Convergence rate for the λ-Medial-Axis estimation under regularity conditions

Catherine Aaron

Convergence rate for the λ-Medial-Axis estimation under regularity conditions

Catherine Aaron
Laboratoire de Mathématiques Blaise Pascal
UMR6620-CNRS
Université Clermont Auvergne
63178 Aubière

July 4, 2019

Abstract

Let $\mathcal{X}_n = \{X_1, \dots X_n\} \subset \mathbb{R}^d$ be a iid random sample of observations drawn with a probability distribution supported by S a compact set satisfying that both S and $\overline{S^c}$ are r_0-convex ($r_0 > 0$). In this paper we study some properties of an estimator of the inner medial axis of S based on the λ-medial axis. The proposed estimator depends on the choices of $\mathcal{Y} \subset \mathcal{X}_n$ an estimator of ∂S and $\hat{S}_n$ an estimator of S. In a first general theorem we prove that our medial axis estimator converges to the medial axis with a rate $O(\max_{y\in\mathcal{Y}} d(y, \partial S), (\max_{y\in\partial S} d(y, \mathcal{Y})^2)$. A corollary being that the choice of $\mathcal{Y}$ as the intersection of the sample and its r-convex hull, $\mathcal{Y} = C_r(\mathcal{X}_n) \cap \mathcal{X}_n$, allows to estimate the medial axis with a convergence rate $O((\ln n/n)^{2/(d+1)})$. In a practical point of view, computational aspects are discussed, algorithms are given and a way to tune the parameters is proposed. A small simulation study is performed to illustrate the results.

Keywords: Geometric Inference, Medial-Axis, Skeleton, r_0-convexity

1 Introduction

Let $S \subset \mathbb{R}^d$ be a compact set, its medial axis, introduced in [6] as the set of points in $\mathbb{R}^d$ that has at least two different projections on ∂S (see Figure 1) has been initially

proposed as a tool for biological shape recognition. Note that the medial axis can be decompose into two parts: its inner part, that is $\mathcal{M}(S) \cap S$ and its outer part that is $\mathcal{M}(S) \cap S^c$ (see Figure 1).

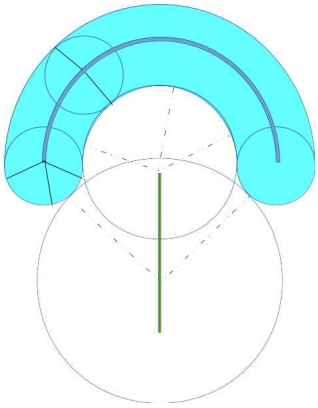

Figure 1: A set, its inner medial axis (blue) and its outer medial axis (green). Some points of the medial axis are presented together with some of there projection onto the boundary

When dealing with compact sets we can only focus on $\mathcal{M}(S)$ the inner part of the medial axis.

Definition 1.

$$\mathcal{M}(S) = \{x \in S, diam(\Gamma(x)) > 0\}$$

where $\Gamma(x) = \{y \in \partial S, ||y-x|| = d(x, \partial S)\}$ *and* $diam(A) = \max\{||x-y||, (x,y) \in A^2\}$.

When S is compact the knowledge of S is equivalent to the knowledge of medial axis transform that is the medial axis and the function $r(x) = d(x, \partial S)$ because we have:

$$S = \bigcup_{x \in \mathcal{M}(S)} \overline{B}(x, r(x)) \tag{1}$$

with $\overline{B}(x, r)$ the closed ball centered in x and of radius r.

The main field of application of this tool is image analysis and compression (see [15] for instance). Roughly speaking: because $\mathcal{M}(S)$ has a lower dimension than S the storage of $(x, r(x))$ is lighter than the one of S. The summary of the shape contained in $\mathcal{M}(S)$ is also useful for shape analysis and pattern recognition.

This two image analysis applications can be transposed to a statistical point of view. The analogous of the image compression being support estimation and the analogous of pattern recognition being the manifold estimation.

Support Estimation: Suppose that we are able to obtain $\mathcal{Z} = \{Z_1, \dots Z_k\}$ a set of points that estimates the medial axis and that we can estimate $d(Z_i, \partial S)$ by some $\hat{r}_i$ then $\hat{S} = \cup_i \overline{B}(Z_i, \hat{r}_i)$ provides a natural estimator of the support. We are going to see, in the Application section, that we can find $\mathcal{Z}$ and $\hat{r}$ such that $\hat{S}$ has minimax convergence rate (under some shape and distribution hypotheses), as the r-convex hull [23] or [1] but with $\hat{S}$ it is computationally much more easy to decide whether a new points belongs to the support or not.

Manifold estimation: Suppose that data are, in fact drawn with the following process. Y is drawn on M a compact $d' < d$ dimensional manifold but we observe X a noisy version of Y with a noise $U|Y$ that is uniform on a ball and we aim to estimate M. See [16] for a study in the "filament case" (that is $d' = 1$) and [17] for the derivation of minimax bounds for M estimation. With infinitely many observations, the manifold estimation problem can be seen as the problem of estimating M through the observation of $S = \cup_{x \in M} \overline{B}(x, r_x)$ with $r_x = d(x, \partial S)$ (indeed if $r_x < d(x, \partial S)$ the problem appear to have infinitely many solution and so is not well-posed). It appears to be very close to the medial axis transform equation (1) and, in Section Applications we are going to give some condition that ensures $M = \mathcal{M}(\cup_{x \in M} \overline{B}(x, r_x))$ so that the manifold estimation problem can be solved using the medial axis estimation.

More recently some more new application of the medial axis appears (for instance in [29] it is applied to wireless networks, In [21] it was applied to endoscopy as the medial axis naturally find a path along the central line of the intestinal system).

Unfortunately, the medial axis is difficult to estimate because it is not continuous with respect to the Hausdorff distance d_h (recall that for A and B two sets $d_h(A, B) = \max\{\sup_{a \in A} d(a, B), \sup_{b \in B} d(b, A)\}$). This is detailed in [19] (see pages $217 - 238$) and illustrated in Figure 2 part a)). This implies that estimating the medial axis using a finite sample of points $\mathcal{X}_n = \{X_1, \dots, X_n\}$ can not be solved using classical plug-in methods (see Figure 2 part b)) and so provides a challenging problem that has been investigated in various papers (see [4] for a state-of-the-art report).

Mainly two different approaches have been investigated. The first one consists in pruning the medial axis of an estimation of S (see [25], [7], [10],[5] or [20]); the second one consists in estimating the λ-medial axis defined as $\mathcal{M}_\lambda(S) = \{x \in \mathcal{M}(S), \Gamma(x) \subset$

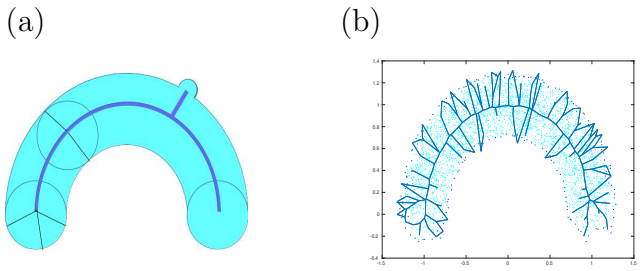

Figure 2: Two sets close to S the one of Figure 2 and their medial axis. (a) $S \cup \overline{B}(x, r_0)$ with $x \in \partial S$: a parasite branch appear whatever is the value of r_0 that illustrates the non continuity of the medial axis with regard to the Hausdorff distance. (b) plug-in estimator of the medial axis computed on a sample points there exist a lot of parasite branches

$\mathcal{B}(a, r) \Rightarrow r \geq \lambda\}$ instead of the medial axis. The λ-media axis has been introduce and studied in [8] where it has been proved to be stable with respect to the Hausdorff distance. More precisely the Authors prove that, if $d_h(S'^c, S^c) = O(\varepsilon)$ then $d_h(\mathcal{M}_\lambda(S), \mathcal{M}_\lambda(S')) = O(\sqrt{\varepsilon})$, then they propose an algorithm to estimate the λ-medial axis given sample points located near the boundary and prove that it converges.

Later on, given a sample point $\mathcal{X}_n$ drawn on S (instead of "near ∂S"), it is proved in [11], under no more shape hypothesis than regularity, that given a support estimator $\hat{S}_n$ such that $d_h(\hat{S}_n, S) \to 0$ a.s. and $d_h(\partial\hat{S}_n, \partial S) \to 0$ a.s. then $d_h(\mathcal{M}_\lambda(\hat{S}_n), \mathcal{M}_\lambda(S)) \to 0$ a.s.

We are going to introduce a new medial axis estimator $\hat{\mathcal{M}}_\lambda(\mathcal{X}_n)$ that is morally very close to the one introduced in [8]. We give conditions on S so that, on one hand we can derive the convergence rates and on the other hand we have convergence toward the medial axis (because, under such hypothesis, the medial axis is the λ-medial axis)

Namely, introduce $\mathcal{M}'_\lambda(S) = \{x \in S, \text{diam}\Gamma(x) > \lambda\}$ instead of λ-medial axis (that looks a bit more natural with regard to definition (1)). Notice that we can also write $\mathcal{M}'_\lambda(S) = \{x \in S, \exists (y, z) \in \partial S^2, d(x, y) = d(x, z) = d(z, \partial S), ||y - z|| > \lambda\}$ so that, if $\hat{S}_n$ is a support estimator and if $\mathcal{Y} = \{Y_1, \ldots Y_N\} \subset \mathcal{X}_n$ is an estimator of ∂S the plug-in estimator or $\mathcal{M}'_\lambda$ is:

$$\hat{\mathcal{M}}_\lambda(S_n, \mathcal{Y}) = \left\{x \in \text{Vor}_{\mathcal{Y}}(y) \cap \text{Vor}_{\mathcal{Y}}(z) \cap \hat{S}_n, (y, z) \in \mathcal{Y}^2, ||y - z|| > \lambda\right\} \qquad (2)$$

where $\text{Vor}_{\mathcal{Y}}(y) = \{z, ||z - y|| \leq d(z, \mathcal{Y})\}$ is the Voronoi cell of y with respect to the set $\mathcal{Y}$.

In a first section we detail the shape hypotheses made on S. then we give the principal theoretical results, the first theorem being a general and deterministic theorem that we applied to the medial axis estimator of type (2) where $\mathcal{Y}$ is obtained by intersection of the observations and there $r-$convex hull and where $\hat{S}_n$ is the Devroye-Wise estimator of S with a well chosen radius. Discussion on application for set estimation and manifold estimation is provided then we give the proofs of the results.

A second section focuses on the practical aspects. First we discuss the algorithmic point of view and provide the algorithm. We also propose a way to tune the parameters. And finally we provide a small simulation study.

2 Shape Hypothesis and main results

2.1 Shape Hypothesis

As mentioned in the introduction the medial axis is not continuous with regard to the Hausdorff distance. This morally implies that the plug-in estimator of the medial axis: $\hat{\mathcal{M}}_0(S_n, \mathcal{Y})$ has some "parasite branch", that are expected to be suppressed considering $\hat{\mathcal{M}}_\lambda(S_n, \mathcal{Y})$. When the support has "corners" there is a conflict between the recognition of the branches induced by the corners and the attempt to erase the parasite branches. See Figure 3 to observe that phenomena.

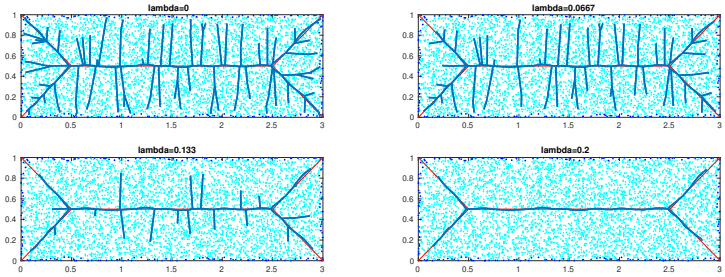

Figure 3: $S = [0; 3] \times [0; 1]$, $\mathcal{M}(S)$ is the union of 5 segments represented in thin red lines, we plot the estimated λ medial axis for some values of λ. If $\lambda = 0$ the corners belong to the estimated medial axis but there is a lot of parasite branches, when $\lambda > 0$ there is no points in the estimated medial axis closer than λ to the corner but the central part of the estimated medial axis looks good

To avoid the corner situation we will requires a rolling ball type condition on S, here and in all the following $B(x,r)$ denotes the open ball of radius r and centered at x.

Definition 2. *Balls of radius r_0 roll freely outside and inside S if, for each $x \in \partial S$ there exists O_x^{out} and O_x^{in} such that $B(O_x^{out}, r_0) \subset S^c$ and $B(O_x^{in}, r_0) \subset S$. In this case, we introduce $u_x = \frac{O_x^{out} - x}{r_0}$ the unit vector, normal to ∂S and pointing outward from S.*

According to Walther [28], it is equivalent to have balls of radius r_0 freely rolling inside and outside S and the r_0-convexity of S, $\overline{S^c}$ and $\mathring{S} \neq \emptyset$, it is also equivalent to ∂S is a $\mathcal{C}_1^1$ manifold. Nevertheless the ball vision of such a notion is the most helpful for the proofs as it allows to give geometric proofs based on euclidean geometry avoiding all the differential geometry tools.. If S is a compact set such that balls roll freely inside and outside S then it is regular enough to have a medial axis satisfying some good properties described in the following lemma.

