Statistic on Manifolds

On boundary detection

Alejandro Cholaquidis[a]

A joint work with: C. Aaron[b]

[a] CMAT - Facultad de Ciencias, UdelaR,
[b] Université Blaise-Pascal Clermont II, France.

Seminario de Probabilidad y Estadística - CMAT

Densities on Manifolds.

General assumption: $\mathcal{M} \subset \mathbb{R}^d$ is a $\mathcal{C}^\infty$, compact, d'-dimensional manifold, with the metric inherit from $\mathbb{R}^d$.

Densities on Manifolds.

General assumption: $\mathcal{M} \subset \mathbb{R}^d$ is a $\mathcal{C}^\infty$, compact, d'-dimensional manifold, with the metric inherit from $\mathbb{R}^d$.

Densities on manifolds.

Let P a probability on $\mathcal{B}(\mathcal{M})$, a random variable on $\mathcal{M}$ is a measurable function $X : \Omega \to \mathcal{M}$. If $\mathcal{M}$ is orientable a density is $f : \mathcal{M} \to \mathbb{R}^+$ which fulfils,

$$P(B) = \int_B f(x)dv(x) \quad \text{being } dv \text{ the volume form.}$$

Another option is to integrate w.r.t. the d'-dimensional Hausdorff measure.

Densities on Manifolds.

General assumption: $\mathcal{M} \subset \mathbb{R}^d$ is a $\mathcal{C}^\infty$, compact, d'-dimensional manifold, with the metric inherit from $\mathbb{R}^d$.

Densities on manifolds.

Let P a probability on $\mathcal{B}(\mathcal{M})$, a random variable on $\mathcal{M}$ is a measurable function $X : \Omega \to \mathcal{M}$. If $\mathcal{M}$ is orientable a density is $f : \mathcal{M} \to \mathbb{R}^+$ which fulfils,

$$P(B) = \int_B f(x)dv(x) \quad \text{being } dv \text{ the volume form.}$$

Another option is to integrate w.r.t. the d'-dimensional Hausdorff measure.

In local coordinates

$$\int_U f dv = \int_{\varphi(U)} f(\varphi^{-1}(x))\sqrt{\det g_{ij}(\varphi^{-1}(x))}dx$$

where g_{ij} are the coefficients of the metric g in the local coordinates (U, φ).

Variance and Expectation

Variance

Let $y \in \mathcal{M}$, and X a r.v on $\mathcal{M}$ with density f, the variance on y, $\sigma_X(y)^2$ is

$$\mathbb{E}(d(y,X)^2) = \int_{\mathcal{M}} d(y,z)^2 f(z) dv(z) \quad \text{being } d \text{ the geodesic distance.}$$

Variance and Expectation

Variance

Let $y \in \mathcal{M}$, and X a r.v on $\mathcal{M}$ with density f, the variance on y, $\sigma_X(y)^2$ is

$$\mathbb{E}(d(y,X)^2) = \int_{\mathcal{M}} d(y,z)^2 f(z)dv(z) \quad \text{being } d \text{ the geodesic distance.}$$

Expectation

If $\sigma_X(y)^2 < \infty$ for all y, the set (possibly empty) of expectations is

$$\mathbb{E}(X) = \text{argmin}_{y\in\mathcal{M}} \sigma_X(y)^2.$$

Kendall, 1990: if $supp(f) \subset B_d(x,r)$ for some regular geodesic ball that does not meet the cutlocus of x, exists a unique $\mathbb{E}(X)$.

Manifold Recovery from a sample of points, filament model

The filament model: $X_i = f(U_i) + Z_i$, where $f : [0,1] \to \mathbb{R}^d$, the U_i are uniform and the Z_i are zero-mean compact supported; Genovese et al. (2012a).

INPUT: $\hat{S}$ and $\hat{\partial S}$ D.W. of radius $\varepsilon > 0$.

OUTPUT: $\hat{\Gamma}$.

