Using the $k-$nearest neighbor restricted Delaunay complex to estimate the density support and its topological properties

Catherine Aaron

Using the $k-$nearest neighbor restricted Delaunay complex to estimate the density support and its topological properties

Catherine Aaron

February 21, 2012

Abstract

We consider random samples in $\mathbb{R}^d$ drawn from an unknown density. This paper is devoted to study the properties of the Delaunay complex restricted to nearest neighbors to estimate the density support and its geometrical properties. We exhibit suitable value for the number of neighbors to be chosen. This value depends on the dimension d, the unknown support S, the unknown density f and the size of the sample. Since f is continuous and the boundary of S is smooth the value only depends on d and the size of the sample. The convergence of the underlying estimator for the support is proved and a lower bound for the convergence speed is given. A procedure to estimate the topological properties of the support is presented.

key words : Delaunay complex, Support estimation, Topological data analysis.

1 Introduction, notations and hypotheses

1.1 Introduction

Let $X_1, ...X_n$ be a random sample in $\mathbb{R}^d$ drawn from f (an unknown density). The density support S is defined by $S = \overline{\{x \in \mathbb{R}^d, f(x) > 0\}}$ (where $\overline{A}$ denotes the closure of the set A). Estimation of the support of the density has various applications in cluster analysis, marketing and econometrics or medical diagnostics (see the discussion in Baillo, Cuevas and Justel [1]). When the support is assumed to be convex, its natural estimator is the convex hull of the sample (see [8] for instance). When the support is no longer assumed

to be convex the most natural estimator (and the most studied one) as been introduced by Chevalier[5] and is:

$$\hat{S}_{n,r_n} = \cup_i \mathcal{B}(X_i, r_n)$$

The properties of this estimator have been studied extensively to reach an exact convergence rate and a central limit theorem ([9] and [10]).

Other estimators have been proposed (as in [6], [15] for instance) but it will be seen, in next paragraph, that topology estimation methods are mainly linked to Chevalier's estimator [5]. In the same time, the main weakness of Chevalier's estimator is that the topological properties of the estimated support are not guaranteed to be the same as those of the support (in particular such an estimator can create unexpected holes).

Estimating the topological properties of the density support is a new and challenging domain that has many applications in fields such as times series, data analysis, image processing and computer vision (see [4] for a review of the applications of topological properties estimation and [14] for a concrete application). By estimating of the topological properties we mean : finding a simplicial complex which is homeomorph to S and can then be used to compute homology groups or homotopy groups.

Computation of persistent homology ([3], [16] and [11]) is a way to estimate the topological properties of S via a set of simplicial complexes. The $\varepsilon-$Rips Complex is the simplicial complex consisting of all simplices that "link" points in a small ball of radius ε. As the support is unknown, the suitable values for ε that assure existence of a a homeomorphism between the simplicial complex and S are unknown. The persistent homology method computes all the $\varepsilon-$Rips complexes and uses a "barcode" graphical indicator to choose an acceptable solution.

Another idea to estimate topological properties is given by the $\alpha-$shape method ([12]). Let us cite the Edlesbrunner's work [7] which proves that: $D_S(X) \approx S$ when a reasonable condition is respected (the closed ball property). To define $D_S(X)$ let us first denote by $V(X_i)$ the Voronoi cell of X_i and by $V_S(X_i) = V(X_i) \cap S$ the restricted Voronoi cell. A simplex $\sigma = (X_{i_1}, ...X_{i_{d'}})$ is a simplex of $D_S(X)$ if and only if $\cap_j V_S(X_{i_j}) \neq \emptyset$. Obviously here S is unknown and the $\alpha-$shape method proposes to use a Chevalier [5] type estimator: $D_{\hat{S}_{n,\alpha}}(X)$ of S instead of S.

Each time, the choice of suitable values for ε (persistent homology) and α ($\alpha-$shapes) is not easy. Moreover choosing of local values ($\varepsilon(x)$ or $\alpha(x)$) may improve the methods when the density is not uniform.

The idea is here to use the well known dual method for a fixed (but may be local) radius: a number of nearest neighbors. It will be proposed to estimate the support and it's topological properties using $D_{k_n}(X)$, the Delaunay complex restricted to the k_n-nearest neighbors.

Definition 1. *$\sigma = (X_{i_1}, ...X_{i_{d'}})$ is a simplex of $D_{k_n}(X)$ if and only if:*

1) σ is a simplex of $D(X)$, the Delaunay complex of X

2) For all (j,k) $[X_{i_j}, X_{i_k}]$ is an edge of the k_n-nearest neighbor graph

Under a set of hypotheses on S that are detailed in the following section, section 2 is dedicated to giving values for k_n. Intuitively k_n should be small enough to help detecting local phenomena : namely k_n/n is expected to converges to 0. But a too small value for k_n may suppress too many simplices to recognize the global underlying shape. According to [7] it is reasonable to think that k_n values such that $D_S(X) \subset D_{k_n}(X)$ are not too small. Section 2 is then devoted to find the smallest number k_n in order to ensure that $D_S(X) \subset D_{k_n}(X)$. We give values for k_n for which this property is almost surely true or asymptotically true

A remarkable result is that k_n can be totally determined from the dimension d and the size of the sample since the density (restricted to S) is uniformly continuous and if the boundary of the support is smooth. Under weaker assumptions, k_n can depends on the extremal values of the density and on the more acute solid angle of S.