Proposition 1. *If S is a compact set such that balls of radius r_0 roll freely inside and outside S then $S = \overline{(\mathring{S})}$ (one then says S is regular which is a common condition when considering the medial axis);*

Proof. The inclusion $\overline{(\mathring{S})} \subset S$ comes from the closeness of S. Considering the second inclusion $S \subset \overline{(\mathring{S})}$, for any $x \in S$, on one hand if $x \in \mathring{S}$ then $x \in \overline{(\mathring{S})}$; on the other hand if $x \in \partial S$, introduce $x_n = x - \frac{r_0}{2n} u_x$. The rolling ball property implies that $x_n \in \mathring{S}$ and as $x_n \to x$ we have $x \in \overline{(\mathring{S})}$. □

In addition to the regularity of the support we need additional assumptions to obtain our results. All the geometric assumptions made on S are listed in Definition 3

Definition 3. *Let $r_0 > 0$ and $K < 1$ be two numbers, S be a compact set in $\mathbb{R}^d$. We say S is (K, r_0)-regular if:*

1. *balls of radius r_0 roll freely inside and outside S;*
2. *$\mathcal{M}(S)$ is closed;*
3. *for all $(x,y) \in \mathcal{M}(S)^2$, $|d(x, \partial S) - d(y, \partial S)|/||x - y|| \leq K$*

The second assumption, that $\mathcal{M}(S)$ is closed, ensures that medial axis and skeleton are the same object. The skeleton, that can be defined, following [30] or [19], by the set of the centers of the maximal balls included in S. More precisely, if $B(x,r)$ denotes the open ball centered in x and of radius r and if $\mathring{S}$ denotes the interior of S the skeleton is defined by:

$$\mathcal{M}^*(S) = \{x, \exists r(x) \text{ such that } B(x,r(x)) \subset \mathring{S} \text{ and } B(x,r(x)) \subsetneq B(x',r') \Rightarrow B(x',r') \nsubseteq \mathring{S}\}. \tag{3}$$

It can be proved (see [19]) that $\mathcal{M}(S) \subset \mathcal{M}^*(S) \subset \overline{\mathcal{M}(S)}$ and an example where the last inclusion is strict can be found in [9]. Nevertheless in this work we will assume that $\mathcal{M}(S)$ is closed which directly implies that $\mathcal{M}(S) = \mathcal{M}^*(S)$. This will be extremely useful in the proof that leads on some properties of the maximal balls.

The third assumption appears necessary in the proof. Notice that it is not so restrictive since the maximality of the balls $B(x, d(x,\partial S))$ and $B(y, d(y,\partial S))$ and the triangular inequality imply that $|d(x,\partial S) - d(y,\partial S)| < ||x-y||$ so that we impose only something a bit more restrictive than a natural property. Moreover it is possible to prove that, when $d = 2$ the third point derives from the first one (closeness of the medial axis being allays satisfied when $d2$). In higher dimension we did not any counter examples of set S satisfying the rolling ball condition and the closeness of the medial axis with the third point not satisfied.

Now recall that we aim to estimate the medial axis of a set S via an estimation based on a finite number of points. It will be seen later that, when dealing with the medial axis, the two parts of the Hausdorff distance between the boundary and its estimator don't have the same importance. This leads us to define (ε, h)-estimations as follows.

Definition 4. *Let S and $\tilde{S}$ be two sets in $\mathbb{R}^d$. Then $\tilde{S}$ is an (ε, h)-estimation of S if:*

$$\max_{y\in\tilde{S}} d(y,S) \leq \varepsilon \text{ and } \max_{x\in S} d(x,\tilde{S}) \leq h.$$

2.2 Main results

The behavior of the estimated medial axis is made explicit in the following theorem. Note that the smoothness conditions on the boundary are close to the one used in [3] where a similar theorem is obtained for a sufficiently dense sample of the boundary that is a $(0,\varepsilon_n)$-estimation of the boundary. Nevertheless, here we can just observe points close to the boundary that is a more realistic assumption in a statistical purpose.

Theorem 1. *Let S be a (K, r_0)-regular compact set. Introduce $\mu_S = diam(S)/r_0$ Suppose that there exists two positive sequences $(\varepsilon_n)_n$ and $(\varepsilon'_n)_n$ such that $\varepsilon_n \to 0$, $\varepsilon'_n \to 0$, $\varepsilon_n \leq \max(1, \frac{r_0}{8}, \frac{5r_0}{8\sqrt{1+\mu_S}})$ and $\varepsilon'_n < r_0/2$. Suppose that we can find $\mathcal{Y} \subset \mathcal{X}_n$ and $\hat{S}_n \subset S$ such that*

1. *$\mathcal{Y} \subset \mathcal{X}_n$ is a $(\varepsilon_n^2, \varepsilon_n)$-estimation of ∂S*
2. *$d_h(\hat{S}_n, S) \leq \varepsilon'_n$ and $d_h(\partial\hat{S}_n, \partial S) \leq \varepsilon'_n$*

There exists an explicit constant λ_0 such that for all $\lambda < \lambda_0$ there exist a constant C, such that, for n large enough:

$$d_h(\hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}), \mathcal{M}) \leq C\varepsilon_n^2$$

In the case where the points are randomly drawn on S satisfying the regularity conditions of Theorem 1 and assuming additional hypothesis on the probability distribution, the following corollary proposes a ways to practically estimate the inner medial axis. For this we use $C_r(\mathcal{X}_n)$, the r-convex hull estimator of S (see [23]), to identify the subset $\mathcal{Y}$ of sample points located close to the boundary. We also use the basic Devroye Wise estimator (see [13]) estimator of the Support. This choices has been done to provide an easy to compute algorithm.

Corollary 1. *Let $\mathcal{X}_n = \{X_1 \dots X_n\}$ be an iid sample of points, drawn on S a (K, r_0)-regular compact set. Assume that the density f of the sample satisfies $f(x) \geq f_0 > 0$ for all $x \in S$. For all $r < r_0$ denote by $\hat{C}_r(\mathcal{X}_n)$ the r-convex hull of $\mathcal{X}_n$ and put $\mathcal{Y}_r = \partial\hat{C}_r(\mathcal{X}_n) \cap \mathcal{X}_n$.*

Let $\hat{S}_n$ be $\hat{S}_n = \bigcup_i \overline{B}(X_i, r_n)$ with $r_n = (4^{1/d} \max_i(\min_{j\neq i} ||X_i - X_j||))$

There exists λ_0 such that, for all $\lambda < \lambda_0$ there exists B_ρ such that

$$d_h(\mathcal{M}, \hat{\mathcal{M}}_\lambda(\hat{C}_r(\mathcal{X}_n), \mathcal{Y}_r)) \leq B_\rho \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1}} \quad e.a.s.$$

2.3 Applications

Support Estimation First let us give a general theorem that gives the convergence rate of the support estimation that directly derives from the medial axis estimation. In a computational purpose we do not propose to estimate the support with

$$S^\circ = \{x, \exists z \in \hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}), d(x, z) \leq d(z, \mathcal{Y})\}$$

but on a Monte-Carlo type estimation of S° based on a finite subset $\mathcal{Z} = \{Z_1, \dots Z_k\} \subset \hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y})$.

Theorem 2. *Under the hypothesis of Theorem 1, for $\lambda < \lambda_0$, if $\mathcal{Z} = \{Z_1, \dots Z_k\}$ is a $(0, \varepsilon_n^2)$ estimation of $\hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}_r)$ then there exits a constant C' such that:*

$$d_h(S, \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C'\varepsilon_n^2 \text{ and } d_h(\partial S, \partial \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C'\varepsilon_n^2.$$

As in the previous section we can derive from this theorem a corollary giving the convergence rate of the support estimator based on a medial axis estimation lying on the r-convex hull to estimate the boundary points. Recall that the interest of such a result consist in the computational easiness of deciding whether a new point belongs to the support or not. This is closely related to k the number of points on the estimated medial axis that should be as small as possible. To control that number introduce a packing definition.

Definition 5. *Let A be a subset of $\mathbb{R}^d$, a finite subset of A $\mathcal{Z} = \{Z_1, \dots Z_k\} \subset A$ is said to pack A with a radius r if*

$$\forall i \neq j : \ \overline{B}(Z_i, r) \cap \overline{B}(Z_j, r) = \emptyset \text{ and } A \subset \bigcup_i \overline{B}(Z_i, 2r)$$

A consequence of well known properties of smooth enough compact d'-dimensional manifold is the following property.

Proposition 2. *If $A \subset \mathbb{R}^d$ is a compact d'-dimensional $\mathcal{C}^1$manifold then it admits packing subsets $\mathcal{Z} = \{Z_1, \dots Z_k\}$ with a radius r and there exists D_A a constants such that $k \leq D_A r^{-d'}$*

The following results says that it is sufficient to obtain $O(n^{\frac{2d'}{d+1}})$ on $\hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}_r)$ to have an estimation of S as good as the r-convex hull. That is specially useful when $d' \ll d$.

Corollary 2. *Under the hypothesis of Theorem 1, Suppose that $\mathcal{M}(S)$ is a $\mathcal{C}^1$ d'-dimensional manifold. For a given $\lambda < \lambda_0$ the constant of Theorem 1. Let $\mathcal{Z} = \{Z_1, \dots Z_k\}$ be a packing subset of $\hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}_r)$ for a radius $\rho_n = \rho_0 \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1}}$ then there exits a constant C' such that, eventually almost surely:*

$$d_h(S, \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C' \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1}} \text{ and } d_h(\partial S, \partial \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C' \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1}}.$$

Remark that this corollary is a direct consequence of Theorem 2 so that we will omit the proof.

Manifold estimation Another application of the estimation of the medial axis is that, under some hypothesis it can be used for manifold estimation. Let us recall that manifold estimation deals with the following problem. Suppose that Y is a random variable drawn with a probability $\mathbb{P}_Y$ supported by M a compact d'-dimensional manifold. Suppose, now that we can not observe Y but $X = Y + U$. The manifold estimation problem consists in estimating M.

The following theorem stands that, if M is smooth enough and, if the noise is such that S the support of X is $\bigcup_{x\in M} \overline{B}(x, \rho_1)$ then the manifold estimation problem and the medial axis estimation problem are the same. Namely, smoothness being characterized by the reach see [14] or [27] defined as follows:

$$\text{reach}(M) = \sup\{r \in \mathbb{R}, d(z, M) < r \Rightarrow \exists! z^* \in M, ||z - z^*|| = d(z, M)\} \qquad (4)$$

we have

Proposition 3. *Let $M \subset \mathbb{R}^d$ be a closed manifold with a positive reach ρ_1, for all $\rho_0 < \rho_1$ $\mathcal{M}(\bigcup_{x\in M} \overline{B}(x, \rho_0)) = M$*

The Following results, stands that, under the same hypothesis than in [17] the medial axis estimator based on the $r-$convex hull is a manifold estimator and that it is minimax when $d' = d - 1$.

Corollary 3. *Let M be a compact d'-dimensional manifold without boundary with positive reach ρ_1. Suppose that $\mathcal{Y} = \{Y_1, \dots Y_n\}$ is a iid sample drawn with a probability $\mathbb{P}_Y$ supported by M with a density f_Y such that, for all $y \in M, f_Y(y) \geq f_0 > 0$. Suppose that we can not observe the random variable Y but that we observe $X = Y + U$ such that the distribution of $U|\{Y = y\}$ is supported by $S_y = \{y + \rho u, u \in N_y M, ||u|| = 1, 0 \leq \rho \leq \rho_0\}$ and has a constant density $f_{U|Y=y}$ on S_y. Finally suppose that $\rho_0 < \rho_1$ we have:*

There exists λ_0 such that, for all $\lambda < \lambda_0$ there exists B such that

$$d_h(M, \hat{\mathcal{M}}_\lambda(\hat{C}_r(\mathcal{X}_n), \mathcal{Y}_r)) \leq B_\rho \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1}} \quad e.a.s.$$

Under a different noise hypothesis we conjecture the following result (in section proof we will present the arguments that make us strongly believe that it is true).

Conjecture. *Let M be a compact d'-dimensional manifold. Suppose that reach$(M) = \rho_1 > 0$. Suppose that $\mathcal{Y} = \{Y_1, \dots Y_n\}$ is a iid sample drawn with a probability $\mathbb{P}_Y$ supported by M with a density f_Y such that, for all $y \in M, f_Y(y) \geq f_0 > 0$. Suppose*

that we can not observe the random variable Y but that we observe $X = Y + U$ such that the distribution of $U|\{Y = y\}$ is supported by $\overline{B}(y, \rho_0)$ and has density $f_{U|y}$. Suppose that there exists $c > 0$ such that for all $y \in M$, for all $x \in S_y$ we have $f_{U|y}(x) \geq c$. Finally suppose that $\rho_0 < \rho_1$ we have:

There exists λ_0 such that, for all $\lambda < \lambda_0$ there exists B such that

$$d_h(M, \hat{\mathcal{M}}_\lambda(\hat{C}_r(\mathcal{X}_n), \mathcal{Y}_r)) \leq B_\rho \left(\frac{\ln n}{n}\right)^{\frac{2}{d+1+2d'}} \quad e.a.s.$$

2.4 Proofs

Before setting the proof let us first introduce some definitions and notations. Let A be an open set and C a closed set such that $A \cap C = \emptyset$ and introduce $\mathcal{B}(A, C) = \{B(x, r), x \in A, B(x, r) \cap C = \emptyset\}$, $\mathcal{C}^{\max}(A, C)$ the set of the maximal balls (for the inclusion) of $\mathcal{B}(A, C)$ and $\overline{\mathcal{M}}(A, C)$ the set of the centers of the balls of $\mathcal{B}(A, C)$. In the following, $\mathcal{S}(x, r)$ denotes the sphere of radius r centered at x, let A and C be two sets $A \Delta C = (A \setminus C) \cup (C \setminus A)$ is there symmetric difference.