ALGORITHM:

1) Compute $\hat{\Delta}(y) = d(y, \hat{\partial S})$ for all $y \in \hat{S}$.
2) $\hat{\sigma} = \max_{y \in \hat{S}} \hat{\Delta}(y)$
3) $\delta = 2\varepsilon$, $\hat{\Gamma} = \{y \in \hat{S} : d(y, \hat{\partial S}) \geq \hat{\sigma} - \delta\}$

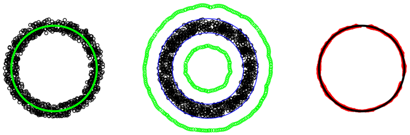

FIG 6. *These plots illustrate the EDT-based estimator. Left: filament and data. Center: Estimated boundary. Right: EDT estimator* $\hat{\Gamma}$.

Inference on the dimension:

Testing the manifold hypothesis, *Fefferman et al 2015*

$\mathcal{G}(d, V, \tau)$: d dimensional $\mathcal{C}^2$ submanifolds of the unit ball in $\mathcal{H}$ a separable Hilbert space, with volume $\leq V$ and reach $\leq \tau < 1$. P a probability with support $B(0, 1)$. **The problem:** decide from a sample of P if there exists $\mathcal{M} \in \mathcal{G}(d, CV, \tau/C)$ such that

$$\int d(M, x)^2 dP(x) < C\varepsilon$$

Inference on the dimension:

Testing the manifold hypothesis, *Fefferman et al 2015*

$\mathcal{G}(d, V, \tau)$: d dimensional $\mathcal{C}^2$ submanifolds of the unit ball in $\mathcal{H}$ a separable Hilbert space, with volume $\leq V$ and reach $\leq \tau < 1$. P a probability with support $B(0,1)$. **The problem:** decide from a sample of P if there exists $\mathcal{M} \in \mathcal{G}(d, CV, \tau/C)$ such that

$$\int d(M,x)^2 dP(x) < C\varepsilon$$

Finding the Homology of Submanifolds, *Smale et al 2008*

Theorem: Let $\mathcal{M} \subset \mathbb{R}^n$ compact with reach τ. $\aleph_n = X_1, \ldots, X_n$ iid uniform on $\mathcal{M}$. Let $0 < \varepsilon < \tau/2$ and $U = \cup_i B(X_i, \varepsilon)$. Then for $n = n(\delta)$, with probability greater than the homology of U equals the homology of $\mathcal{M}$ with probability $> 1 - \delta$.

Estimation of the dimension:

MLE of Intrinsic Dimension, *Bickel and Levina 2005*

Heuristic: $X_1, \ldots, X_n$ iid in $\mathbb{R}^p$, $X_i = g(Y_i)$ where Y_i are sampled from a unknown density f on $\mathbb{R}^m$ and $f(x)$ ∼constant on $B(x,R)$ for some R, with $m \leq p$ and g is *smooth*. If we consider the process

$$N(t,x) = \sum_{i=1}^{n} \mathbb{I}_{\{X_i \in B(x,t)\}} \sim Poisson(\lambda(t)) \quad 0 \leq t \leq R,$$

with $\lambda(t) = f(x) Vol(B_m(0,1)) m t^{m-1}$.
If $\theta = \log(f(x))$, the log-likelihood of $N(t)$ is

$$L(m,\theta) = \int_0^R \log(\lambda(t)) dN(t) - \int_0^R \lambda(t) dt,$$

then the MLE for m is $\hat{m}_k(x) = \left[\frac{1}{k-1} \sum_{j=1}^{k-1} \log \frac{d(x, X_k(x))}{d(x, X_j(x))} \right]^{-1}.$

Models

- *Noiseless model*: the data $X_1, \dots, X_n$ are taken from a distribution whose support is the manifold $\mathcal{M}$; Aamari and Levrard (2015), Amenta et al. (2002).

Models

- *Noiseless model*: the data $X_1, \ldots, X_n$ are taken from a distribution whose support is the manifold $\mathcal{M}$; Aamari and Levrard (2015), Amenta et al. (2002).
- *The clutter noise model*: $X_i \sim (1-\pi)U + \pi G$, where U is a uniform distribution on a compact set $K \subset \mathbb{R}^d$ with nonempty interior, and G is supported on $\mathcal{M}$; Genovese et al (2012c).