In section 3 we prove that $D_{k_n}(X)$ is a convergent estimator of S. By abuse of notation, we write $D_{k_n}(X) = \{x \in \mathbb{R}^d, \exists \sigma \in D_{k_n}(X), x \in \sigma\}$ here. The convergence speed is proved to be at least $(n/\log(n))^{1/d}$ and that is better than the convergence speed for the Chevalier's estimator.

Finally, section 4 presents an algorithm to find a complex homeomorph to S. Unfortunately k_n cannot be chosen in such a way that $D_{k_n}(X) = D_S(X)$ which would have allow us to conclude that $D_{k_n}(X) \approx S$ (which is the empirically observed behavior for $D_{k_n}(X)$). However, a corollary to the sections 2 and 3 is that, if $D_{k_n}(X)$ is not homeomorph to S the unexpected phenomena are localized near the boundary. Such a corollary allows us to propose a procedure to find a sub-complex of $D_{k_n}(X)$ homeomorph to S.

1.2 Hypotheses and notations

Throughout the paper $X = \{X_1, ...X_n\}$, a sample of n independent and identically distributed random variables observed in $\mathbb{R}^d$ $(X_i \in \mathbb{R}^d)$. If f is the associated density, S its support, is defined by $S = \overline{\{x \in \mathbb{R}^d, f(x) > 0\}}$.

Throughout the paper $V(A)$ will denote the volume of the set A and $\mathcal{B}(x,r)$ the open ball centered on x and of radius r.

Throughout the paper S will be supposed to have the following properties:

- **P1:** $\{x \in \mathbb{R}^d, f(x) > 0\}$ is a d-dimensional compact manifold (and so the closure can be removed $S = \{x \in \mathbb{R}^d, f(x) > 0\}$).
- **P2:** S has the α property, i.e. there exists ν such that for all $\varepsilon < \nu$ and for all $x \in S$ $V(\mathcal{B}(x,\varepsilon) \cap S) > \alpha\theta_d\varepsilon^d$ (with θ_d the volume of the unit $d-$ball).
- **P3:** There exists $\mu(\delta(S))$ a constant such that:
  $V(\cup_{x\in S}\mathcal{B}(x,\varepsilon)) - V(S) \sim \varepsilon\mu(\delta(S))$ and $V(S) - V(S_\varepsilon^-) \sim \varepsilon\mu(\delta(S))$
  with $S_\varepsilon^- = \{x \in S \text{ such that } \mathcal{B}(x,\varepsilon) \subset S\}$.
- **P4:** There exists ν such that for all $\varepsilon < \nu$, $S_\varepsilon^- \approx S$.
- **P5:** f must satisfy the following property: $f_0 = \inf_S(f) > 0$ and $f_1 = \sup_S(f) < \infty$.

Let us remark that the most restrictive hypothesis is the first one. Firstly the dimensional hypothesis restricts the field of applications, and it should be said that a challenging future problem should be to suppress this hypothesis. Secondly, the compactness of $\{x \in \mathbb{R}^d, f(x) > 0\}$ is also a strong hypothesis, and it would be interesting to enlarge our results by suppressing that hypothesis (and allow f to be null on the boundary of the support).

Let us briefly remark that, if $\delta(S)$ is a smooth $d-1$ manifold then **P3** (via Minkowski-Steiner formula) and **P4** (via existence of a collar neighborhood [13]) are automatically satisfied. Under this hypothesis **P2** is also satisfy for all $\alpha < 0.5$. With this remark it can be easily proved the same results as those proved in following sections replacing α by 0.5 when smoothness of the boundary is assumed (using a slice modification in the definitions of $r_n(\varepsilon), \rho_n(\varepsilon), r_n^\star(x,\varepsilon), \rho_n^\star(x,\varepsilon)$ in section 2).

It will not be assumed that f (restricted to S) is continuous (in fact uniform continuity of f will be an additional hypothesis) but it is required that f be bounded on S (with a lower bound strictly greater than to 0).

Moreover we define :

- For A a set, $\delta(A)$ is the boundary of the set A.
- The complex restricted to a set A, denoted $D\mid_A$, is composed of all the simplices that are in D and have a vertex in A.

2 Choosing the number of neighbors

The aim of this section is to determine suitable values for k_n the number of neighbors.

Since the density support S is unknown, it is impossible to compute the Edelsbrunner restricted Delaunay complex $D_S(X)$.

Theorem 1 gives a lower bound k_n^* such that $D_S(X) \subset D_{k_n^*}(X)$ almost surely.

This ensures that, if $k_n \geq k_n^*$, not too many simplices are removed when restricting to the k_n-nearest neighbors. This bound is proved to be independent of f in the case when f is uniformly continuous on S.

In general $D_{k_n^*}(X) \nsubseteq D_S(X)$ but a weaker result of this type is established : we prove that the simplices which are in $D_{k_n^*}(X)$ and not in $D_S(X)$ are almost surely located near the boundary of S.