Proof of theorem 1

Proof. First note that for all $x \in \mathcal{M}$ we have $r(x) \leq \text{diam}(S) \leq \mu_S r_0$. Second remark that the inner rolling ball condition implies that for all $x \in \mathcal{M}$, $r(x) = d(x, \partial S) \geq r_0$.

Notice that, for any $x \in \mathcal{M}(S)$ we have that $x \in \hat{S}_n$. Indeed suppose the reverse. As there exists $x' \in \hat{S}_n$ with $||x' - x|| \leq \varepsilon'_n$, there exists $x'' \in [x, x'] \cap \partial \hat{S}_n$, $||x - x''|| \leq \varepsilon'_n$. As $B(x, r(x)) \subset S$ we also have $B(x'', r(x) - \varepsilon'_n) \subset S$ and $d(x'', \partial S) \geq r(x) - \varepsilon'_n \geq r_0 - \varepsilon'_n$. Thus $0 \geq r_0 - \varepsilon'_n$ that is impossible because $\varepsilon'_n < r_0/2$. We also have that $B(x, r(x) - 2\varepsilon_n^2) \cap \mathcal{Y} = \emptyset$. Thus $B(x, r(x) - 2\varepsilon_n^2) \in \mathcal{B}(\hat{S}_n, \mathcal{Y})$ and there exist $x' \in \hat{S}_n$ and r' such that:

$$B(x, r(x) - 2\varepsilon_n^2) \subset B(x', r') \text{ with } B(x', r') \in \mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y}). \tag{5}$$

We are now going to prove that for all (x', r') such that $B(x, r(x) - 2\varepsilon_n^2) \subset B(x', r')$ and $B(x', r') \in \mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$ we have :

$$||x - x'|| \leq \frac{1 + K}{1 - K} \left(2 + \frac{8(1 + \mu_S)}{5r_0}\right) \varepsilon_n^2. \tag{6}$$

Introduce x'^* a point of ∂S such that $d(x', \partial S) = ||x' - x'^*||$ and $\gamma = r' - ||x' - x'^*||$. Notice that

1. Because $x'^* \in \partial S$, there exits $y_i \in \mathcal{Y}$ such that $||y_i - x'^*|| \leq \varepsilon_n$. Thus $B(x', ||x' - x'^*|| + \varepsilon_n) \notin \mathcal{B}(\hat{S}_n, \mathcal{Y})$ and we obtain : $r' \leq ||x' - x'^*|| + \varepsilon_n$ thus $\gamma < \varepsilon_n$.

2. Conversely, because $B(x', r')$ is in $\mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$ there exists $y \in \mathcal{Y}$ such that $||x' - y|| = r'$ thus there exists $z \in \partial S$ such that $||y - z|| \leq \varepsilon_n^2$ so we have $||x' - x'^*|| \leq r' + \varepsilon_n^2$ so that $\gamma \geq -\varepsilon_n^2$.

Finally, we have that

$$-\varepsilon_n^2 \leq \gamma \leq \varepsilon_n \tag{7}$$

Remark now that there exists $y_i \in \mathcal{Y}$ such that:

1. $||y_i - x'^*|| \leq \varepsilon_n$, because $\mathcal{Y}$ is a $(\varepsilon, \varepsilon_n^2)$-estimation of ∂S;

2. $||y_i - O^{\text{out}}_{x'^*}|| \geq r_0$, indeed $y_i \notin \mathcal{B}(O^{\text{out}}_{x'^*}, r_0)$, because $y_i \in S$;

3. $||y_i - x'|| \geq r'$, because $B(x', r')$ is a ball of $\mathcal{B}(\hat{S}_n, \mathcal{Y})$.

Notice that x', x'^* and $O^{\text{out}}_{x'^*}$ are on a same line directed by $u = \frac{x'^* - x'}{||x'^* - x'||}$. We so have $x' = x'^* - (r' - \gamma)u$ and $O^{\text{out}}_{x'^*} = x'^* + r_0 u$. Let us write $y_i - x'^* = au + bw$ where and w is a unit vector of $u^\perp$. Notice that w can be chosen such that $b \geq 0$. See Figure 4 for the position of the different points.

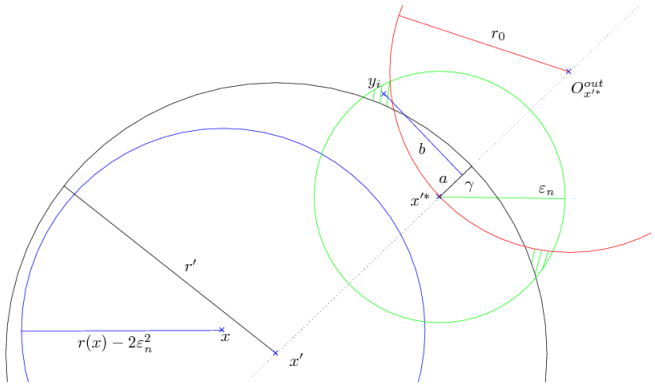

Figure 4: Let $B(x', r') \in \mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$ and x'^* be a point of ∂S such that such that $d(x', \partial S) = ||x' - x'^*||$. Then there exists a $y_i \in \mathcal{Y}$ in the green dashed area.

The previous considerations on the location of y_i gives:

$$\begin{cases} a^2 + b^2 \leq \varepsilon_n^2 \\ b^2 + (r_0 - a)^2 \geq r_0^2 \\ (r' - \gamma + a)^2 + b^2 \geq r'^2 \end{cases}$$

That, in turns gives:

$$\begin{cases} a^2+b^2 \leq \varepsilon_n^2 \\ 2ar_0 \leq \varepsilon_n^2 \\ \gamma^2 - 2(r'+a)\gamma + (1+\frac{r'}{r_0})\varepsilon_n^2 \geq 0 \end{cases}$$

Now we focus mainly on the last inequality. Notice first that $r'+a \geq r' - \varepsilon_n \geq r_0 - 2\varepsilon_n^2 - \varepsilon_n \geq 5r_0/8$. Notice second that $r' \leq \text{diam}(S)$ so that, since $\varepsilon_n^2(1+\mu_S) \leq (5r_0/8)^2$ that is guaranteed by the hypotheses, the discriminant is positive and we must have $\gamma \leq r'+a-\sqrt{(r'+a)^2-(1+\frac{r'}{r_0})\varepsilon_n^2}$ or $\gamma \geq r'+a+\sqrt{(r'+a)^2-(1+\frac{r'}{r_0})\varepsilon_n^2}$.

We are now going to prove that the second case is impossible. Indeed, if $\gamma \geq r'+a+\sqrt{(r'+a)^2-(1+\frac{r'}{r_0})\varepsilon_n^2}$ we also have $\gamma \geq r'+a \geq r'-\varepsilon_n \geq r_0 - 2\varepsilon_n^2 - \varepsilon_n$ (because $|a| \leq \varepsilon_n$ by the first inequality and because $r' \geq r(x) - 2\varepsilon_n^2 \geq r_0 - 2\varepsilon_n^2$ by initial inclusion. Thus, by (7) we have $r_0 - 2\varepsilon_n^2 \leq 2\varepsilon_n$ so $r_0 \leq 2\varepsilon_n + 2\varepsilon_n^2 \leq 4\varepsilon_n \leq r_0/2$ (because $\varepsilon_n \leq 1$ then because $\varepsilon_n \leq r_0/8$) that is impossible.

Now we have $\gamma \leq r'+a-\sqrt{(r'+a)^2-(1+\frac{r'}{r_0})\varepsilon_n^2}$. Let us recall that $1-\sqrt{1-x} \leq x$ when $x \in [0,1]$ so that $\gamma \leq \frac{1+\frac{r'}{r_0}}{r'+a}\varepsilon_n^2 \leq \frac{8(1+\mu_S)}{5r_0}\varepsilon_n^2$. Introduce $c_S = \frac{8(1+\mu_S)}{5r_0}$ we have $B(x', r' - c_S\varepsilon_n^2) \subset \mathcal{B}(\mathring{S}, \partial S)$, so there exists $B(x'', r(x'')) \in \mathcal{C}^{\max}(\mathring{S}, \partial S)$ with $B(x', r' - c_S\varepsilon_n^2) \subset B(x'', r(x''))$, that is:

$$\exists x'' \in \mathcal{M} \text{ such that } B(x', r' - c_S\varepsilon_n^2) \subset B(x'', r(x'')). \tag{8}$$

Now, by (5) and (8) it follows that $B(x, r(x) - (2+c_S)\varepsilon_n^2) \subset B(x'', r(x''))$. As a consequence, by the triangular inequality, we obtain $r(x'') \geq ||x''-x|| + r(x) - (2+c_S)\varepsilon_n^2$. Thus, using now the K-regularity: $(2+c_S)\varepsilon_n^2 \geq ||x''-x|| + r(x) - r(x'') \geq (1-K)||x-x''||$ and we finally obtain:

$$||x-x''|| \leq \frac{(2+c_S)\varepsilon_n^2}{1-K}. \tag{9}$$

By (5) we also have $r' \geq ||x-x'|| + r(x) - 2\varepsilon_n^2$ and by (8) we have $r(x'') \geq ||x''-x'|| + r' - c_S\varepsilon_n^2$. Summing these two inequalities gives: $r(x'') \geq r(x) + ||x-x'|| + ||x'-x''|| - (2+c_S)\varepsilon_n^2$ so

$$||x-x'|| \leq r(x'') - r(x) + (2+c_S)\varepsilon_n^2$$

So again using the K-regularity of the support and (9) we obtain $||x-x'|| \leq \frac{1+K}{1-K}(2+c_S)\varepsilon_n^2$. This concludes the proof of (6).

Assuming that, if $B(x', r')$ is a ball of $\mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$, then the maximality impose that there exists at least two points in $\overline{B(x', r')} \cap \mathcal{Y}$ and so $x' \in \mathcal{M}_{0,\hat{S}_n}(\mathcal{Y})$. That is not sufficient since we want to guarantee the existence of a x' in some $\mathcal{M}_{C,\hat{S}_n}(\mathcal{Y})$, with C large enough and close to x. Namely, introduce $\lambda_0 = \min\{\sqrt{\frac{1-K}{(1+K)(1+2c_S)}} r_0, \frac{r_0}{2}\}$. We now aim to prove that, for all $x \in \mathcal{M}(S)$ there exits x' and r'_x such that

$$x' \in \mathcal{M}_{\lambda_0,\hat{S}_n}(\mathcal{Y}) \ , \ B(x', r'_x) \in \mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y}) \ , \text{ and } B(x, r(x) - 2\varepsilon_n^2) \subset B(x', r'_x). \quad (10)$$

More precisely we will show that this is realized for x', such that $B(x', r'_x)$ is a ball of $\mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$ that has a maximum radius, that is such that:

$$r'_x = \max\{r', B(x, r(x) - 2\varepsilon_n^2) \subset B(x', r'), B(x', r') \in \mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})\}.$$

Clearly there exists $y \in \mathcal{S}(x', r'_x) \cap \mathcal{Y}$. Suppose that $\max\{||y - z||, z \in S(x', r'_x) \cap \mathcal{Y}\} = l \leq r_0/2$. Introduce a point $z_0 \in \mathcal{S}(x', r'_x) \cap \mathcal{S}(y, l)$, $u = \frac{x'-y}{||x'-y||}$, $x''_t = x' + tu$ and $r''_t = ||z_0 - x''_t||$ (See Figure 5).

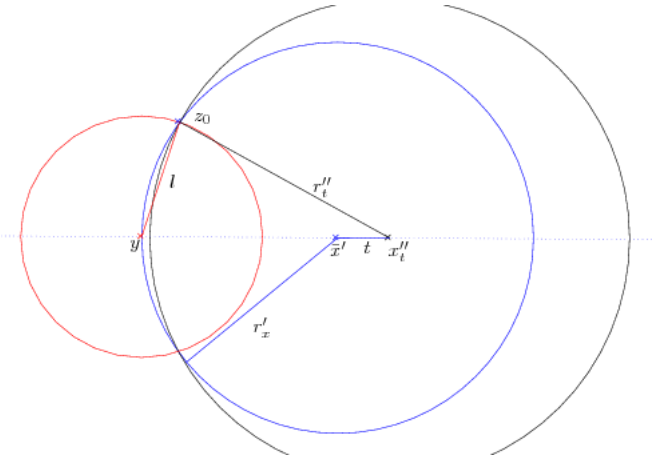

Figure 5: Construction of x''_t.

Note first that $\langle x' - z_0, u\rangle = \frac{2(r'_x)^2 - l^2}{2r'_x} \geq \frac{3}{4} > 0$ (because $l^2 \leq r_0^2/4 \leq (r'_x)^2$). Thus, when $t > 0$ we have:

$$(r''_t)^2 = ||x' - z_0 + tu||^2 = (r'_x)^2 + t^2 + 2t\langle x' - z_0, u\rangle = (r'_x)^2 + t^2 + t\frac{2(r'_x)^2 - l^2}{r'_x} > (r'_x)^2 \quad (11)$$

and

$$B(x''_t, r''_t) \cap B^c(x', r'_x) \cap B(y, l) = \emptyset \quad (12)$$

Indeed, consider $x \in B(x''_t, r''_t) \cap B^c(x', r'_x) \cap B(y, l)$, $x = x' + au + bw$ with w a unit vector of $u^\perp$, we have:

$$\begin{cases} a^2 + b^2 \geq (r'_x)^2 \\ (a-t)^2 + b^2 < (r'_x)^2 + t^2 + t\frac{2(r'_x)^2 - l^2}{r'_x} \\ (a + r'_x)^2 + b^2 < l^2 \end{cases}$$

combining first and second inequalities, then first and third inequalities gives:

$$\begin{cases} a > \frac{l^2}{2r} - r \\ a < \frac{l^2}{2r} - r \end{cases}$$

That is impossible.