Models

- *Noiseless model*: the data $X_1, \ldots, X_n$ are taken from a distribution whose support is the manifold $\mathcal{M}$; Aamari and Levrard (2015), Amenta et al. (2002).
- *The clutter noise model*: $X_i \sim (1-\pi)U + \pi G$, where U is a uniform distribution on a compact set $K \subset \mathbb{R}^d$ with nonempty interior, and G is supported on $\mathcal{M}$; Genovese et al (2012c).
- *The additive noise model*: $X_i = Y_i + Z_i$, where the Y_i are supported on $\mathcal{M}$ and $Z_i|Y_i$ is uniform on a segment orthogonal to $\mathcal{M}$ on Y_i; Genovese et al. (2012b).

Models

- *Noiseless model*: the data $X_1, \dots, X_n$ are taken from a distribution whose support is the manifold $\mathcal{M}$; Aamari and Levrard (2015), Amenta et al. (2002).
- *The clutter noise model*: $X_i \sim (1-\pi)U + \pi G$, where U is a uniform distribution on a compact set $K \subset \mathbb{R}^d$ with nonempty interior, and G is supported on $\mathcal{M}$; Genovese et al (2012c).
- *The additive noise model*: $X_i = Y_i + Z_i$, where the Y_i are supported on $\mathcal{M}$ and $Z_i|Y_i$ is uniform on a segment orthogonal to $\mathcal{M}$ on Y_i; Genovese et al. (2012b).
- *The parallel model*: The X_i have a distribution whose support is the parallel set $B(\mathcal{M}, r)$; Berrendero et al. (2014).

Open problems

- Things to define:
  - depths.
  - outlier.
  - classical distributions.
- To estimate/test:
  - Positive reach/condition number
  - Is $\mathcal{M}$ orientable?
  - It has empty interior? In general, detect a lower (non-linear) dimensional structure.
  - Estimate $\mu_{d'}(\partial M)$ being $\mu_{d'}$ de d' Lebesgue measure.

Model and problem

The model and the problem

We will assume the noisless model wit $f > f_0 > 0$ Lipschitz. The problem

$$\begin{cases} H_0 : & \partial M = \emptyset \\ H_1 : & \partial M \neq \emptyset \end{cases}$$

Model and problem

The model and the problem

We will assume the noisless model wit $f > f_0 > 0$ Lipschitz. The problem

$$\begin{cases} H_0 : & \partial M = \emptyset \\ H_1 : & \partial M \neq \emptyset \end{cases}$$

Why not just estimate the manifold?

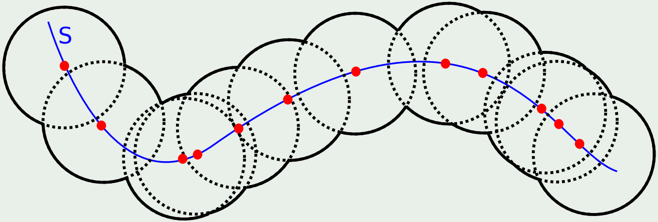

Figure: The boundary of D.W. estimator is not a good estimation of ∂M

The heuristic idea

$\mathcal{X}_{k_n,x} = \{X_{1(x)}, \dots, X_{k_n(x)}\}$; $r_{x,n} = \max_{y \in \mathcal{X}_{k_n,x}} \|y - x\|$; $\overline{X}_{x,k_n} = \frac{1}{k_n} \sum_{k=1}^{k_n} X_{k(x)}$.
Assume that $k_n \to +\infty$ slowly enough to have $\max_{x \in S} r_{x,n} \stackrel{a.s.}{\to} 0$.