Theorem 1. *Let us define*

$$k_n^* = \frac{f_1}{f_0}\frac{2}{\alpha}(2^d + \alpha 2^{d/2})\ln(n)$$

$$k_n^\star = \frac{2}{\alpha}(2^d + \alpha 2^{d/2})\ln(n)$$

$$r_n = \left(\frac{2\ln(n)}{f_0\theta_d\alpha n}\right)^{1/d}$$

$$\rho_n = \left(\frac{\ln(n)}{f_0\theta_d\alpha n}\right)^{1/d}$$

(i) If $k_n \geq k_n^*$ then $D_S(X) \subset D_{k_n}(X)$ almost surely.

(ii) If $k_n \geq ck_n^*/2$ with any constant $c > 1$ then

$$P(D_S(X) \subset D_{k_n}(X)) \to 1$$

.

(iii) For all k_n, $D_{k_n}(X)\mid_{S_{r_n}^-} \subset D_S(X)\mid_{S_{r_n}^-}$ almost surely.

(iv) For all k_n, and for all constant $c > 1$:

$$P(D_{k_n}(X)\mid_{S_{c\rho n}^-} \subset D_S(X)\mid_{S_{c\rho n}^-}) \to 1$$

.

Adding the hypothesis that f (restricted to S) is uniformly continuous gives:

(i') If $k_n \geq k_n^\star$ then $D_S(X) \subset D_{k_n}(X)$ almost surely.

(ii') For all $c > 1$, if $k_n \geq ck_n^\star/2$ then

$$P(D_S(X) \subset D_{k_n}(X)) \to 1$$

Before proving this theorem, let us remark that, as mentioned in section 1.2., if the boundary of S is a smooth $(d-1)-$manifold α can be replaced by all values strictly inferior to 0.5. In this case, if f is continuous, the lower bound $k_n^\star$ only depends on d and n.

The sequel of this section is devoted to the proof of Theorem 1.

We first define the following quantities:

$$\begin{array}{rcl} r_n(\varepsilon) = \left(\frac{(2+\varepsilon)\ln(n)}{f_0\theta_d\alpha n}\right)^{1/d} & , & r_n^\star(x,\varepsilon) = \left(\frac{(2+\varepsilon)\ln(n)}{f(x)\theta_d\alpha n}\right)^{1/d} \\ \rho_n(\varepsilon) = \left(\frac{(1+\varepsilon)\ln(n)}{f_0\theta_d\alpha n}\right)^{1/d} & , & \rho_n^\star(x,\varepsilon) = \left(\frac{(1+\varepsilon)\ln(n)}{f(x)\theta_d\alpha n}\right)^{1/d} \\ k_n(\varepsilon,\lambda) = \frac{f_1}{f_0}\frac{2+\varepsilon}{\alpha}(\lambda^d + \alpha\lambda^{d/2})\ln(n) & , & k_n^\star(\varepsilon,\lambda) = \frac{2+\varepsilon}{\alpha}(\lambda^d + \alpha\lambda^{d/2})\ln(n). \end{array}$$

Lemma 1. *Let us pick $x \in S$ deterministically. Let $N_x(X,\lambda)$ be the number of observations that are in $\mathcal{B}(x, \lambda r_n(\varepsilon))$ and $M_x(X,\lambda)$ be the number of observations that are in $\mathcal{B}(x,\lambda\rho_n(\varepsilon))$. Then there exists $n_0(\varepsilon)$ and $n_1(\varepsilon,\lambda)$ such that:*

- *For all $n > n_0$, $P(N_x(X,1) = 0) \leq q_n(\varepsilon) \sim n^{-2-\varepsilon}$*
- *For all $n > n_0$, $P(M_x(X,1) = 0) \leq q'_n(\varepsilon) \sim xn^{-1-\varepsilon}$*
- *For all $n > n_1$, $P(N_x(X,\lambda) \geq k_n(\varepsilon,\lambda)) \sim q_n^* \leq \frac{1}{2\sqrt{\pi\ln(n)}} n^{-2-\varepsilon}$*
- *For all $n > n_1$, $P(M_x(X,\lambda) \geq k_n(\varepsilon,\lambda)/2) \sim q_n^{*\prime} \leq \frac{1}{2\sqrt{\pi\ln(n)}} n^{-1-\varepsilon}$*

Proof. **Proof of the first point:** Let q denote the probability of having one observation in $\mathcal{B}(x, r_n(\varepsilon))$. It is clear that **P2** and **P5** imply that for all n such that $r_n(\varepsilon) < \nu$:

$$q \geq \frac{(2+\varepsilon)\ln(n)}{n}$$

Then

$$P(N_x(X,1)=0) = (1-q)^n \leq \left(1 - \frac{(2+\varepsilon)\ln(n)}{n}\right)^n \sim n^{-2-\varepsilon}$$

The proof of the second point is exactly the same.

Proof of the third point point: Let q denote the probability to have one observation in $\mathcal{B}(x, \lambda r_n(\varepsilon))$.