Consider now $\rho(t) = d(x''_t, \mathcal{Y} \cap B^c(y, l)) - r''_t$. For $t = 0$ we have $\rho(0) > 0$ so that, using continuity arguments, there exists $t_0 > 0$ such that, for all $t \in [0, t_0]$ we have $\rho(t) > 0$ and thus, by (12) $B(x''_t, r''_t) \cap \mathcal{Y} = \emptyset$. Observe that, because x' is far enough from the boundary of S, we can can choose t_0 such that for all $t < t_0$ we have $x''_t \in \hat{S}_n$ and $B(x''_t, r''_t) \cap \mathcal{Y} = \emptyset$.

As $B(x', r'_x)$ is a ball containing $B(x, r(x) - 2\varepsilon_n^2)$, in $\mathcal{C}^{\max}(\hat{S}_n, \mathcal{Y})$ that has a maximal radius we must have that, for all $0 < t < t_0$ there exists $y_t \in B^c(x''_t, r''_t) \cap B(x, r(x) - 2\varepsilon_n^2)$. Observe that we also have $y_t \in (B^c(x''_t, r''_t) \cap \overline{B(x', r')}) \cap \overline{B(x, r(x) - 2\varepsilon_n^2)}$.

So, taking $t \to 0$ and using compactness arguments we obtain that there exists $y' \in \mathcal{S}(x', r') \cap \overline{B(y, l)} \cap \overline{B(x, r(x) - 2\varepsilon_n^2)}$. Notice also that, if $y' \in \mathcal{S}(x', r') \cap \overline{B(x, r(x) - 2\varepsilon_n^2)}$, the inclusion $B(x, r(x) - 2\varepsilon_n^2) \subset B(x', r')$ implies that x', x and y' are on the same line and that $||x - y'|| = r(x) - 2\varepsilon_n^2$. Introduce now $\gamma = ||x - y|| - (r(x) - 2\varepsilon_n^2)$ and $\theta = \angle yx'y'$. See Figure 6 for the general configuration of x, x', y, y', γ and θ .

Recall that $y \in \mathcal{Y}$ so that there exists $z \in \partial S$ such that $||y - z|| \leq \varepsilon_n^2$, recall also that $B(x, r(x)) \subset S$ so that $||z - x|| \geq r(x)$ thus $||x - y|| \geq r(x) - \varepsilon_n^2$ and $\gamma \geq \varepsilon_n^2$.

By $||y - y'||^2 = ||y - x||^2 + ||x - y'||^2 + 2\langle y - x, x - y' \rangle$ it follows that $||y - y'||^2 = (r(x) - 2\varepsilon_n^2)^2 + (r(x) - 2\varepsilon_n^2 + \gamma)^2 - 2(r(x) - 2\varepsilon_n^2)(r(x) - 2\varepsilon_n^2 + \gamma)\cos(\theta)$. Thus we have:

$$\cos(\theta) = \frac{(r(x) - 2\varepsilon_n^2)^2 + (r(x) - 2\varepsilon_n^2 + \gamma)^2 - ||y - y'||^2}{2(r(x) - 2\varepsilon_n^2)(r(x) - 2\varepsilon_n^2 + \gamma)}. \tag{13}$$

We also have $x' - x = ||x' - x||(\cos(\theta), -\sin(\theta))$. Thus

$$||x' - y||^2 = (r(x) - 2\varepsilon_n^2 + \gamma + ||x' - x||\cos(\theta))^2 + (||x' - x||\sin(\theta))^2. \tag{14}$$

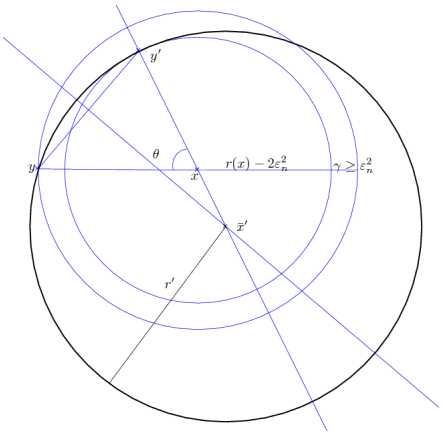

Figure 6:

But we also have $||x' - y||^2 = ||x' - y'||^2 = (r')^2 = (r(x) - 2\varepsilon_n^2 + ||x' - x||)^2$. Thus, from (14) we obtain:

$$(r(x)-2\varepsilon_n^2+\gamma)^2+2||x'-x||(r(x)-2\varepsilon_n^2+\gamma)\cos(\theta) = (r(x)-2\varepsilon_n^2)^2+2||x'-x||(r(x)-2\varepsilon_n^2),$$

and thus :

$$||x' - x|| = \frac{1}{2}\frac{2\gamma(r(x) - 2\varepsilon_n^2) + \gamma^2}{(r(x) - 2\varepsilon_n^2) - (r(x) - 2\varepsilon_n^2 + \gamma)\cos(\theta)}.$$

That, combined with (13) gives

$$||x' - x|| = \frac{2\gamma(r(x) - 2\varepsilon_n^2)^2 + \gamma^2(r(x) - 2\varepsilon_n^2)}{||y - y'||^2 - 2\gamma(r(x) - 2\varepsilon_n^2) - \gamma^2}. \tag{15}$$

Finally, recall that $r(x) - 2\varepsilon_n^2 \geq r(x) - 2\varepsilon_n \geq 3r_0/4$ and $\gamma \geq \varepsilon_n^2$ so that (15) gives: $||x - x'|| \geq \frac{r_0^2\varepsilon_n^2}{||y-y'||^2}$. Observe that because of (6) we have $||x' - x|| \leq \frac{K+1}{1-K}(2 + c_S)\varepsilon_n^2$. Therefore we finally obtain:

$$||y - y'|| \geq \sqrt{\frac{1 - K}{(1 + K)(1 + 2c_S)}}r_0$$

As, we have $l \geq ||y - y'||$ we have finally proved that, if $l \leq r_0/2$ we have $l \geq \sqrt{\frac{1-K}{(1+K)(1+2c_S)}}r_0$. This concludes the proof of (10).

Now from (6) and (10) we have that, for all $\lambda < \lambda_0$:

$$\text{for all } x \in \mathcal{M}(S) \text{ there exists } x' \in \hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}) \text{ such that } ||x - x'|| \leq \frac{K+1}{1-K}(2 + c_S)\varepsilon_n^2. \tag{16}$$

Consider points $x \in S$ such that $d(x, \mathcal{M}) \geq e$. In this last part of the proof we put $l = d(x, \partial S)$. Consider a point $x'^* \in \partial S$ such that $||x - x'^*|| = l$. As $B(x, l) \subset \mathring{S}$ one can introduce $B(x', r(x'))$, a ball of $\mathcal{C}^{\max}(\mathring{S}, \partial S)$ containing $B(x, l)$. Recall that the regularity condition on S allows the existence of $O^{\text{out}}_{x'^*}$ such that $B(O^{\text{out}}_{x'^*}, r_0) \subset S^c$ and observe that x, x', x'^* and $O^{\text{out}}_{x'^*}$ are on the same line, and that $r(x') = ||x' - x|| + l$ with $e' = ||x' - x|| \geq e$.

Because $\mathcal{Y}$ is a $(\varepsilon_n^2, \varepsilon_n)$-estimation of ∂S there exists $y \in \mathcal{Y}$ such that $||x'^* - y|| \leq \varepsilon_n$. As $d(y, \partial S) \leq \varepsilon_n^2$, we also have $||y - x'|| \geq r(x') - \varepsilon_n^2$, and since $y \in S$ we have $||y - O^{\text{out}}_{x'^*}|| \geq r_0$, that is $y \in B(x'^*, \varepsilon_n) \cap B^c(x', r(x') - \varepsilon_n^2) \cap B^c(O^{\text{out}}_{x'^*}, r_0)$. See Figure 7 for the construction (y being in the blue zone).

Let us write $y = x'^* + au + bw$, where $u = \frac{x - x'}{||x - x'||}$ and $w \in u^\perp$, since $||x'^* - y|| \leq \varepsilon_n$ and $||y - O^{\text{out}}_{x^*}|| \geq r_0$ we have:

$$\begin{cases} a^2 + b^2 \leq \varepsilon_n^2 \\ a^2 - 2r_0 a + b^2 \geq 0 \end{cases}$$

Thus

$$a^2 \left(1 - \frac{r(x') - e'}{r_0 + r(x') - e'}\right) + 2\frac{r_0(r(x') - e')}{r_0 + r(x') - e'} a + b^2 \left(1 - \frac{r(x') - e'}{r_0 + r(x') - e'}\right) \leq \varepsilon_n^2.$$

That is

$$\frac{r_0}{r_0 + r(x') - e'} \left((a + r(x') - e')^2 + b^2 - (r(x') - e')^2\right) \leq \varepsilon_n^2.$$

And Finally

$$||x - y||^2 \leq (r(x') - e')^2 + \frac{r_0 + r(x') - e'}{r_0} \varepsilon_n^2 \leq (1 + \mu_S)\varepsilon_n^2.$$

Thus, for all $y_i \in \mathcal{Y}$ such that $x \in \text{Vor}(y_i)$, we have : $y_i \in B(x, \sqrt{(r(x') - e')^2 + (1 + \mu_S)\varepsilon_n^2})$. Because, $y_i \in \mathcal{Y}$ we also have $||y_i - x'|| \geq r(x') - \varepsilon_n^2$. One can introduce $E_x = B(x, \sqrt{(r(x') - e')^2 + C_1\varepsilon_n^2}) \cap B^c(x', r(x') - \varepsilon_n^2)$. See Figure 7 again (the possible location for the y_i being in he brown zone).

For all $y_i = x'^* + au + bw$, where $u = \frac{x - x'}{||x - x'||}$ and $w \in u^\perp$, $y \in E_x$ we have

$$\begin{cases} (a - r(x') + e')^2 + b^2 \leq (r(x') - e')^2 + (1 + \mu_S)\varepsilon_n^2 \\ (a - r(x'))^2 + b^2 \geq (r(x') - \varepsilon_n^2)^2. \end{cases}$$

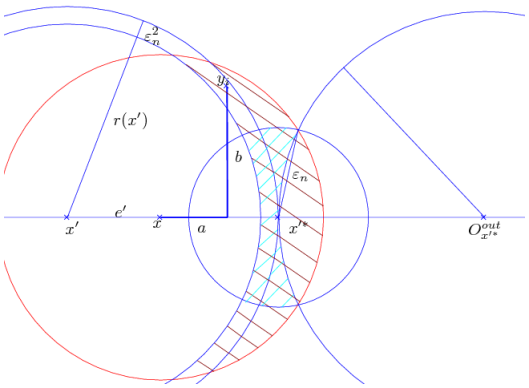

Figure 7: As there exists an observation in the blue crosshatched domain, all the y_i such that $x \in \mathrm{Vor}_{\mathcal{Y}}(y_i)$ are in E_x, the brown crosshatched domain.

By subtraction we have: $2ae' \leq (1+\mu_S+2r(x'))\varepsilon_n^2$, thus considering the first inequality it comes that: $a^2+b^2 \leq (1+\mu_S)\varepsilon_n^2 + 2a(r(x')-e') \leq (1+\mu_S)\varepsilon_n^2 + (r(x')-e')(1+\mu_S+2r(x'))\frac{\varepsilon_n^2}{e'}$ Thus:

$$\mathrm{diam}(E_x) \leq 2\sqrt{(1+\mu_S)\varepsilon_n^2 + \mu_S r_0(1+\mu_S+2\mu_S r_0)\frac{\varepsilon_n^2}{e'}}$$

And, for all $\mu < \frac{\lambda}{2}$ we have that obtained that, for all $x \in S$, if $d(x,\mathcal{M}) \geq \frac{\mu_S r_0(1+\mu_S+2\mu_S r_0)\varepsilon_n^2}{\mu^2}$ then $x \notin \hat{\mathcal{M}}_{2\sqrt{\mu^2+(1+\mu_S)\varepsilon_n^2},\hat{S}_n}(\mathcal{Y}))$, and for n large enough, $x \notin \hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y})$. Choosing $\mu = \lambda/4$ it comes that, for n large enough for all $x \in S \cap \hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y})$, $d(x,\mathcal{M}) \leq 4\frac{\mu_S r_0(1+\mu_S+2\mu_S r_0)\varepsilon_n^2}{\lambda^2}$.

To conclude the proof it only remains to prove that, for n large enough, $S^c \cap \hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y}) = \emptyset$ that is easy since, if $x \in S^c\hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y})$ then $x \in S^c \cap \hat{S}_n$ so $d(x,\partial S) \leq \varepsilon_n'$ and there exists $y_i \in \mathcal{Y}$ such that $||x-y_i|| \leq \varepsilon_n + \varepsilon_n'$, so that, for all $y \in \mathcal{Y}$ such that $x \in \mathrm{Vor}(y_i)$ we have $||x-y_i|| \leq \varepsilon_n + \varepsilon_n'$ and thus $\mathrm{diam}\{y, x \in \mathrm{Vor}(y_i)\} \leq 2(\varepsilon_n+\varepsilon_n')$ that is impossible, for n large enough because we should have $\mathrm{diam}\{y, x \in \mathrm{Vor}(y_i)\} \geq \lambda$ as $x \in \hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y})$.

