Local Convex Hull density and level set estimation

Catherine Aaron

Université Clermont-Auvergne

September 17, 2021

Context

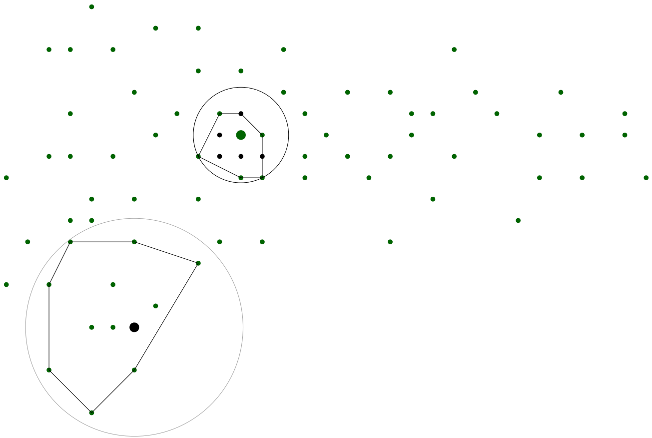
1. Statistics : YES
2. Dependence : NO
3. Ecology : Applicable

LcH Original paper

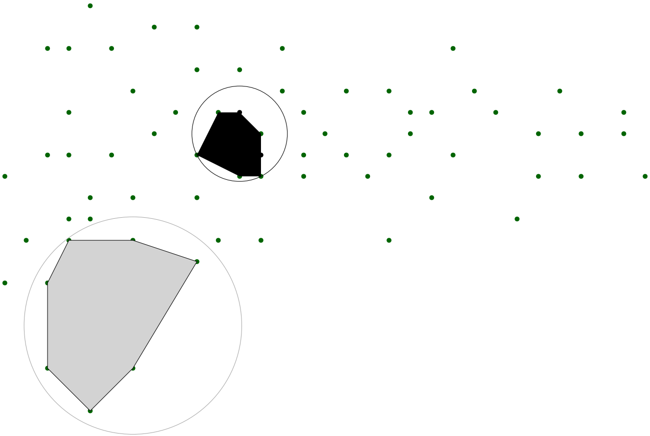
Original paper : A local nearest-neighbor convex-hull construction of home ranges and utilization distributions (2004) Getz and Wilmer Ecography : 475 citations

1. Let $X_1, \ldots, X_n$ be the observations, for all i compute $C_i = H(B(X_i, r_{k(i)})$ the convex hull of the k-NN of X_i and $|C_i|$ its volume.
2. Sort by decreasing $|C_i|$
3. Plot $C_1, \ldots, C_n$ filled with an increasing level of grey

LcH Original paper



LcH Original paper



LcH Original translated in math

$$\text{Density estimator : } \hat{f}_n(x) = \max_{x \in C_i} \frac{k}{n|C_i|}$$

$$\text{Level Set estimator : } \hat{L}_t = \{x, \hat{f}_n(x) \geq t\}$$

LcH Original : Some critics

$$\text{Density estimator : } \hat{f}_n(x) = \max_{x \in C_i} \frac{k}{n|C_i|}$$

1. Good points : the division by $|C_i|$ correct the bias near the boundary of the support in case of compact support and bounded bellow density
2. Problem : There is a "double" overestimation of the density
   1. Mainly $|C_i|$ overestimates $|B(X_i, r_{k(i)})|$
   2. The max in $\hat{f}_n$ emphasizes this problem.

Proposition

The aim of this talk is to propose new, support, density and level set estimators, based on the very good original idea of the Local convex hulls.
For simplicity we work with fix radius instead of nearest neighbors approach

First Hypotheses and notations

Let $X_1, \ldots, X_n$ be iid, in $\mathbb{R}^d$ drawn with a density f bounded bellow by a positive f_0 on S the support of the distribution that is supposed to be compact and to satisfy the inside and outside (r_0)-rolling ball property.