P2 and **P5** now imply that for all n such that $\lambda r_n(\varepsilon) < \nu$,

$$q_m = (2+\varepsilon)\lambda^d \frac{\ln(n)}{n} \leq q \leq \frac{(2+\varepsilon)f_1}{f_0\alpha}\lambda^d \frac{\ln(n)}{n} = q_M.$$

Using the Gaussian approximation of the binomial law,

$$P(N_x(X,\lambda) \geq k_n(\varepsilon,\lambda)) \sim \Phi\left(\frac{nq - k_n}{\sqrt{nq(1-q)}}\right) = p_1.$$

Now Using the previous equalities,

$$p_1 \leq \Phi\left(\frac{nq_M - k_n(\varepsilon,\lambda)}{\sqrt{nq_m(1-q_M)}}\right) = p_2 \sim \Phi\left(\frac{nq_M - k_n(\varepsilon,\lambda)}{\sqrt{nq_m}}\right)$$

Replacing $k_n(\varepsilon,\lambda)$, q_M and q_m by their values,

$$p_2 \sim \Phi\left(-\frac{f_1}{f_0}\sqrt{(2+\varepsilon)\ln(n)}\right) = p_3.$$

And as $\frac{f_1}{f_0} \geq 1$ and using the well known inequality $\Phi(-x) \leq \frac{1}{\sqrt{2\pi}x}\exp(-x)$,

$$p_3 \leq \Phi\left(-\sqrt{(2+\varepsilon)\ln(n)}\right) \leq \frac{1}{2\sqrt{\pi\ln(n)}} n^{-2-\varepsilon}.$$

The proof of the fourth point is exactly the same.

□

Lemma 2. *(Lemma 1 for the continuous case) Let us pick $x \in S$ deterministically. Let $N_x^\star(X,\lambda)$ be the number of observations that are in $\mathcal{B}(x, \lambda r_n^\star(x,\varepsilon))$ and $M_x^\star(X,\lambda)$ be the number of observations that are in $\mathcal{B}(x, \lambda\rho_n(x,\varepsilon))$.*

If f is uniformly continuous then:

- *there exists $n_0(\varepsilon)$ such that, for all $n > n_0$, $P(N_x(X,1)=0) \leq q_n^\star(\varepsilon) \sim n^{-2-\varepsilon/2+\varepsilon^2/4}$,*

- *there exists* $n_1(\varepsilon)$ *such that, for all* $n > n_1$, $P(M_x(X,1) = 0) \leq q_n^{\star\prime}(\varepsilon) \sim n^{-1-\varepsilon/2+\varepsilon^2/4}$,
- *there exists* $n_2(\varepsilon,\lambda)$ *such that, for all* $n > n_2$, $P(N_x(X,\lambda) \geq k_n(\varepsilon,\lambda)) \sim q_n^{\star\star} \leq \frac{1}{2\sqrt{\pi \ln(n)}} n^{-2-\varepsilon/2+\varepsilon^2/4}$,
- *there exists* $n_3(\varepsilon,\lambda)$ *such that, for all* $n > n_3$, $P(M_x(X,\lambda) \geq k_n(\varepsilon,\lambda)/2) \sim q_n^{\star\star\prime} \leq \frac{1}{2\sqrt{\pi \ln(n)}} n^{-1-\varepsilon/2+\varepsilon^2/4}$.

Proof. The proof is exactly the same than of Lemma 1. We will give details only for the first point. Since f is absolutely continuous, there exists $n_0(\varepsilon)$ such that for all $n > n_0$ and all $z \in \mathcal{B}(x, r_n^\star(x,\varepsilon))$, $f(z) \in [f(x) - f_0\varepsilon/4, f(x) + f_0\varepsilon/4]$, and the proof for the first point is the same as previously using the inequality: $q \geq \frac{f(x)-2\varepsilon f_0}{f(x)} \frac{(2+\varepsilon)\ln(n)}{n} \geq (2 + \varepsilon/2 - \varepsilon^2/4)\frac{\ln n}{n}$. □

Lemma 3. *(Classical geometric lemma) Points* $x_1, ...x_{\nu_n(\varepsilon)}$ *in* S *can be deterministically found such that:*

- $S \subset \cup \mathcal{B}(x_i, \varepsilon r_n(\varepsilon)/2)$
- $\nu_n(\varepsilon) \leq \frac{c(S)}{f_0 \varepsilon^d} \frac{n}{\ln(n)}$

with $c(S)$ *a constant that only depends on* S

This is a classical lemma and the proof is let to the reader.