We so obtain that, for n large enough, for all $x \in \hat{\mathcal{M}}_\lambda(\hat{S}_n,\mathcal{Y})$, $d(x,\mathcal{M}) \leq 4\frac{\mu_S r_0(1+\mu_S+2\mu_S r_0)\varepsilon_n^2}{\lambda^2}$. that, combined to (10) concludes the proof. □

Proof of Corollary 1

Proof. We refer the reader to [23] to be convinced that there exist constants a_r, b_r and c_r such that, eventually almost surely $\mathcal{Y}_r$ is (eventually almost surely) a $(a_r(\ln n/n)^{\frac{2}{d+1}}, b_r(\ln n/n)^{\frac{1}{d+1}})$-

estimation of S. Considering the Devroye Wise estimator we refer the reader to [12]. Roughly our shape hypothesis implies standardness with $\delta = f_0/2$, it also implies partly expandable hypothesis so that by application of theorems 2 and 3 the choice of $r_n = (4 \ln n/(n f_0 \omega_d f_0))^{1/d}$ allow to have convergent estimator for the support and the boundary. By application of result of Jansen in [22] the proposed data driven r_n sequence ensures that, eventually almost surely, $d_h(\hat{S}_n, S)$ and $d_h(\partial\hat{S}_n, \partial S)$ converge to 0. □

Proof of Theorem 2

Proof. We are going to prove that, for all $(Z_i, x) \in \hat{\mathcal{M}}_\lambda(\hat{S}_n, \mathcal{Y}_r) \times \mathcal{M}(S)$ such that$||Z_i - x|| \leq A\varepsilon_n^2$ we have $r_i = d(Z_i, \mathcal{Y})$ that satisfies:

$$r(x) - (A+1)\varepsilon_n^2 \leq r_i \leq r(x) + (A + r_0^{-1})\varepsilon_n^2. \tag{17}$$

Consider first the part $r(x) - (A+1)\varepsilon_n^2 \leq r_i$. Recall that $B(x, r(x)) \subset S$ so $\mathcal{Y} \cap B(x, r(x) - \varepsilon_n^2) = \emptyset$ (otherwise there exists Y_i with $d(Y_i, \partial S) > \varepsilon_n^2$). Thus $d(x, \mathcal{Y}) \geq r(x) - \varepsilon_n^2$ and, by triangular inequality $r_i \geq r(x) - (A+1)\varepsilon_n^2$.

Second Consider the part $r_i \leq r(x) + (A + r_0^{-1})\varepsilon_n^2$.

i. there exists $x^* \in \partial S$ such that $d(x, \partial S) = r(x) = ||x - x^*||$ and there exists $y \in \mathcal{Y}$ with $||y - x^*|| \leq \varepsilon_n$ (because $\mathcal{Y}$ is a $(\varepsilon_n^2, \varepsilon_n)$-estimation of ∂S);

ii. by the rolling ball condition, $y \notin B(O_{x^*}^{\text{out}}, r_0)$ with $O_{x^*}^{\text{out}} = r_0 \frac{x^* - x}{||x^* - x||}$,

iii. if $r_i > r(x') + (A + r_0^{-1})\varepsilon_n^2$ then $y \notin B(x, r(x) + r_0^{-1}\varepsilon_n^2)$

Introduce $u = \frac{x^* - x}{||x^* - x||}$, if $z = x^* + au + bw$ with $w \in u^\perp$ then the previous conditions give;

$$\begin{cases} a^2 + b^2 \leq \varepsilon_n^2 \\ (a + r(x))^2 + b^2 > (r(x) + r_0^{-1}\varepsilon_n^2)^2 \\ (a - r_0)^2 + b^2 > r_0^2 \end{cases}$$

Combining inequalities 1 and 2 and inequalities 1 and 3 we obtain:

$$\begin{cases} a > \left(\frac{1}{r_0} - \frac{1}{2r(x)}\right)\varepsilon_n^2 + \frac{\varepsilon_n^2}{r_0^2} \geq \left(\frac{1}{r_0} - \frac{1}{2r_0}\right)\varepsilon_n^2 \geq \frac{\varepsilon_n^2}{2r_0} \\ a < \frac{\varepsilon_n^2}{2r_0} \end{cases}$$

That is impossible that concludes the proof of Equation (17).

We are now going to prove that:

$$\bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))\Delta S \subset \partial S \oplus C''\varepsilon_n^2 \overline{B} \text{ with } C'' = \max(2C + r_0^{-1}, 2C + 3) \qquad (18)$$

Suppose that $x \in \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))\backslash S$ then there exists i such that $x \in \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ and, because $Z_i \in \hat{\mathcal{M}}_\lambda$ and $d_h(\hat{\mathcal{M}}_\lambda, \mathcal{M}) \leq C\varepsilon_n^2$ there exists $z \in \mathcal{M}$ with $||z - Z_i|| \leq C\varepsilon_n^2$ so that by (17) (with $A = C$) and triangular inequality $x \in \overline{B}(z, r(z) + (2C + r_0^{-1})\varepsilon_n^2)$ and $x \notin S$. Consider $x' = \mathcal{S}(z, r(z)) \cap [z, x]$ we have $[x', x] \cap \partial S \neq \emptyset$ (indeed $B(z, r(z)) \subset S$ so or $x' \in \partial S$ or $x' \in \mathring{S}$ and $]x', x] \cap \partial S \neq \emptyset$). From which we can deduce that there exits $x^* \in \partial S$ such that $||x - x^*|| \leq (2C + r_0^{-1})\varepsilon_n^2)$.

Suppose now that $x \in S \setminus \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ for all $z \in \mathcal{M}(S)$ such that $x \in \overline{B}(z, r(z))$ and there exits $Z_i \in \mathcal{Z}$ such that $||Z_i - z|| \leq (C+1)\varepsilon_n^2$, as $x \notin \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ we also have $||x - Z_i|| > r_i$ so that, by (17) (with $A = C+1$) $||x - Z_i|| > r(z) - (C+2)\varepsilon_n^2$ and, finally triangular inequality implies that $||x - z|| > r(z) - (2C+3)\varepsilon_n^2$. If $x \notin \partial S$, let us now introduce $l = d(x, \partial S)$ and $x^* \in \partial S$ such that $||x - x^*|| = l$ we have $B(x, l) \subset \mathring{S}$ and one can introduce $B(z_0, r(z_0))$ be a ball of $\mathcal{C}^{\max}(\mathring{S}, \partial S)$ containing $B(x, l)$ (obviously $x \in B(z_0, r(z_0))$. Notice that the regularity condition implies that there exists a unique z_0 and that $x \in [z_0, x^*]$ so that $||x - z_0|| = r(z_0) - l$. Finally, this and previous consideration gives that, if $x \notin \partial S$ then $d(x, \partial S) \leq (2C+3)\varepsilon_n^2$ that concludes the proof of (18).

Finally, (18) directly implies that: For all $x \in \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ and $x \notin S$ there exist $x' \in \partial S$ such that $||x - x'|| \leq C''\varepsilon_n^2$.

Now, suppose that n is large enough to have $C''\varepsilon_n^2 < r_0$, for all x such that $d(x, \partial S) \leq C''\varepsilon_n^2 < r_0$ denote by x^* its unique (due to the regularity condition) projection onto ∂S and introduce u_{x^*} the unit vector, tangent to ∂S pointing outward S. If $x \in S$ and $x \notin \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ then by (18) $x \in [x^*, x^* - C''\varepsilon_n^2 u_{x^*}]$ and for all $t \in]C''\varepsilon_n^2, r_0]$ we have $x_t = x^* - t u_{x^*} \in \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))$ so that (doing $t \to C''\varepsilon_n^2$) exists $x' \in \partial\left(\bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y}))\right)$ with $||x' - x|| \leq C''\varepsilon_n^2$.

So that it comes that

$$d_h(S, \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C'\varepsilon_n^2 \text{ and } d_h(\partial S, \partial \bigcup_i \overline{B}(Z_i, d(Z_i, \mathcal{Y})) \leq C'\varepsilon_n^2.$$

that concludes the proof. □

Proof of Proposition 3

Proof. First let us prove that balls roll freely inside and outside S. Indeed a direct consequence of Corollary 4.9 in [14] is that $\text{reach}(S) \geq \rho_0 - \rho_1 > 0$ that balls of radius $\rho_0 - \rho 1$ roll freely outside S. Now for all $x \in \partial S$, as $d(x, M) \leq \rho_1 \leq rho_0$ there exists (a unique) $x^* \in M$ such that $||x - x^*|| = d(x, M)$, because $x \in \partial S$ we must have $||x - x^*|| = \rho_1$ and, by definition of S, $B(x^*, \rho_1) \subset S$ so that balls of radius ρ_1 roll freely inside S. We so have

$$\text{Balls of radius } r_0 = \min(\rho_0 - \rho_1, \rho_1) > 0 \text{ roll freely inside and outside} S. \tag{19}$$

Introduce $n(x, M)$ the space normal to M at the point x $\text{nor}(M) = \{(x, u), x \in M, u \in n(x, M), ||u|| = 1\}$ Now, proposition 16 in [27] says that: $\varphi : \partial S \to \text{norn}(M)$, $\varphi(x) = (x^*, \frac{x - x^*}{\rho_1})$, where x^* is the projection of x onto M, is bijective. This bijectivity implies that $\mathcal{M}(S) = M$. Indeed, it first implies the existence of normal vectors, for $x \in M$, let u be a unit vector of $n(x, M)$ then $x + \rho_1 u$ and $x - \rho_1 u$ are two different points of ∂S and there is no points of ∂S closer of x than ρ_1 so that x has at least two different projections and $x \in \mathcal{M}(S)$. Reversely if $x \in \mathcal{M}(S)$ then there exists x_1 and x_2 two different projection of x on ∂S so that, by the inside rolling ball property we have $d(x, \partial S) = \rho_1$ that finally implies that $x \in M$.

$$M = \mathcal{M}(S) \tag{20}$$

□

About Corollary 3 and Conjecture 2.3 The corollary 3 is a direct consequence of Proposition 3, Corollary 1 and section 2.2 in [17] where it is said that S the support of X as a density bounded away from 0 on its support $\bigcup_{x, \in M} \overline{B}(x, \rho_0)$.

About Conjecture 2.3 we think that the density of X satisfy that $f_X(x) \geq X d(x, \partial S)^{d'}$ for all $x \in \bigcup_{x, \in M} \overline{B}(x, \rho_0)$. Indeed we have take $x \in \mathring{S}$ such that $B(x, \varepsilon) \subset \mathring{S}$ we have $\mathbb{P}_X(B(x, \varepsilon)) \geq \int_{y \in M, ||x-y|| \leq \rho_1} c\omega_d \varepsilon^d dy$ so that $\mathbb{P}_X(B(x, \varepsilon)) \geq |\{y \in M, ||x - y|| \leq \rho_1\}|_{d'} c\omega_d \varepsilon^d$ (where ω_k is the volume of the k-dimensional unit ball and $|A|_k$ the k-dimensional measure). Now using Pythagoras, introducing x^* the projection of x onto M we have

$$\mathbb{P}_X(B(x, \varepsilon)) \geq |M \cap B(x^*, \sqrt{\rho_1^2 - ||x - x^*||^2})|_{d'} c\omega_d \varepsilon^d$$

And because $\rho_1^2 - ||x - x^*||^2 = d(\partial S, x)$ we obtain

$$\mathbb{P}_X(B(x, \varepsilon)) \geq |M \cap B(x^*, d(\partial S, x))|_{d'} c\omega_d \varepsilon^d$$

. when $d(\partial S, x)$ is small enough $M \cap B(x^*, d(\partial S, x)$ looks like a part of a d'-dimensional ball the "worst" case being when x^* is on the ∂M, but the assumed regularity of ∂M ensures that in this case $|M \cap B(x^*, d(\partial S, x))|_{d'} \sim \frac{\omega_{d'}}{2} d(\partial S, x)^{d'}$.

Such a density is not sufficient to apply the results on the r-convex hull obtained in [23] but we really think that the asymptotic is the same than in [1] so that, if $f_X(x) \geq X d(x, \partial S)^\alpha$ we will have $\mathcal{X}_n \cap C_r(\mathcal{X}_n)$ is a $((\ln n/n)^{2/(d+1+2\alpha)}, (\ln n/n)^{1/(d+1+2\alpha)})$ estimation of ∂S.

3 Practical aspects

Now we are going to detail the practical aspect related to the estimator proposed in Corollary 1 that is we aim to detail the algorithmic point of view and propose some tools for the choice of the parameters. At the end of each section we propose the associated algorithms in pseudo-code. Note that they require the use of a programming language that can compute Voronoi cells in any dimension. More precisely for $X = \{X_1, \ldots, X_n\}$ a set of n points in $\mathbb{R}^d$ we need to obtain a set of points $V = \{V_1, \ldots V_m\}$ and a set of set of index $\{J_1, \ldots J_n\}$ ($J_k \subset \{1, \ldots m\}$) such that $\mathrm{Vor}_X(X_i) =$ Convex Hull of $\{V_j; j \in J_i\}$. That is the usual output for such a topic. For instance, using matlab it is the result of $[V, J] =$voronoin(X). Using Python it can be obtained by $VV =$VoronoiTess(X), $V = VV$.vertices, $J = VV$.regions. In the proposed algorithm we will choose the matlab denomination: $[V, J] =$voronoin(X) .

It also requires a Delaunay triangulation function that allows to obtain from X a set $T = \{T_1, \ldots T_{m'}\}$, $T_i \in \{1, \ldots n\}^{d+1}$ where $\{X_t, T \in T_i\}$ is a Delaunay simplex of X. In matlab it is obtained via $T =$delaunayn(X), in Python $TT =$DelaunayTri(X) and $T = TT$.vertices. Once again, in the proposed algorithm we will use the matlab denomination. Notice that: exists k such that $\{i, j\} \subset T_k$ is equivalent to $\mathrm{Vor}_X(X_i) \cap \mathrm{Vor}_X(X_j) \neq \emptyset$

It also needs a ncc(G) function that returns the number of connected components of a graph represented by its adjacency matrix G, $G_{i,j} > 0$ if i and j are connected on the graph.

