_X is standard w.r.t the d'-dimensional Lebesgue measure, there exists $L_0(\mathcal{M})$. Let r_n such that $r_n \to 0$ and $(\log(n)/n)^{1/d'} = o(r_n)$, then

(b) If $\text{reach}(M) = R_0 > 0$, 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 $\beta_n := d_H(\mathcal{X}_n, \mathcal{M}) = \mathcal{O}\big(\log(n)/n\big)^{1/d'}$.

Minkowski content estimation

Noisy model

If $\max(a_n^{1/3}, e_n, \varepsilon_n) = o(r_n)$ where $\varepsilon_n = c(\log(n)/n)^{1/d}$ with $c > (4/(\omega_d f_0))^{1/d}$, then,

$$\lim_{n\to\infty} \frac{\mu_d(B(\mathcal{Z}_m, r_n))}{\omega_{d-d'} r_n^{d-d'}} = L_0(\mathcal{M}) \quad a.s.. \tag{7}$$

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, 84?116.

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 Poincaré Conjecture*. CRC Press, Boca Raton.