Corollary 1. *Let us pick deterministically* $x_1, ...x_{\nu_n(\varepsilon)}$ *points in* S *as in the previous lemma:*

- *For all* i, $\mathcal{B}(x_i, r_n(\varepsilon))$ *contains at least one point (except for finitely many* n*)*
- *For all* i, $\mathcal{B}(x_i, \lambda r_n(\varepsilon))$ *contains at most* $k_n(\varepsilon,\lambda)$ *points (except for finitely many* n*)*
- $P(\forall i,\ \mathcal{B}(x_i, \rho_n(\varepsilon))$ *contains at least 1 point*$) \to 1$
- $P(\forall i,\ \mathcal{B}(x_i, \lambda\rho_n(\varepsilon))$ *contains at most* $k_n(\varepsilon,\lambda)/2$ *points* $) \to 1$

Proof. The proof is a direct application of Lemmas 1 and 3 in addition to the Borel-Cantelli lemma. □

Corollary 2. *(continuous case) If* f *is uniformly continuous, let us pick deterministically* $x_1, ...x_{\nu_n(\varepsilon)}$ *points in* S *as in the lemma* 3*:*

- *For all i, $\mathcal{B}(x_i, r_n^\star(x_i, \varepsilon))$ contains at least one point (except for finitely many n)*
- *For all i, $\mathcal{B}(x_i, \lambda r_n^\star(x_i, \varepsilon))$ contains at least most $k_n^\star(\varepsilon, \lambda)$ points (except for finitely many n)*
- $P(\forall i,\ \mathcal{B}(x_i, \rho_n^\star(\varepsilon))$ *contains at least 1 point* $) \to 1$
- $P(\forall i,\ \mathcal{B}(x_i, \lambda\rho_n^\star(\varepsilon))$ *contains at most* $k_n^\star(\varepsilon, \lambda)/2$ *points*$) \to 1$

Proof. The proof is a direct application of lemmas 2 and 3 in addition to the Borel-Cantelli lemma. □

Lemma 4. *(First point of theorem 1) If $k_n \geq k_n^*$ then $D_S(X) \subset D_{k_n}(X)$ almost surely*

Proof. The proof is divided into two parts. The first part assures that for all edges of $D_S(X)$, the points of the edges are not too far one from each other, or more precisely: for all X_i, X_j with $[X_i, X_j]$ an edge of $D_S(X)$, $d(X_i, X_j) \leq (2+\varepsilon)r_n(\varepsilon)$ (except for finitely many n). The second point uses this upper bound to prove that points are no more than $(1+\varepsilon')k_n^*$ neighbors (except for finitely many n).

Throughout this proof, $x_1, ...x_{\nu_n(\varepsilon)}$ are deterministically picked in S as in Lemma 2.

First step: It will be proved here that:

For all X_i, X_j with $[X_i, X_j]$ an edge of $D_S(X)$, $d(X_i, X_j) \leq (2+\varepsilon)r_n(\varepsilon)$ (except for finitely many n).

Let us suppose the contrary, and pick X_i, X_j with $[X_i, X_j]$ an edge of $D_S(X)$ and $d(X_i, X_j) \geq (2+\varepsilon)r_n(\varepsilon)$.

As $[X_i, X_j]$ is an edge of $D_S(X)$, there exists $C \in S$ such that:

- $r = d(C, X_i) = d(C, X_j) \geq d(X_i, X_j)/2 \geq (1+\varepsilon/2)r_n(\varepsilon)$
- $\mathcal{B}(C, r)$ does not contain any point of X.

As C is in S, there exists an x_{i^*} with $C \in \mathcal{B}(x_{i^*}, \varepsilon r_n(\varepsilon)/2)$.

With our conditions, $\mathcal{B}(x_{i^*}, r_n(\varepsilon)) \subset \mathcal{B}(C, r)$ and so doesn't contain any observations. This is impossible, according to first point of corollary 1 (except for finitely many n).

Second step: It is now assumed that for all X_i, X_j with $[X_i, X_j]$ an edge of $D_S(X)$, $d(X_i, X_j) \leq (2+\varepsilon)r_n(\varepsilon)$ and it will be proved that X_j can not be a more than $k_n(\varepsilon)$ nearest neighbor of X_i with:

$$k_n(\varepsilon) = \frac{1}{\alpha}((2+2\varepsilon)^d + \alpha(2+2\varepsilon)^{d/2})\frac{f_1}{f_0}(2+\varepsilon)\ln(n)$$

Once again let us suppose the contrary and suppose that there exist X_i and X_j with $d(X_i, X_j) \leq (2+\varepsilon)r_n(\varepsilon)$ and with $\mathcal{B}(X_i, d(X_i, X_j))$ containing more than $k_n(\varepsilon)$ points. As $X_i \in S$ there exists x_{i^*} with $X_i \in \mathcal{B}(x_{i^*}, \varepsilon r_n(\varepsilon)/2)$. As $\mathcal{B}(X_i, d(X_i, X_j)) \subset \mathcal{B}(x_{i^*}, (2+2\varepsilon)r_n(\varepsilon))$, $\mathcal{B}(x_{i^*}, (2+2\varepsilon)r_n(\varepsilon))$ contains more than $k_n(\varepsilon)$ points and that is impossible (except for finitely many n) according to second point of corollary 1 (with $\lambda = 2(1+\varepsilon)$).

□

Remark: The proofs for points *(ii)*, *(i')* and *(ii')* of theorem 1 are exactly the same and will not be written here.

Lemma 5. *Third point of Theorem* 1*:*

For all k_n, $D_{k_n}(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}} \subset D_S(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}}$ except for finitely many n, and so for all k_n, $D_{k_n}(X) \mid_{S^-_{r_n}} \subset D_S(X) \mid_{S^-_{r_n}}$ almost surely.