3.1 Identification of points close to the boundary

Notice that the identification of $\mathcal{Y}_r = C_r(\mathcal{X}_n) \cap \mathcal{X}_n$ can be easily computed. Indeed, the following proposition stands that points located close to the boundary are the one that have "large" Voronoi cells.

Proposition 4. *$X_i \in \mathcal{Y}_r$ if and only if* $\max\{||y - X_i||, y \in (Vor_{\mathcal{X}_n}(X_i)) \geq r\}$.

Proof. Observe that $\max\{||y - X_i||, y \in (\mathrm{Vor}_{\mathcal{X}_n}(X_i))\} \geq r$. This implies that there exists x such that $B(x,r) \cap \mathcal{X}_n = \emptyset$. Thus, by definition of $\hat{C}_r$, $B(x,r) \cap \hat{C}_r(\mathcal{X}_n) = \emptyset$ and $x_n = (1/n)x + (r - 1/n)X_i \to X_i \in \hat{C}_r^c(\mathcal{X}_n)$ with $X_i \in \hat{C}_r(\mathcal{X}_n)$, so $X_i \in \partial\hat{C}_r(\mathcal{X}_n)$. Conversely, if $X_i \in \partial\hat{C}_r(\mathcal{X}_n)$ then there exist two sequences x_n and $y_n \in \hat{C}_r^c(\mathcal{X}_n)$ such that $x_n \to X_i$, $x_n \in B(y_n, r)$ and $B(y_n, r) \cap \mathcal{X}_n = \emptyset$. We have $y_n \in S \oplus rB$, which is compact. Thus, up to an extraction we can suppose that $y_n \to y$. As $r < ||y_n - X_i|| \leq r + ||x_n - x||$ we have in the limit $||y - X_i|| \geq r$. Moreover, since for all n, $B(y_n, r) \cap \mathcal{X}_n = \emptyset$, we have $B(y, r) \cap \mathcal{X}_n = \emptyset$ and therefore $y \in \mathrm{Vor}_{\mathcal{X}_n}(X_i)$. □

The crucial point when choosing a value for r is to identify observations that are really close to the boundary of S. In [24] one can find fully data driven way to select r. Unfortunately this method is based on the fact that the data is uniformly drawn, and, more annoying it appears very difficult to compute it when the dimension is higher than two. It is why we will propose a more rough way to choose r. Let us introduce $r_i = \sup_{x \in V_i} ||X_i - x||$. One can clearly guess that, for all i such that X_i is far enough from the boundary, r_i is small. For instance, under our last hypotheses (S compact r_0-smooth and f bounded away from 0 on S) the maximal spacing theory ([18] for the original paper and [2] for the extension to the same hypotheses as in Corollary 1) ensures that there exists μ such that for all i such that $d(X_i, \partial S) \geq \mu(\ln n/n)^{1/d}$ we have $r_i \leq \mu(\ln n/n)^{1/d}/2$. Reversely, by [**?**] we now that for all $r_1 < r_0$, for n large enough there exists a $\mathcal{Y}_{r_1}$ is an accurate estimation of the boundary. The number of observation N_{r_1} in $\mathcal{Y}_{r_1}$ can be neglected with regard to the total number of observation n (roughly because $\mathbb{E}(N_{r_1}/n) \leq \max f \mu(\ln n/n)^{1/d}$) we propose to consider the large values of r_i as outsiders and to detect them using a classical outliers detection based on quartile. More precisely we propose the following algorithm for the choice of r:

Algorithm 1: Identification of Boundary observations

Data: $X = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ the observations
Result: Y=boundaryobs(X) such that $Y = X \cap \partial C_r(X)$
$[V, J]$ =voronoin(X) ;
for $i = 1$ *to* n **do**
  $r_i := \max_{j \in J_j} ||V_j - X_i||$
end
$R = \{r_1, \ldots, r_n\}$, ;
while $\max(R) \geq Q_{75}(R) + 3(Q_{75}(R) - Q_{25}(R))$ **do**
  $\rho = Q_{75}(R) + 3(Q_{75}(R) - Q_{25}(R))$, $R = \{r_i, r_i \leq \rho\}$
end
$Y = \{X_i$ such that $r_i \geq \rho\}$;

3.2 Computation of the estimated medial axis

Let us recall that the proposed medial axis estimator is the following:

$$\hat{\mathcal{M}}_\lambda(S_n, \mathcal{Y}) = \left\{ x \in \mathrm{Vor}_{\mathcal{Y}}(y) \cap \mathrm{Vor}_{\mathcal{Y}}(z) \cap \hat{S}_n, (y,z) \in \mathcal{Y}^2, ||y-z|| \geq \lambda \right\}.$$

To get an easy computation of $\hat{\mathcal{M}}_\lambda(S_n, \mathcal{Y})$ one have to notice that:

Proposition 5. *under the hypothesis of theorem 1, for n large enough, if* $(y,z) \in \mathcal{Y}^2$ *such that* $||y-z|| \geq \lambda$ *we have :*

$$Vor_{\mathcal{Y}}(y) \cap Vor_{\mathcal{Y}}(z) \cap \hat{S}_n \neq \emptyset \Longleftrightarrow (Vor_{\mathcal{Y}}(y) \cap Vor_{\mathcal{Y}}(z)) \subset \hat{S}_n$$

Proof. Proceeding by contradiction, let us suppose that there exists $(y,z) \in \mathcal{Y}^2$ such that $||y-z|| \geq \lambda$ with $\mathrm{Vor}_{\mathcal{Y}}(y) \cap \mathrm{Vor}_{\mathcal{Y}}(z) \cap \hat{S}_n \neq \emptyset$ and $\mathrm{Vor}_{\mathcal{Y}}(y) \cap \mathrm{Vor}_{\mathcal{Y}}(z) \cap \hat{S}_n^c \neq \emptyset$. Because $\mathrm{Vor}_{\mathcal{Y}}(y) \cap \mathrm{Vor}_{\mathcal{Y}}(z)$ is connected there exists a point $x \in \mathrm{Vor}_{\mathcal{Y}}(y) \cap \mathrm{Vor}_{\mathcal{Y}}(z) \cap \partial\hat{S}_n$. There also exists a point $x^* \in \partial S$ with $||x - x^*|| \leq \varepsilon_n'$ and a point $y' \in \mathcal{Y}$ such that $||y' - x^*|| \leq \varepsilon_n$ so that, if $x \in \mathrm{Vor}_{\mathcal{Y}}(z')$ then $||z' - x|| \leq \varepsilon_n + \varepsilon_n'$. That, in turns imply that $\lambda \leq 2(\varepsilon_n + \varepsilon_n')$ that is impossible for n large enough. □

Thus, one can decide to compute

$$\hat{\mathcal{M}}_\lambda^*(S_n, \mathcal{Y}) = \bigcup_{(i,j)\in I} \mathrm{Vor}_{\mathcal{Y}}(Y_i) \cap \mathrm{Vor}_{\mathcal{Y}}(Y_j)$$

$$I = \{(i,j), ||Y_i - Y_j|| \geq \lambda, \text{ All the vertex of } \mathrm{Vor}_{\mathcal{Y}}(Y_i) \cap \mathrm{Vor}_{\mathcal{Y}}(Y_j) \text{ are in } \hat{S}_n\}$$

instead of $\hat{\mathcal{M}}_\lambda(S_n, \mathcal{Y})$ since, for n large enough both estimator coincide.

Now, we can detail the proposed algorithm. We propose first algorithm that returns:

- all the $\mathrm{Vor}_{\mathcal{Y}}(Y_i) \cap \mathrm{Vor}_{\mathcal{Y}}(Y_j)$ such that all its vertex are in $\hat{S}_n$
- the associated value of $||Y_i - Y_j||$ denoted $\lambda_{(i,j)}$

So that $\hat{\mathcal{M}}^*_{\lambda,S_n}(\mathcal{Y}) = \cup_{(i,j),\lambda_{i,j}\geq\lambda} \mathrm{Vor}_{\mathcal{Y}}(Y_i) \cap \mathrm{Vor}_{\mathcal{Y}}(Y_j)$ can be easily computed in a second step. This is described in Algorithm 2.

3.3 Choice of the λ parameter

The choice of λ is primordial to obtain satisfying results. We so propose different graphical tools to help the user to, a posteriori, choose λ. Note that, most of this tools are extremely costly so that we propose to compute them only on 20 different values of λ. Throughout this section we are going to illustrate our indicators on the following example. 50000 points has been uniformly drawn on a simplified version of the Da Vinci Vitruvian man see Figure 8.

Algorithm 2: Computation of $\hat{\mathcal{M}}^*$

Data: $X = \{X_1, \dots, X_n\} \subset \mathbb{R}^d$ the observations , $Y = \{Y_1, \dots, Y_m\}$ the boundary observation

Result: $Z = \{Z_1, \dots, Z_m\} \subset \mathbb{R}^d$ a list of point, $I = \{I_1, \dots I_k\}$ a list of sets with $I_j \subset \{1, \dots m\}$ and $\Lambda = \{\lambda_1, \dots \lambda_k\} \subset \mathbb{R}$ such that:

$$\hat{\mathcal{M}}^*_\lambda(X \cap \partial C_r(X), \hat{S}_\varepsilon) = \bigcup_{i, \lambda_i \geq \lambda} \text{Convex Hull of}(\{Z_j, j \in I_i\}).$$

$[Z, J]$ =voronoin(Y), TT =delaunayn(Y);
$\varepsilon = 4^{1/d} \min_i(\max_{j \neq i} ||X_i - X_j||)$;
Initialization: $T = \emptyset$, $\Lambda = \emptyset$ and $I = \emptyset$;
for $i = 1$ *to the number of simplices in* TT **do**
  for $j = 1$ *to* d **do**
    for $k = j + 1$ *to* d **do**
      $(i_0, j_0) = (TT_i(j), TT_i(k))$;
      if $\{i_0, j_0\} \cap T = \emptyset$ **then**
        $T = T \cup \{i_0, j_0\}$;
        if $\max_{j \in J_{i_0} \cap J_{j_0}}(\min_i ||Z_j - X_i||) \leq \varepsilon$ **then**
          $\Lambda = \Lambda \cup \{||Y_{i_0} - Y_{j_0}||\}$;
          $I = I \cup \{J_{i_0} \cap J_{j_0}\}$
        end
      end
    end
  end
end

Figure 8: The original image, S, the sample of points (in black the one identified as boundary point by our algorithm)

1: Reconstruction performances First recall that the medial axis is linked to image compression and that one of its property is that $S = \bigcup_{x \in \mathcal{M}(S)} \overline{B}(x, r(x))$. The most natural idea should be to consider

$$D_H(\lambda) = d_h(\mathcal{X}_n, \bigcup_{x \in \hat{\mathcal{M}}_\lambda^*}, \overline{B}(x, d(x, \mathcal{Y})).$$

Unfortunately that is computationally very expensive. Indeed instead of considering a infinite union of balls and we propose to consider only points located at the vertices of the intersections of the Voronoi cells in $\hat{\mathcal{M}}_\lambda^*$. Moreover the maximum distance from a point of the reconstructed set to an observation is difficult to compute. Hopefully it is not the most important part of the distance. We so propose to compute the following indicator. Let Z_λ be the set of vertices of the $\text{Vor}_{\mathcal{Y}}(Y_i) \cap \text{Vor}_{\mathcal{Y}}(Y_j)$ that are a face of of $\hat{\mathcal{M}}_\lambda^*$. Introduce

$$\tilde{S}_\lambda = \bigcup_{z \in Z_\lambda} \overline{B}(z, d(z, \mathcal{Y}))$$

and compute:

$$C_1(\lambda) = \max_i \left(d(X_i, \tilde{S}_\lambda) \right).$$

See Figure 9 to observe that:

i The part $\max_{x \in \tilde{S}_\lambda}(d(x, \mathcal{X}_n))$ can be neglected.

ii The reconstruction only through the vertex is a good approximation

iii The indicator helps to detect too large values for λ, here clearly $\lambda = 0.07$ is too large but is not that useful to detect the "best" value for λ (here we should expect a reconstruction with $\lambda \in [0.04, 0.05]$.

The way to compute $C_1(\lambda)$ is given in Algorithm 3.