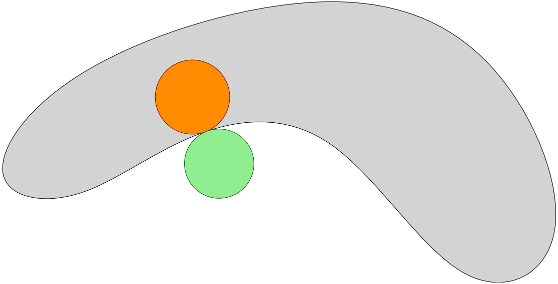
Definition

A set compact S satisfies the inside and outside (r_0)-rolling ball property if, for all $x \in \partial S$ there exits O_x^+ and O_x^- such that $||O_x^+ - x|| = ||O_x^- - x|| = r_0$ and $\mathring{B}(O_x^+, r_0) \subset S$ and $\mathring{B}(O_x^-, r_0) \subset S^c$

First Hypotheses and notations

Definition

A set compact S satisfies the inside and outside (r_0)-rolling ball property if, for all $x \in \partial S$ there exits O_x^+ and O_x^- such that $||O_x^+ - x|| = ||O_x^- - x|| = r_0$ and $\mathring{B}(O_x^+, r_0) \subset S$ and $\mathring{B}(O_x^-, r_0) \subset S^c$



First Hypotheses and notations

$\mathbb{X}_n = \{X_1, \ldots, X_n\}$, $H(A)$ denotes the convex hull of a set A.

$$C_{x,r_n} = H(B(x, r_n) \cap \mathbb{X}_n)$$

and

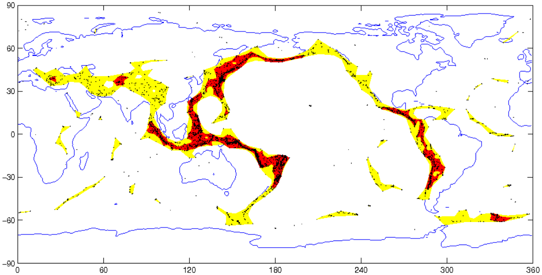
$$\hat{S}_{r_n}(\mathbb{X}_n) = \bigcup_i C_{X_i,r_n}$$

First results on support estimation

Theorem (A. and Bodart 2016)

under the above mentionned hypotheses, for $r_n = c(\ln n/n)^{2/(d+1)}$ *we have*

1. $d_h(\hat{S}_{r_n}(\mathbb{X}_n), S) = O((\ln n/n)^{2/(d+1)})$ *e.a.s.*
2. $d_h(\partial\hat{S}_{r_n}(\mathbb{X}_n), \partial S) = O((\ln n/n)^{2/(d+1)})$ *e.a.s.*
3. $\hat{S}_{r_n}(\mathbb{X}_n) \approx S$ *and* $\partial\hat{S}_{r_n}(\mathbb{X}_n) \approx \partial S$ *e.a.s.*



New results on density and level set estimation

New notations:

1. N_{x,r_n} is the number of X_i that are in $B(x, r_n)$
2. N^{∂}_{x,r_n} is the number of X_i that are in $\partial C_{x,r_n}$
3. N^{o}_{x,r_n} is the number of X_i that are in $\mathring{C}_{x,r_n}$

Density estimator(s)

$$\hat{f}_{r_n,A}(x) = \frac{N^{o}_{x,r_n}}{(n - N^{\partial}_{x,r_n})|C_{x,r_n}|}\mathbb{I}_{|C_{x,r_n}|\geq A\omega_d r_n^d}\mathbb{I}_{N^{\partial}_{x,r_n}\leq n/2}, \quad (1)$$

$$\hat{f}_{r_n,A,S}(x) = \frac{N^{o}_{x,r_n}}{(n - N^{\partial}_{x,r_n})|C_{x,r_n}|}\mathbb{I}_{|C_{x,r_n}|\geq A\omega_d r_n^d}\mathbb{I}_{N^{\partial}_{x,r_n}\leq n/2}\mathbb{I}_S(x), \quad (2)$$

$$\hat{f}_{r_n,A,\hat{S}}(x) = \frac{N^{o}_{x,r_n}}{(n - N^{\partial}_{x,r_n})|C_{x,r_n}|}\mathbb{I}_{|C_{x,r_n}|\geq A\omega_d r_n^d}\mathbb{I}_{N^{\partial}_{x,r_n}\leq n/2}\mathbb{I}_{\hat{S}}(x). \quad (3)$$

The correction by N^{∂}_{x,r_n} is a consequence of Efron (1965) then Baldin and Reiss (2015) work

Hypotheses

Let $X_1, \ldots, X_n$ be iid, in $\mathbb{R}^d$ drawn with a density f bounded bellow by a positive f_0 and $\mathcal{C}^2$ on S the support of the distribution that is supposed to be compact and to satisfy the inside and outside (r_0)-rolling ball property.