Proof. It is sufficient to prove that $D(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}} \subset D_S(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}}$ except for finitely many n.

As usual let us suppose the contrary, and pick a simplex of $D(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}}$ which is not in $D_S(X) \mid_{S^-_{(1+\varepsilon)r_n(\varepsilon)}}$. Let us pick X_i a vertex of the simplex that is in $S^-_{(1+\varepsilon)r_n(\varepsilon)}$ (this is possible by the definition of the restriction). As the simplex is in $D(X)$ and not in $D_S(X)$, there exists $C \notin S$ such that $\mathcal{B}(C, d(X_i, C))$ does not contain any other points of X. Now $C \notin S$ implies that $d(X_i, C) > (1+\varepsilon)r_n(\varepsilon)$, and there exists $C' \in S$, with also $C' \in [X_i, C]$ and $d(X_i, C') = (1+\varepsilon)r_n(\varepsilon)$. Let us now pick deterministic points as in Lemma 2. There exists x_{i^*} with $C' \in \mathcal{B}(x_{i^*}, \varepsilon r_n(\varepsilon)/2)$ and $\mathcal{B}(x_{i^*}, r_n(\varepsilon)) \subset \mathcal{B}(C', r_n(1+\varepsilon)) \subset \mathcal{B}(C, d(X_i, C))$ and so $\mathcal{B}(x_{i^*}, r_n(\varepsilon))$ does not contain any observations. That is impossible (except for finitely many n).

□

Remark: The proof for *(iii)* is clearly the same.

3 Density support estimation

Throughout this section, for the sake of simplicity, we write

$$D_{k_n}(X) = \{x \in \mathbb{R}^d, \exists \sigma \in D_{k_n}(X), x \in \sigma\}.$$

When the support is convex, the convex hull of the observations (i.e $D_n(X) = D(X)$) is known to be a good estimator of the support S (cite).

The restricted Delaunay complexes $D_{k_n}(X)$, $k_n \in \{1, \dots, n\}$ can be seen as a generalization of the convex-hull which can be suitable when S is no longer assumed to be convex.

In this section we focus on the study of $D_{k_n^*}(X)$ as an estimator of the density support.

As in [9] the volume of the symmetric difference

$$A \Delta B = (A \setminus B) \cup (B \setminus A)$$

will be used to measure the distance between two sets A and B.

The main result of this section is the following:

Theorem 2. *There exists a constant $A(S, f)$ that only depends on S and the extreme values of f such that:*

$$\left(\frac{n}{\ln(n)}\right)^{1/d} V(D_{k_n^*} \Delta S) < A(S, f) \textit{almost surely.}$$

If f is assumed to be uniformly continuous,

$$\left(\frac{n}{\ln(n)}\right)^{1/d} V(D_{k_n^\star} \Delta S) < A(S, f) \textit{almost surely.}$$

This result does not give general properties for $D_{k_n}(X)$ as an estimator. No optimal value for k_n is exhibited. However, in view of Lemma 9 and remarks that follow, one can empirically see that k_n^* is close to be optimal.

We now prove Theorem 2.

Let us denote $k_0 = (2f_1/(\alpha f_0))(2^d + \alpha 2^{d/2})$. Let us remark that $k_0 \geq 2^{d+1}$. We will also define $M_n = (k_n^*/(f_0 \alpha n \theta_d))^{1/d}$.

Lemma 6. *Let us denote $M_n(\varepsilon, k_n^*) = ((1+\varepsilon)k_n^*/(f_0 \alpha n))^{1/d}$. Let us choose deterministically $x_1, ...x_{\nu_n(\varepsilon)}$ such that $S \subset \cup \mathcal{B}(x_i, \varepsilon M_n)$, all x_i are in S and $\nu_n(\varepsilon) < c(S) n \varepsilon^d / \ln(n)$. There are at least k_n^* points in each ball $\mathcal{B}(x_i, M_n(\varepsilon, k_n^*))$ (except for finitely many n).*

Proof. Let N_i denote the number of points of X that are in $\mathcal{B}(x_i, M_n(\varepsilon, k_n^*))$. For a point X_j the probability to fall in $\mathcal{B}(x_i, M_n(\varepsilon, k_n^*))$ is q.

As in the proof of Lemma 1, $q \leq (1+\varepsilon)k_n^*/n$ (using **P1**, **P2** and **P4**) and using the Gaussian approximation of the binomial law

$$P(N_i < k_n^*) \sim \Phi\left(\frac{k_n^* - nq}{\sqrt{nq(1-q)}}\right) = p_1$$

$$p_1 \leq \Phi\left(\frac{k_n^* - nq^*}{\sqrt{nq^*(1-q^*)}}\right) = p_2$$

$$p_2 = \Phi\left(-\frac{\varepsilon\sqrt{k_n^*}}{\sqrt{1+\varepsilon}}\right) \leq \frac{\sqrt{1+\varepsilon}}{\varepsilon\sqrt{2\pi k_n^*}}\exp(-k_n^*) = \frac{\sqrt{1+\varepsilon}}{\varepsilon\sqrt{2\pi k_n^*}}n^{-k_0}$$

The probability P_n that there exists a ball that contains less than k_n^* points therefore satisfies $\sum P_n < \infty$. This conclude the demonstration.