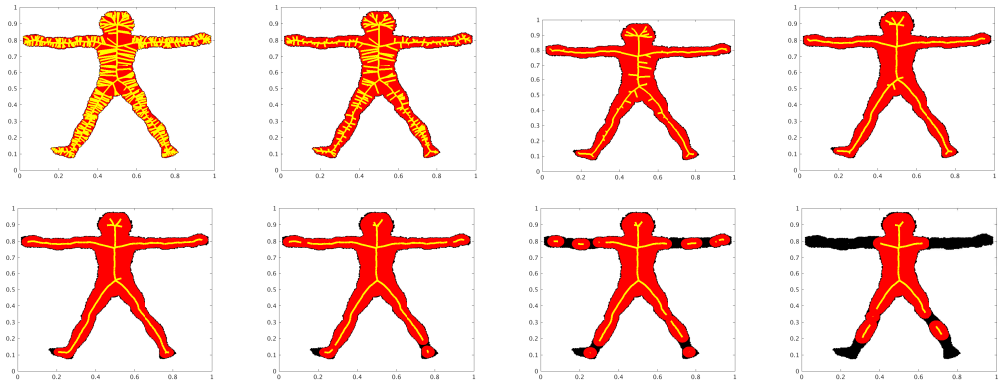

Figure 9: For $\lambda \in \{0, 0.01, \ldots, 0.07\}$; Yellow the associated $\hat{\mathcal{M}}^*_\lambda$, in red the reconstructed Sets $\tilde{S}_\lambda$ and in black the samples points, when λ is small the reconstructed set hides the sample points but when λ decreases $\tilde{S}_\lambda$ underestimates S

Algorithm 3: Computation of $C_1(\lambda)$

Data: $X = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ the observations, $Y = \{Y_1, \ldots, Y_m\}$ the boundary observations Z, I and Λ the outputs of Algorithm 2 and λ a real number

Result: $C_1(\lambda)$

Collect the vertices $C = \bigcup_{i,\lambda_i > \lambda} \left(\bigcup_{j \in I_i} \{Z_j\} \right)$;

$$C_1(\lambda) = \max_i \left(\min_j \left[\max \left(||X_i - C_j|| - \max_k ||C_j - Y_k||, 0 \right) \right] \right)$$

2: Number of connected components of $\hat{\mathcal{M}}^*_\lambda$. If we previously said that we should expect a reconstruction with $\lambda \in [0.04, 0.05]$ it is because we expect that the estimated medial axis is, not only close, in distance, to the medial axis but, but we also want it to have topological properties closed to the one of the original medial axis. For instance we aim to recover the number of connected components of the medial axis. Throughout our simulation study we observed the following, in general the number of connected components start to the number of connected components of S for $\lambda = 0$ then slowly increases (or stay constant) to the number of connected components of $\mathcal{M}(S)$ then increases because of suppression of faces, then decreases when we have only few remaining faces. We so propose to compute $C_2(\lambda)$ the number of connected components of $\hat{\mathcal{M}}^*_\lambda$. The way to compute $C_2(\lambda)$ is given in Algorithm

4.

3: Number of connected components of the "extremities" of $\hat{\mathcal{M}}_\lambda^*$ We also can compute $C_3(\lambda)$ the number of connected components of the extremities of $\hat{\mathcal{M}}_\lambda^*$. Here we define the extremities as a generalization of the boundary in a "manifold" sense: Let A be a compact set in $\mathbb{R}^d$ one can define $E_k(A)$ the set of its k-dimensional extremities $x \in A$ such that, there exists r_0 such that $B(x, r_0) \cap A$ is homeomorphic to $H_k = \{(x_1, \ldots, x_k), \sum x_i^2 < 1, x_1 \geq 0\}$. Obviously, if A is a k dimensional manifold then $E_k(A) = \partial A$.

Here, as we are interested in $\hat{\mathcal{M}}_\lambda^*$ which is a union of $(d-1)$-dimensional convex compact polygons we propose to compute $E_{d-1}(\hat{\mathcal{M}}_\lambda^*)$ the set of its $(d-1)$-dimensional extremities.

Ones again consider our example and observe that, when λ is too small the $\hat{\mathcal{M}}_\lambda^*$ as a lot of "extremities" that roughly counts the number of parasite branch of the medial axis. Increasing λ we expect, this number decreases to reach a minimum value for suitable λ then $C_3(\lambda)$ increases as $\hat{M}_\lambda^*$ may have different connected components or because "holes" in $\hat{M}_\lambda^*$.

The way to compute $C_3(\lambda)$ is given in Algorithm 4.

Algorithm 4: Computation of $C_2(\lambda)$ and $C_3(\lambda)$

Data: $X = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ the observations, $Y = \{Y_1, \ldots, Y_m\}$ the boundary observations Z, I and Λ the outputs of Algorithm 2 and λ a real number

Result: $C_2(\lambda)$ and $C_3(\lambda)$

Triangulate the estimated medial axis, initialization: $T = \emptyset$;

for $i = 1$ *to size*(Λ) **do**

  if $\lambda_i \geq \lambda$ **then**

    $t =$delaunayn$(\{\pi(Z_i), i \in I_i\})$;

    $t = T \cup t$

  end

end

$G =$graphtri(T) and $C_2(\lambda) =$ncc(G);

if $d = 2$ **then**

  $T_e =$extremities(T), $C_3(\lambda) = \#(T_e)$

else

  $T_e =$extremities(T), $G_e =$graphtri(T_e) and $C_3(\lambda) =$ncc(G_e)

end

Algorithm 5: Function graphtri

Data: $T = \{T_1, \dots T_k\}$ a list of d' dimensional simplex, i.e $T_i \subset \{1, \dots, N\}^{d'+1}$

Result: G a graph

N is the maximum integer in T, G_0 is the null $N \times N$ matrix, n_0 is the null N vector ;

for $i = 1$ *to* k **do**
  for $j = 1$ *to* d' **do**
    for $k = j + 1$ *to* d' **do**
      $G_0(T_i(j), T_i(k)) = 1$; $G_0(T_i(k), T_i(j)) = 1$; $n_0(T_i(j)) = 1$ and $n_0(T_i, k) = 1$
    end
  end
end

$N_{\text{pt}} = \sum_i n_0(i)$, G is the null $N_{\text{pt}} \times N_{\text{pt}}$ matrix ;

$i0$=0;

for $i = 1$ *to* N **do**
  if $n_0(i) > 0$ **then**
    $i_0 = i0 + 1$, $j_0 = 0$;
    for $j = 1$ *to* N **do**
      if $n_0(j) > 0$ **then**
        $j_0 = j_0 + 1$, $G(i_0, j_0) = 1$
      end
    end
  end
end

Algorithm 6: Function extremities

Data: $T = \{T_1, \dots T_k\}$ a list of d' dimensional simplices, i.e
$T_i \subset \{1, \dots, N\}^{d'+1}$

Result: T' the list of $d'-1$ dimensional extremity simplices

$T'' = \emptyset$, nb $= \emptyset$;

for $i = 1$ *to* k **do**
  for $j = 1$ *to* d' **do**
    $t = \{T_i(1), \dots, T_i(j-1), T_i(j+1), \dots, T_i(d'+1)\}$; **if** $t \in T'$ **then**
      find j such that $t = T'_j$, nb$(j) :=$ nb$(j) + 1$
    else
      $T'' := T'' \cup \{t\}$, nb $=$ (nb, 1)
    end
  end
end

$T' = \emptyset$ **for** $i = 1$ *to* $\#(T")$ **do**
  if $nb(i) = 1$ **then**
    $T' = T' \cup \{T''(j)\}$
  end
end

4: Estimation of the density. The previous algorithm provides Λ, the list of the different values for λ as output. We can see the points of Λ as the results of the mixture of two laws, one containing the small values (parasite branches) and the other one containing the largest values (stable part of the medial axis). Here we have no a priori idea of the proportion in each part of the mixture and the outliers detection approach is no more convenient. We propose to estimate the density of the λ (in the simulation part we used a Kernel density estimator with the Sheater and Jones [26] procedure for the bandwidth selection that exists in main programming languages : this bandwidth is the default one in matlab function ksdensity and can be obtain with hsj function in python). If a multi-modality is observed one can guess that a suitable value for λ is located near a local minima of the density.

This is clearly the indicator that is the less motivated but, as illustrated by Figure 10 it gives some accurate results (in this Figure one can observe that the most significant local minima of the density is located for $\lambda \sim 0.04$ that is a satisfying value for the estimation of the medial axis).

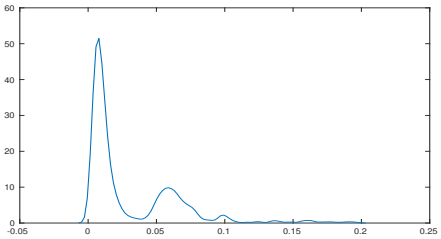

Figure 10: Estimated density of on Λ

3.4 Some simulations

The aim of this simulation section is first to illustrate the convergence of the proposed estimator theoretically obtained, and second to observe how behave the different indicators that has been proposed to help the user in choosing a suitable λ parameter. This second points being, in our opinion, the most important one. Indeed there exists other medial axis estimator and ours is closely related to them. It is so expected to behave as other ones and the novelty is in the theoretical study of the estimator based on a boundary estimation. But, up to our knowledge, there exits no way to help in choosing the smoothing parameter which is our main contribution.

First we present some results related to the manifold estimation. In the three first examples Y is drawn on $M \subset \mathbb{R}^d$ then we observe $X = Y + 0.5U$ where U is uniformly drawn on the d-dimensional unit ball.

a. $d = 2$ and M is the unit circle in $\mathbb{R}^2$, see Figure 11. Since $n = 500$ we can observe that the density, the number of connected components and the number of the connected components of the extremities has a local associated to a small reconstruction error. The choice of the larger λ such that the two numbers of connected components are in a local minima then provide a medial axis estimation that is close to the initial circle. For smaller values of n we have not reach the convergence. When $n = 100$ the procedure fails in recognize the boundary, when $n = 200$ the choice of a suitable λ according to the number of connected components of the extremities provides almost good results expect that there exists a residual part of the outer medial axis.

b. $d = 3$ and M is a trefoil knot (which is a 1-dimensional manifold), see Figure 12. Since $n = 5.10^3$ we start to observe local minima in the numbers of connected components and a second mode apparition in the density. Unfortunately if the different values are quite close they are not located at the same place before $n = 2.10^4$. Nevertheless we chose to select λ such that the sum of the connected components is locally minimum that gives correct medial axis estimation since $n = 5.10^3$.

c. $d = 3$ and M is a Moebus ring (which is a 2-dimensional manifold), see Figure 13. The analyze of the different indicators that help in choosing λ is similar to the trefoil knot case. The difference here is that the medial axis estimator only gives correct results since $n = 2.10^4$.

We also test our program on different images that do not satisfy the regularity assumption. For such images the choice of the λ parameter is more difficult. It start to be "easy" for sample size much larger than the one of the disk of Figure 11 and most of the time we have to neglect one of the indicators. The bird image Figure 14. We have chosen λ close to the first local minima of the number of connected components. This number is also a local minima of the number of connected components of the extremities (or has a small associated number). Its only correspond to a local minima of the density for sample sizes $n \geq 5000$. and it is always associated to a quite large reconstruction error (due to the sharpness of the wings). The leaf image in Figure **??** example is more or less similar to the bird one. It is a bit easier because angles are more soft.

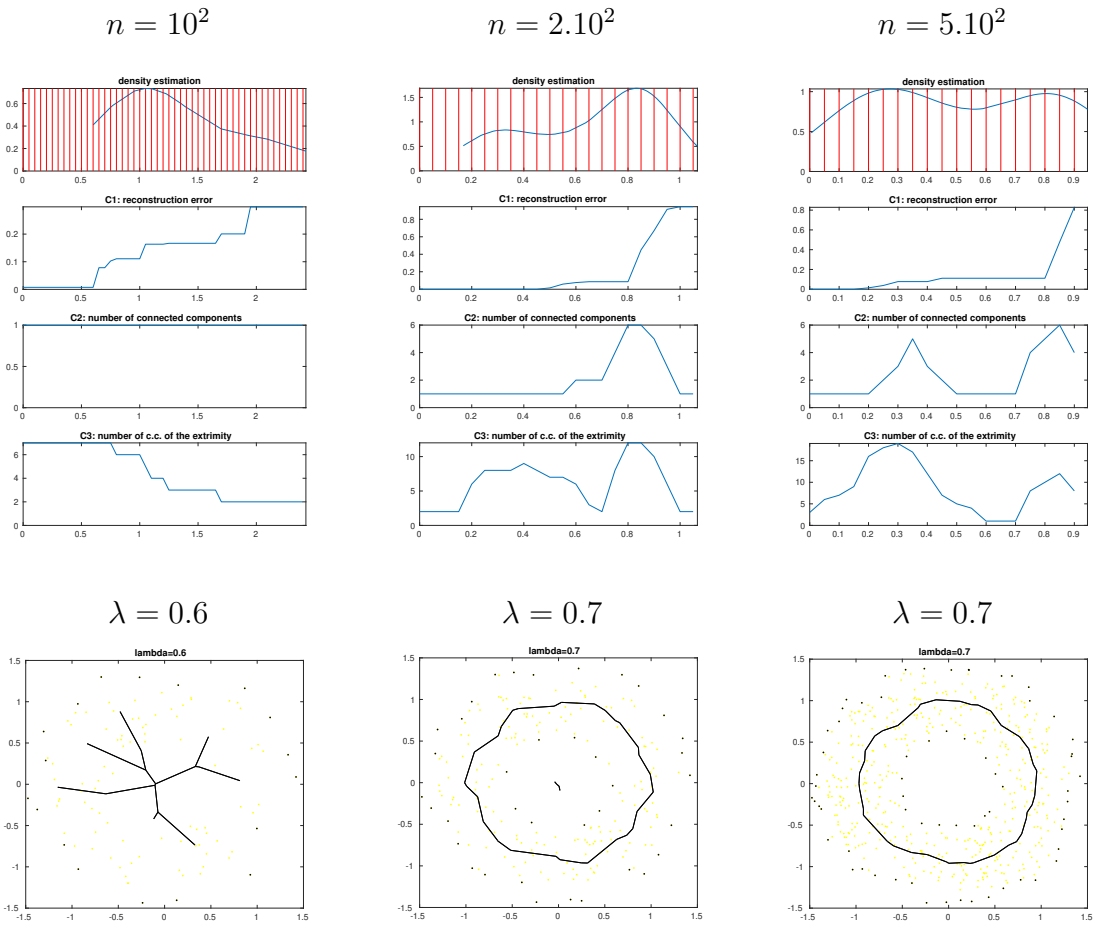

Figure 11: Data drawn on a $\overline{B}(O, 1.5) \setminus \overline{B}(O, 0.5)$. Each column correspond to a sample size. For each sample size we plot the density estimator computed on Λ (the red vertical lines are the tested values for λ), the reconstruction error function C_1, the number of connected component of the estimated medial axis function C_2 and the number of connected components of the extremities of the estimated medial axis C_3. The chosen λ is indicated then we plot the data sample, the boundary observations are highlighted in black, and the medial axis graph is plotted.