If $\partial M = \emptyset$

$\left\{(X_{1(x)} - x)/r_{x,n} \dots (X_{k_n(x)} - x)/r_{x,n}\right\}$ is “close” to a sample uniformly distributed on $B(x,1) \subset \mathbb{R}^{d'}$ with $d' = dim(\mathcal{M})$.
As $k_n \to \infty$ we expect $\|\overline{X}_{x,k_n} - x\| \stackrel{a.s.}{\to} 0$, then $\max_i \|\overline{X}_{X_i,k_n} - x\| \stackrel{a.s.}{\to} 0$.

The heuristic idea

$\mathcal{X}_{k_n,x} = \{X_{1(x)}, \ldots, X_{k_n(x)}\}$; $r_{x,n} = \max_{y \in \mathcal{X}_{k_n,x}} \|y - x\|$; $\overline{X}_{x,k_n} = \frac{1}{k_n} \sum_{k=1}^{k_n} X_{k(x)}$.
Assume that $k_n \to +\infty$ slowly enough to have $\max_{x \in S} r_{x,n} \overset{a.s.}{\to} 0$.

If $\partial M = \emptyset$

$\left\{(X_{1(x)} - x)/r_{x,n} \ldots (X_{k_n(x)} - x)/r_{x,n}\right\}$ is "close" to a sample uniformly distributed on $B(x,1) \subset \mathbb{R}^{d'}$ with $d' = dim(\mathcal{M})$.
As $k_n \to \infty$ we expect $\|\overline{X}_{x,k_n} - x\| \overset{a.s.}{\to} 0$, then $\max_i \|\overline{X}_{X_i,k_n} - x\| \overset{a.s.}{\to} 0$.

If ∂M is a $\mathcal{C}^2$ manifold

If $x \in \partial M$, the "locally rescaled sample" is close to sample on a half unit ball and $\|\overline{X}_{x,k_n} - x\| \to \alpha_{d'}$ with $\alpha_{d'}$ a positive constant. Then $\max_i \|\overline{X}_{X_i,k_n} - x\| \overset{a.s.}{\to} \alpha_{d'}$.

We decide $\partial M = \emptyset$ if $\max_i \|\overline{X}_{X_i,k_n} - x\|$ is small enough.

Some definitions.

Definition

Let us define: $r_{i,k_n} = \|X_i - X_{k_n(i)}\|$; $r_n = \max_{i \leq n} r_{i,k_n}$

$$\mathcal{X}_{i,k_n} = \begin{pmatrix} X_{1(i)} - X_i \\ \vdots \\ X_{k_n(i)} - X_i \end{pmatrix}; \quad \hat{S}_{i,k_n} = \frac{1}{k_n}(\mathcal{X}_{i,k_n})'(\mathcal{X}_{i,k_n}).$$

- Q_{i,k_n} is the plane spanned by the d' eigenvectors of $\hat{S}_{i,k_n}$ associated to the d' largest eigenvalues.

Some definitions.

Definition

Let us define: $r_{i,k_n} = \|X_i - X_{k_n(i)}\|$; $r_n = \max_{i \le n} r_{i,k_n}$

$$\mathcal{X}_{i,k_n} = \begin{pmatrix} X_{1(i)} - X_i \\ \vdots \\ X_{k_n(i)} - X_i \end{pmatrix} ; \quad \hat{S}_{i,k_n} = \frac{1}{k_n}(\mathcal{X}_{i,k_n})'(\mathcal{X}_{i,k_n}).$$

- Q_{i,k_n} is the plane spanned by the d' eigenvectors of $\hat{S}_{i,k_n}$ associated to the d' largest eigenvalues.
- $X^*_{k(i)}$ the normal projection of $X_{k(i)} - X_i$ on Q_{i,k_n} and $\overline{X}_{k_n,i} = \frac{1}{k_n} \sum_{j=1}^{k_n} X^*_{j(i)}$.

Some definitions.