□

Corollary 3. *Let $X_{(k_n^*(i))}$ denote the k_n^{*the} neighbor of X_i.*

$$max_i(d(X_i, X_{k_n^*(i)}) \leq (k_n^*/(f_0\alpha n))^{1/d}$$

almost surely.

Proof. Let us first choose deterministically $x_1, ...x_{\nu_n(\varepsilon)}$ as in previous lemma. Let us suppose that there exists X_i such that $d(X_i, X_{k_n^*(i)}) \geq (1+2\varepsilon)M_n$. There exists a x_i such that X_i is in $\mathcal{B}(x, \varepsilon M_n))$. As $B_1 = \mathcal{B}(x_i, M_n)[(1+2\varepsilon)^{1/d} - \varepsilon^{1/d}]) \subset \mathcal{B}(X_i, d(X_i, X_j))$, B_1 contains less than k_n^* points. This is impossible (except for finitely many n). Finally the inequality $(1+2\varepsilon)^{1/d} - \varepsilon^{1/d} > (1+\varepsilon^{1/d})$ implies that B_1 contains more than k_n^* points. This is impossible (except for finitely many n).

□

Corollary 4. *$D_{k_n^*}(X) \subset \cup_{x\in S}\mathcal{B}(x, M_n)$ almost surely.*

Proof. If x is in $D_{k_n^*}(X)$, it is in a simplex $\sigma = (X_{1(x)}, .., X_{d'(x)})$ $(d' \leq d+1)$ and $d(x, X_{1(x)}) < M_n$ almost surely using the previous corollary, so $x \in \mathcal{B}(X_{1(x)}, M_n)$. □

Lemma 7. *$S^-_{2r_n} \subset D(X)$ almost surely.*

Proof. Let us suppose that there exist $x \in S^-_{2(1+\varepsilon)r_n}$ and $x \notin D(X)$. As $D(X)$ is the convex hull of X, this implies that there exists a hyperplane H which contains x, and $\overrightarrow{u}$ such that:

- $||\overrightarrow{u}|| = 1$
- For all $X_i \in X$, $\overrightarrow{u}.\overrightarrow{xX_i} \geq 0$.

The half ball $\mathcal{B}^-(x, 2(1+\varepsilon)r_n, u) = \{y \in \mathcal{B}(x, 2(1+\varepsilon)), \overrightarrow{u}.\overrightarrow{xy} < 0\}$ does not contain any observations. Let us now define x_0 such that $\overrightarrow{xx_0} = -(1+\varepsilon)r_n\overrightarrow{u}$. Then $x_0 \in S$ because $x \in S^-_{2(1+\varepsilon)r_n}$ and $\mathcal{B}(x_0, (1+\varepsilon)r_n \subset \mathcal{B}^-(x, 2(1+\varepsilon)r_n, u)$ and so does not contain any observations. This is impossible except for finitely many n. □

Corollary 5. *$S^-_{2r_n} \subset D_{k_n^*}(X)$ almost surely.*

Proof. This is a direct application of Lemmas 8, 4 and 6: since x is in $D(X)$ and in $S^-_{2r_n}$, it is in $D_S(X)$ as in Lemma 6 and so it is in $D_{k_n^*}(X)$ via Lemma 4. □

Lemma 8. *$P(S^-_{2\rho_n} \subset D_{k_n^*}(X)) \to 1$.*

Moreover if f is assumed to be uniformly continuous then $S^-_{2r_n} \subset D_{k_n^\star}(X)$ almost surely and $P(S^-_{2\rho_n} \subset D_{k_n^\star}(X)) \to 1$

The proof is similar to the one of Corollary 5.

Lemma 9. *Both parts of the symmetric difference of S and $D_{k_n^*}(X)$ can be almost surely bounded as follows:*

$$\left(\frac{n}{\ln(n)}\right)^{1/d} V(D_{k_n^*}(X)\backslash S) \leq \left(\frac{n}{\ln(n)}\right)^{1/d} V(\cup_{x\in S}\mathcal{B}(x, M_n)\backslash S) \sim (f_0\alpha\theta_d)^{1/d}k_0\mu(\delta(S))$$

$$\left(\frac{n}{\ln(n)}\right)^{1/d} V(S \setminus D_{k_n^*}(X)) \leq \left(\frac{n}{\ln(n)}\right)^{1/d} V(S \setminus S^-_{2r_n}) \sim (f_0\alpha\theta_d)^{1/d}\mu(\delta(S))$$

Proof. lemma 9 is a direct consequence of Corollaries 4 and 5 and Hypothesis **P3**. □

Remark 1 : Increasing k_n increases $V(D_{k_n} \setminus S)$. Moreover, decreasing k_n increases $V(S\backslash D_{k_n})$. Lemma 9 states that the convergence rates of $V(D_{k_n^*}\backslash S)$ and $V(S \setminus D_{k_n^*})$ are of the same order. This suggest that an optimal choice for k_n should be of order $\log(n)$ as k_n^*.