4 Future Work

Some various questions are still open, specially dealing with the manifold estimation problem. Recall that there exists a random variable Y whose distribution is sup-

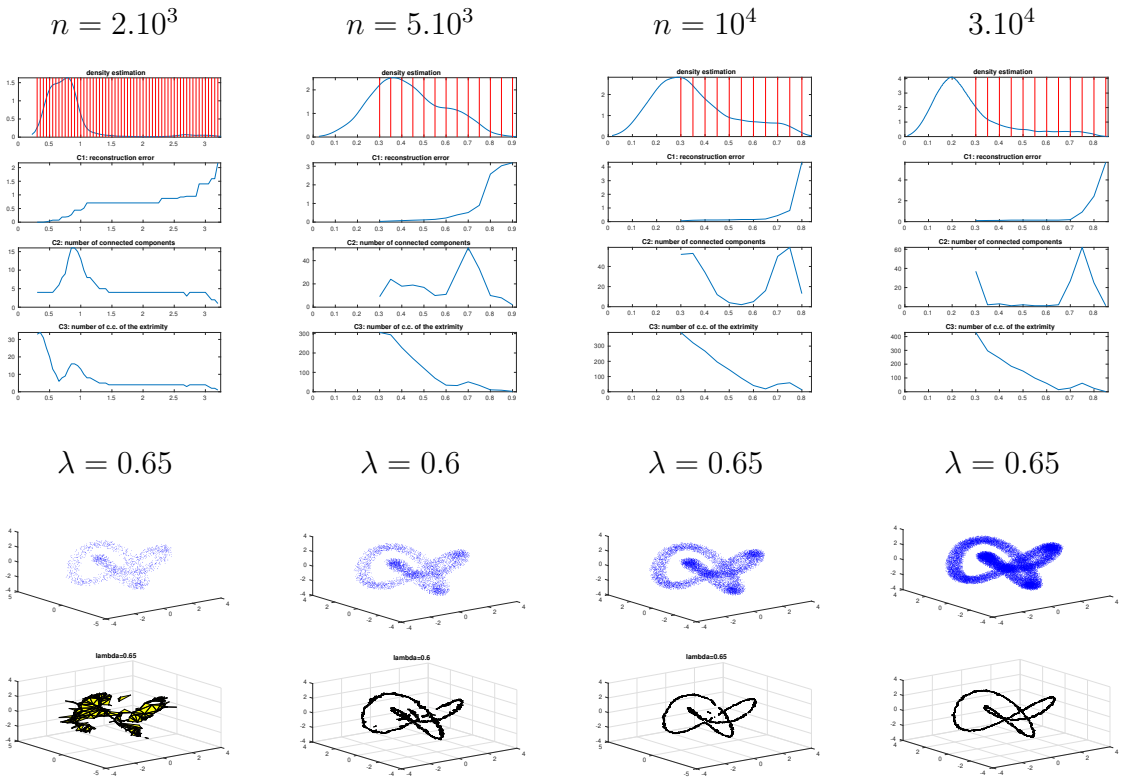

Figure 12: Data drawn on a noised Trefoil knot. Each column correspond to a sample size. For each sample size we plot the density estimator computed on Λ (the red vertical lines are the tested values for λ), the reconstruction error function C_1, the number of connected component of the estimated medial axis function C_2 and the number of connected components of the extremities of the estimated medial axis C_3. The chosen λ is indicated. The data is plotted. The last graph is the medial axis estimator for the chosen λ

ported by M a d'-dimensional sub-manifold of $\mathbb{R}^d$ and we observe $X = Y + U$ with $U|Y = y$ that has a support $B(y, \rho_y)$. We have seen that, when ρ_y is constant and inferior to a regularity measure then we had $\mathcal{M}(\{z \in \mathbb{R}^d, \exists y \in M, ||z - y|| \leq \rho_y\} = M$. We expect that we can allow smooth variation of ρ_y and still have this good property that can be assimilated to the fact that M is identifiable. We are yet almost sure that when $d = 2$ the $K-$Lipschitz continuity (with $K < 1$) of ρ_y allows such a good property. We also wonder if we can find a optimal method (with the rate obtained

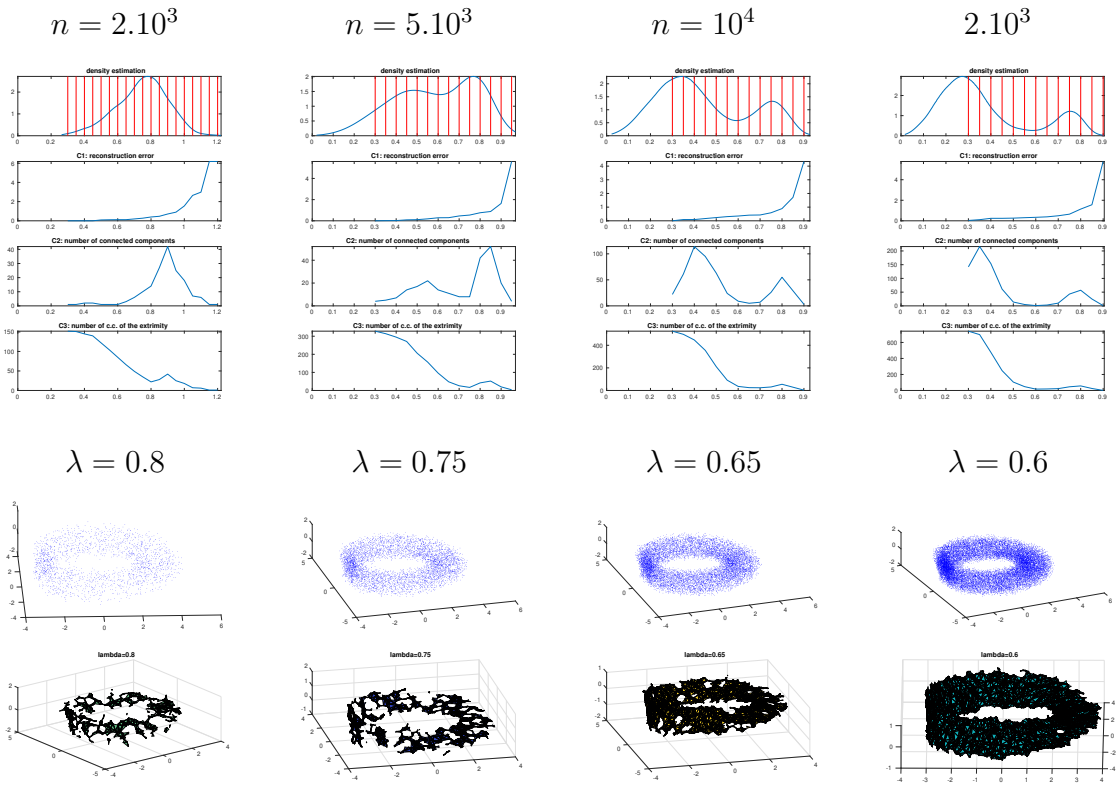

Figure 13: Data drawn on a noised Moebus ring.

in [17]) running preliminary a medial axis algorithm, that allows to estimate d' then, use the knowledge of d' to obtain a optimal method when $d' < d - 1$.

References

[1] C. Aaron and O. Bodart. Local convex hull support and boundary estimation. *J. Multivariate Anal.*, 147:82–101, 2016.

[2] C. Aaron, A. Cholaquidis, and A. Fraiman. A generalization of the maximal-spacings in several dimensions and a convexity test. *Extremes*, 10:605–634, 2017.

[3] N. Amenta, S. Choi, and R. Kolluri. The power crust, unions of balls, and the medial axis transform. *Computational Geometry*, 19:127–153, 2001.

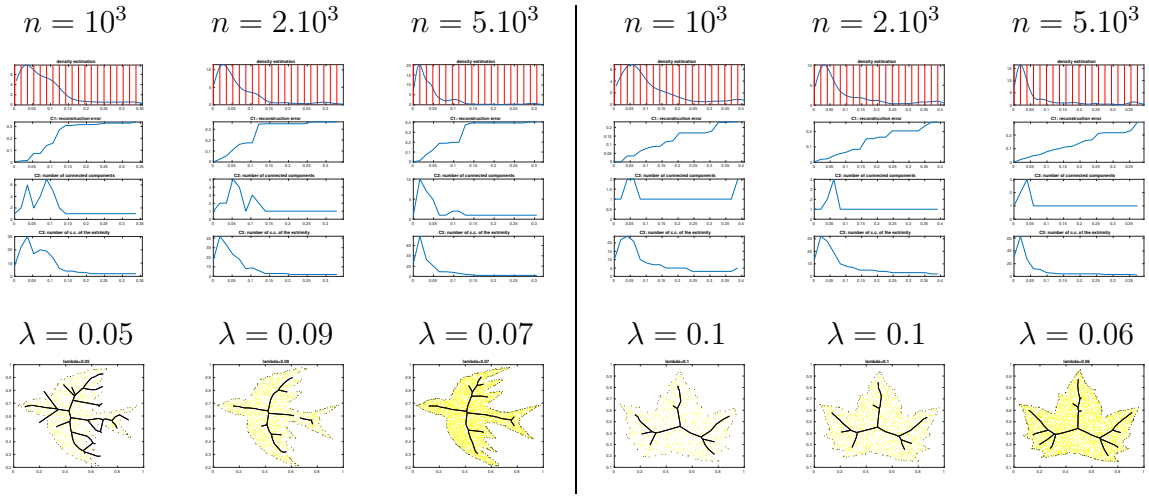

Figure 14: Data uniformly drawn on bird image.

[4] D. Attali, J. Boissonnat, and E. Edelsbrunner. *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration.* Springer, 2009.

[5] D. Attali and A. Montanvert. Modeling noise for a better simplification of skeletons. In *Proc. of 3rd IEE Internat. Conf. Impage Process*, 1996.

[6] Harry Blum. A Transformation for Extracting New Descriptors of Shape. In Weiant Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pages 362–380. MIT Press, Cambridge, 1967.

[7] J. W. Brandt and V. R. Algazi. Continuous skeleton computation by voronoi diagram. *CVGIP: Image Understanding*, 55:329–338, 1992.

[8] F. Chazal and A. Lieuter. The "$\lambda-$medial axis". *Graphical Models*, 67(4):304–331, 2005.

[9] F. Chazal and R. Soufflet. Stability and finiteness properties of medial axis and skeleton. *J. Dyn. Control Syst.*, 10:149–170, 2004.

[10] S. W. Choi and H.-P. Seidel. Linear one-sided stability of mat for weakly injective 3d domain. In *Proc. 7th ACM Sympos. Solid Modeling Appl.*, pages 344–355, 2002.

[11] A. Cuevas, P. Llop, and B. Pateiro-Lopez. On the estimation of the medial axis and inner parallel body. *J. Multivariate Anal.*, 129:171–185, 2014.

[12] A. Cuevas and A. Rodriguez-Casal. On boundary estimation. *Adv. Appl. Probab.*, 36(2):340–354, 2004.

[13] L. Devroye and G.L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *Siam J. Appl. Math.*, 38(3):480–488, 1980.

[14] H. Federer. *Geometric measure theory*. Springer-Verlag, 1969.

[15] D. Fritsch, S. Pizer, B. Morse, D. Eberly, and A. Liu. The multiscale medial axis and its applications in image registration. *Pattern Recognit Lett.*, 15:445–452, 1994.

[16] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. The geometry of nonparametric filament estimation. *J. Am. Stat. Assoc.*, 107(498), 2012.

[17] C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.*, 13:1562–1582, 2012.

[18] S. Janson. Maximal spacings in several dimensions. *Ann. Prob.*, 15:274–280, 1987.

[19] G. Matheron. *Image Analysis and Mathematical Morphology, Volume 2: Theoretical Advances*. Academic Press,London, 1988.

[20] A. Montero and J. Lang. Skeleton pruning by contour approximation and the integer medial axis transform. *Comput. Graphics*, 36:477–487, 2012.

[21] D. S. Paik, C. F. Beaulieu, R. Brooke Jeffrey, G. D. Rubin, and S. Napel. Automated flight path planning for virtual endoscopy. *Medical Physics*, 25:629–637, 1998.

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

[23] A. Rodríguez-Casal. Set estimation under convexity type assumptions. *Ann. I. H. Poincaré B.*, 43(6):763–774, 2007.

[24] A. Rodríguez-Casal and P. Saavedra-Nieves. A fully data-driven method for estimating the shape of a point cloud. *Esaim probab. stat.*, 20:332–348, 2016.

[25] D. Shaked and A. Bruckstein. Pruning medial axes. *Comput Vis. Image Underst.*, 69(2):156–169, 1998.

[26] S.J. Sheather and M.C. Jones. A reliable data-based bandwidht selection method for kernel density estimation. *J. R. Stat. Soc. B*, 53:683–690, 2012.

[27] C. Thäle. 50 years sets with positive reach -a survey-. *Surveys in Mathematics and its Applications*, 3:123–165, 2008.

[28] G. Walther. On a generalization of Blaschke's rolling theorem and the smoothing of surfaces. *Math. Methods Appl. S.*, 22(4):301–316, 1999.

[29] S. Xia, N. Ding, M. Jin, H. Wu, and Y. Yang. Medial axis construction and applications in 3d. In *Proc. IEE Infocom*, pages 305–309, 2013.

[30] L. Younes. *Shapes and Diffeomorphisms*. Springer-Verlag Berlin Heidelberg, 2010.