Definition

Let us define: $r_{i,k_n} = \|X_i - X_{k_n(i)}\|$; $r_n = \max_{i \leq n} r_{i,k_n}$

$$\mathcal{X}_{i,k_n} = \begin{pmatrix} X_{1(i)} - X_i \\ \vdots \\ X_{k_n(i)} - X_i \end{pmatrix}; \quad \hat{S}_{i,k_n} = \frac{1}{k_n}(\mathcal{X}_{i,k_n})'(\mathcal{X}_{i,k_n}).$$

- Q_{i,k_n} is the plane spanned by the d' eigenvectors of $\hat{S}_{i,k_n}$ associated to the d' largest eigenvalues.
- $X^*_{k(i)}$ the normal projection of $X_{k(i)} - X_i$ on Q_{i,k_n} and $\overline{X}_{k_n,i} = \frac{1}{k_n}\sum_{j=1}^{k_n} X^*_{j(i)}$.
- $\delta_{i,k_n} = \frac{(d'+2)k_n}{r^2_{i,k_n}}\|\overline{X}_{k_n,i}\|^2$, for $i = 1, \ldots, n$.

The proposed test statistic is: $\Delta_{n,k_n} = \max_i \delta_{i,k_n}$.

Some results

We will denote by $\Psi_{d'}(t)$ the cumulative distribution function of a $\chi^2(d')$ distribution and $F_{d'}(t) = 1 - \Psi_{d'}(t)$.

Theorem

$\mathcal{M}$ is $\mathcal{C}^2$, compact, the density f is Lipschitz and $f(x) > f_0$ on $\mathcal{M}$. $\partial\mathcal{M} = \emptyset$ or $\mathcal{C}^2$. If $k_n/(\ln(n))^4 \to \infty$ and $(\ln(n))k_n^{1+d'}/n \to 0$, the test

$$\begin{cases} H_0: & \partial M = \emptyset \\ H_1: & \partial M \neq \emptyset \end{cases} \tag{1}$$

with the rejection zone

$$W_n = \left\{\Delta_{n,k_n} \geq F_{d'}^{-1}(9\alpha/(2e^3 n))\right\}, \tag{2}$$

fulfills: $\mathbb{P}_{H_0}(W_n) \leq \alpha + o(1)$. The test (1) with rejection zone (2) has power 1 for n large enough.

More Results

Theorem

Under previous conditions, if we define

$$\hat{\Psi}_{n,k_n}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{\delta_{i,k_n} \leq x\}},$$

then, for all $x \in M$,

$$\mathbb{E}\left(\hat{\Psi}_{n,k_n}(x) - \Psi_{d'}(x)\right)^2 \to 0 \quad \text{as } n \to \infty.$$

Some probabilistic results for the proof

Lemma

Let $X_1, \ldots, X_n$ be an i.i.d. sample uniformly on $\mathcal{B}(x,r) \subset \mathbb{R}^d$. Let us denote $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then we have:

$$\frac{(d+2)n\|\overline{X}_n - x\|^2}{r^2} \xrightarrow{\mathcal{L}} \chi^2(d), \tag{3}$$

Some probabilistic results for the proof

Lemma

Let $X_1, \dots, X_n$ be an i.i.d. sample uniformly on $\mathcal{B}(x,r) \subset \mathbb{R}^d$. Let us denote $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, then we have:

$$\frac{(d+2)n\|\overline{X}_n - x\|^2}{r^2} \xrightarrow{\mathcal{L}} \chi^2(d), \tag{3}$$

Lemma

Let X be uniformly distributed on $\mathcal{B}_u(x,r) = \mathcal{B}(x,r) \cap \{z \in \mathbb{R}^d : \langle z - x, u\rangle \geq 0\}$ where u is a unit vector, then

$$\mathbb{E}\left(\frac{\langle X - x, u\rangle}{r}\right) = \alpha_d, \text{ where } \alpha_d = \left(\frac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\Gamma(\frac{d+3}{2})}\right). \tag{4}$$

Some key results for the proofs

Theorem

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact, d'-dimensional $\mathcal{C}^2$ manifold without boundary. X with Lipschitz density f. There exist positive constants R_1 and C_1 such that: if $r \leq R_1$, then $\left|\mathbb{P}_X(\mathcal{B}(x,r)) - f(x)\sigma_{d'} r^{d'}\right| \leq C_1 r^{d'+1}$, with $\sigma_{d'} = vol(\mathcal{B}_d(0,1))$