This hypotheses can be relax for the pointwise $L2$ result but not for the pointwise probabilistic result

Pointwise $L2$ result

Theorem (A. and Fraiman (?))

if $r_n = cn^{-1/(d+4)}$ for all $x \in S$ we have that

$$\mathbb{E}(\hat{f}_{r_n,A}(x) - f(x))^2 = O_{\mathcal{M}}(n^{-2/(d+4)})$$

If $d \leq 7$ for all $x \in \mathring{S}$, when n is large enough

$$\mathbb{E}(\hat{f}_{r_n,A}(x) - f(x))^2 \leq O_{\mathcal{M},x}(n^{-4/(d+4)})$$

Pointwise probabilistic result

Theorem (A. and Fraiman (?))

if $r_n = cn^{-1/(d+4)}$ for all $x \in S$ we have that for n large enough

$$\mathbb{P}\left(|\hat{f}_{r_n,A}(x) - f(x)| \geq c_1 \ln n n^{-1/(d+4)}\right) \leq 3n^{-2.4}$$

When $d \leq 7$ for all $x \in S \setminus rB$ we have that for n is large enough

$$\mathbb{P}\left(|\hat{f}_{r_n,A}(x) - f(x)| \geq c_2 \ln n n^{-2/(d+4)}\right) \leq 3n^{-2.4}$$

Uniform on X_i result

Corollary

if $r_n = cn^{-1/(d+4)}$ we have that for n large enough

$$\max_i |\hat{f}_{r_n,A}(X_i) - f(X_i)| \leq c_1 \ln n.n^{-1/(d+4)} \ e.a.s.$$

And, when $d \leq 7$

$$\max_{i, X_i \in S \ominus rB} |\hat{f}_{r_n,A}(X_i) - f(X_i)| \leq c_1 \ln n.n^{-2/(d+4)} \ e.a.s.$$

Sketch of proof

$$\hat{f}_{r_n,A}(x) = \frac{N^o_{x,r_n}}{(n - N^\partial_{x,r_n})|C_{x,r_n}|}\mathbb{I}_{|C_{x,r_n}|\geq A\omega_d r_n^d}\mathbb{I}_{N^\partial_{x,r_n}\leq n/2}$$

First conditions on S and $f \geq f_0$ allows to have the following bounds $\mathbb{P}(|C_{r,x}| \leq A\omega_d r_n^d)$ (classical technics in set estimation) and $\mathbb{P}(N^\partial_{x,r_n} \geq n/2)$ (with Hoeffding inequality)

Sketch of proof

Introduce $\tilde{\Gamma}_{x,r_n} = \int_{C_{x,r_n}} f(z)dz$

$$\hat{f}_{r_n,A}(x) - f(x) = \frac{1}{|C_{x,r_n}|(n - N^{\partial}_{x,r_n})}\left(N^{o}_{x,r_n} - \tilde{\Gamma}_{x,r_n}(n - N^{\partial}_{x,r_n})\right)$$

$$+\left(\frac{\tilde{\Gamma}_{x,r_n}}{|C_{x,r_n}|} - f(x)\right)$$

$\varepsilon_1 = \frac{1}{|C_{x,r_n}|(n-N^{\partial}_{x,r_n})}\left(N^{o}_{x,r_n} - \tilde{\Gamma}_{x,r_n}(n - N^{\partial}_{x,r_n})\right)$

$N^{o}_{x,r_n}|C_{x,r_n} \sim Binom(n - N^{\partial}_{x,r_n}, \tilde{\Gamma}_{x,r_n})$ (cf Baldin and Reiss 2015)

Thus $\mathbb{E}(\varepsilon_1^2\,|C_{x,r_n}) \leq \frac{\Gamma_{x,r_n}}{|C_{x,r_n}|^2(n-N^{\partial}_{x,r_n})}$

Use Bennett inequality to give bound on $\mathbb{P}(\varepsilon_1 \geq t|C_{r,x})$

Sketch of proof

Introduce $\tilde{\Gamma}_{x,r_n} = \int_{C_{x,r_n}} f(z)dz$

$$\hat{f}_{r_n,A}(x) - f(x) = \frac{1}{|C_{x,r_n}|(n - N^{\partial}_{x,r_n})} \left(N^{o}_{x,r_n} - \tilde{\Gamma}_{x,r_n}(n - N^{\partial}_{x,r_n}) \right)$$

$$+ \left(\frac{\tilde{\Gamma}_{x,r_n}}{|C_{x,r_n}|} - f(x) \right)$$

$\varepsilon_2 = \frac{\int_{C_{x,r_n}} f(z)dz - \int_{C_{x,r_n}} f(x)dz}{|C_{x,r_n}|}$ for all x : $\varepsilon_1 = O(r)$

If $B(x, r_n) \subset S$ then $\varepsilon_2 = \frac{\int_{B(x,r_n)} (f(z)-f(x))dz - \int_{B(x,r_n)\setminus C_{x,r_n}} (f(z)-f(x))dz}{|C_{x,r_n}|}$

$\int_{B(x,r_n)} (f(z) - f(x))dz = O(r^2)$ (Taylor expension, deterministic) and bounds on $\int_{B(x,r_n)\setminus C_{x,r_n}} (f(z) - f(x))dz$ (moments and proba) are given by V.E. Brunel (2017)

Level set estimation

We now aim at estimating $L_t = \{x, f(x) \geq t\}$.
In all the following we will suppose that the level set is γ_0 regular that is, for some $\varepsilon_0 > 0$ for all t' such that $|t - t'| \leq \varepsilon_0$, $d_h(L_t, L_{t'}) \leq \gamma_0 |t - t'|$ (morally $\Delta f(x) \geq \gamma_0 > 0$ for all $x \in \partial L_t$) and $L_{t'}$ satisfies the r_0 inside and outside rolling ball condition (see Rodriguez Casal (2019) and Walther (1997) for some discussion and sufficient conditions)

Estimator

$$\mathbb{X}_{t,n,r_n,A} = \{X_i, \hat{f}_{r_n,A}(X_i) \geq t\}$$

$$\hat{L}_{r_n,A}(t) = \bigcup_{X_i \in \mathbb{X}_{t,n,r,A}} H(B(X_i, r_n) \cap \mathbb{X}_{t,n,r,A})$$

The same r_n for the density estimation and the Level set estimation
It is not the original estimator which is

$$\tilde{L}_{r_n,A}(t) = \bigcup_{X_i \in \mathbb{X}_{t,n,r_n,A}} H(B(X_i, r_n) \cap \mathbb{X}_n)$$

Results

Next results are under the asumption of all the previous theorems and for a regular level set

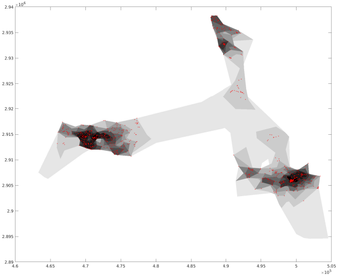
Theorem

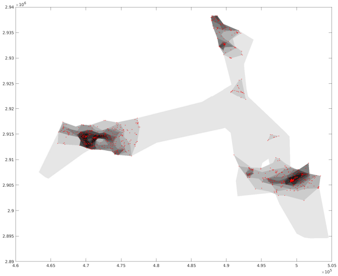
If $L_t \subset S \ominus r_n B$ *then* $d_H(L_t, \hat{L}_{r_n,A}(t)) = O(n^{-2/(d+4)})$

Theorem

$d_H(L_t, \hat{L}_{r_n,A}(t)) = O(n^{-1/(d+4)})$

Illustration (Panther Jitter data Pennstates university)





Further developpements

1. if $d_H(L_t, \hat{L}_{r_n,A}(t)) = O(n^{-1/(d+4)})$ we nevertheless have $d_\Delta(L_t, \hat{L}_{r_n,A}(t))$ that is much smaller
2. k-NN instead of fixed radius
3. Density on manifold

Bibliogarphy

1. Baldin N and Reiss M. **Unbiased estimation of the volume of a convex body** *Stochastic Processes and their Applications* 126(12) (2015)
2. Brunel V.E. **Uniform deviation and moment inequalities for random polytopes with general densities in arbitrary convex bodies** *arXiv:1704.01620*
3. Efron B. **The Convex Hull of a Random Set of Points** *Biometrika* Vol. 52, No. 3/4 (Dec., 1965)
4. Walther G. **Granulometric smoothing** *Ann. Statist.* 25(6): 2273-2299 (December 1997)