Some key results for the proofs

Theorem

Let $\mathcal{M} \subset \mathbb{R}^d$ be a compact, d'-dimensional $\mathcal{C}^2$ manifold without boundary. X with Lipschitz density f. There exist positive constants R_1 and C_1 such that: if $r \leq R_1$, then $\left|\mathbb{P}_X(\mathcal{B}(x,r)) - f(x)\sigma_{d'} r^{d'}\right| \leq C_1 r^{d'+1}$, with $\sigma_{d'} = vol(\mathcal{B}_d(0,1))$

Lemma

Let $X_1, \ldots, X_n$ be an i.i.d. of $\mathbb{P}_X$, with $\partial\mathcal{M} = \emptyset$. Then there exists a constant A_d such that

$$X^*_{k_n(i)} = (I_d + E_{i,n})\varphi_{X_i}(X_{k_n(i)}) - X_i \text{ with: } \max_i \|E_{i,n}\|_\infty \leq A_d \sqrt{\frac{\ln(n)}{k_n}} \text{ e.a.s.}$$

Simulations

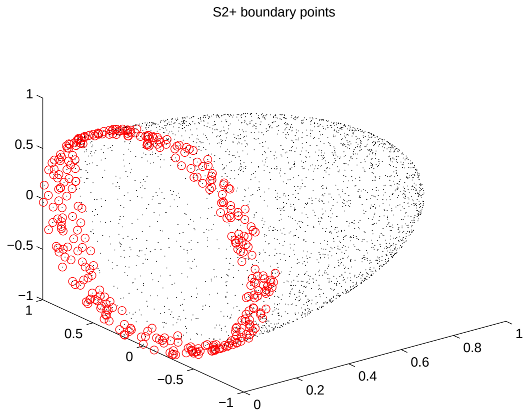

Figure: $n = 3000$ points, X_i, boundary point if $\frac{2e^3}{9} F_{d'}(\delta_{i,k}) \leq 5\%$

Simulations

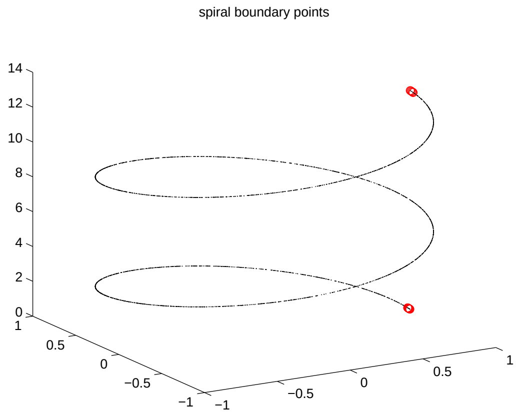

Figure: $n = 3000$ points, X_i, boundary point if $\frac{2e^3}{9} F_{d'}(\delta_{i,k}) \leq 5\%$

Simulations

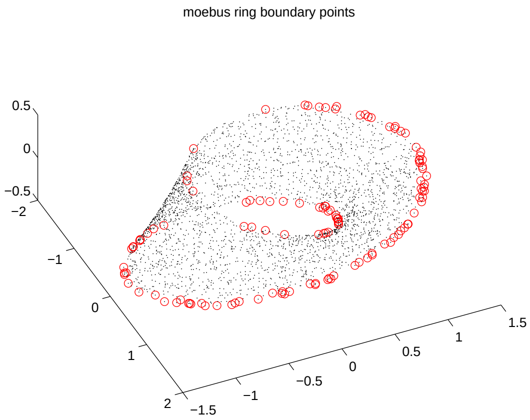

Figure: $n = 3000$ points, X_i, boundary point if $\frac{2e^3}{9} F_{d'}(\delta_{i,k}) \leq 5\%$

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

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.

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.

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

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

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

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.

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

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

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

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

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

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

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.

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*

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.

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

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.

Walther, G. (1997) Granulometric smoothing. *Ann. Statist.* **25** 2273–2299.

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