Remark 2: It can be easily seen that such a convergence speed is better than the convergence speed for Chevalier's support estimator. Biau, Cadre

and Pelletier proved in [9] that the convergence speed is $\sqrt{nr_n^d}$ when $nr_n^d \to \infty$ and $nr_n^{d+2} \to 0$. Obtaining a speed an higher than $\left(\frac{n}{\ln(n)}\right)^{1/d}$ gives r_n that can not satisfy $nr_n^{d+2} \to 0$.

Remark 3 : Theorem 2 is a direct consequence of Lemma 9.

4 Using $D_{k_n^0}(X)$ to estimate topological properties of S

Here k_n^0 can denote or k_n^*, $ck_n^*/2$, $k_n^\star$, $ck_n^{\star/2}$ $(c > 1)$ according to the expected hypothesis on f and the chosen convergence mode. Also r_n^0 will be the associated radius (r_n when expected an almost surely convergence and ρ_n if the probability convergence sufficient). The expression "almost surely or in probability" should end most of the following equalities, inclusion of sentence. It has been removed for ease of reading.

It would have been preferable to obtain (in section 2) the equality $D_S(X) = D_{k_n^0}(X)$ to prove that $D_{k_n^0} \approx S$ (since X has the closed ball property [**?**]). Unfortunately such an equality is false because there exist small simplices (in the k_n^0 nearest neighbor graphs) that are not simplices of $D_S(X)$ (the center of the circumscribed sphere is not in S).
Here, all that can be assured is that

$$S^-_{2r_n^0} \subset D_{k_n^0}(X)$$

Property **P4** assures this, since n is large enough that $D_{k_n^0}(X)$ contains a set homeomorph to S. That localizes the presence of unexpected non contractible cycles (of any dimension) of $D_{k_n^0}(X)$ in $S \setminus S^-_{2r_n^0}$.

Now, $S \setminus S^-_{2r_n^0}$ has a decreasing measure that converges towards 0 and even if this doesn't assure that there is a homeomorphism between $D_{k_n^0}(X)$ and S, it proves that the unexpected phenomena are concentrated in a "small" set.

In practice, topological differences between $D_S(X)$ and $D_{k_n^0}(X)$ have never been observed when computing examples. And what has been observed is that the simplices that are in $D_{k_n^0}(X)$ and not in $D_S(X)$ are localized on the boundary (or, if a simplex is not on the boundary it is surrounded by another simplex of $D_{k_n^0}(X)$ which is not a simplex of $D_S(X)$) (see figure 1).

Unfortunately the fact that only such configurations can happen has not been proved. (Moreover it is possible to construct example of the opposite:

it can be quite easily proved that, when boundary is smooth and $d = 2$ the probability of having $D_{k_n^0}(X)$ not homeomorphic to $D_S(X)$ vanishes, but extension to higher dimension is a challenging further problem).

Figure 1 presents a simulated restricted Delaunay complex on a toy example: points uniformly drawn on a holed disk (a Cd Rom) for an increasing number of observations (first 100 then 200, 500 and finally 1000). Here it has been chosen to use $k_n^\star/2-$nearest neighbors (i.e. only to assure the convergence in probability, but assuming that the density is uniformly continuous and that $\delta(S)$ is smooth so $\alpha = 0.5$).

The plain large red complex is $D_S(X)$ and the plain thin and blue one is $D_{k_n^\star/2}(X)$. The boundary of $S^-_{2\rho_n^\star}$ is represented by large dashed circles.The boundary of $S^-_{\rho_n^\star}$ is represented by large circles.. Finally the boundary of S is presented by thin dashed circles.

The convergence of $D_{k_n^\star/2}(X)$ towards S can be clearly observed.

In the first graph (100 observations) $D_{k_n^\star/2}(X)$ does not allow one to recognize the hole in the disk: the hole is filled by simplices.

From 200 observations the $D_{k_n^\star/2}(X)$ allows one to recognize topologically a holed disk.

For $n = 200$ and $n = 500$ one can observe simplices of $D_{k_n^\star/2}(X) \mid_{S^-_{\rho_n^\star}}$ that are not in $D_S(X) \mid_{S^-_{\rho_n^\star}}$ (a case of this is circled by a gray ellipse for the 500 points example). Such situations vanish for $n = 1000$.

Even if it has never been observed that, for n is large enough, there is topological difference between S and $D_{k_n^0}(X)$ an algorithm must be proposed to correct potential problems (unexpected non contractible cycles of any dimension in $S \setminus S^-_{2r_n}$).

We now propose to compute the set of peeled complexes (a peeling operation consists of removing all the simplices that intersect the boundary).

Definition 2. *The peeling function $P : \mathcal{C} \to \mathcal{C}$ (with $\mathcal{C}$ the set of all the simplicial complexes) can be defined as follows:*

- $P(c) \subset c$
- $\sigma \in c$ *and* $\sigma \notin P(c) \Leftrightarrow \sigma \cap \delta(c) \neq \emptyset$

The effect of peeling can be seen in figure 2.

It is thus proposed to compute topological invariants as homology groups for all the $P^{(i)}(D_{k_n^0})$ and proceed as in the persistent homology method.

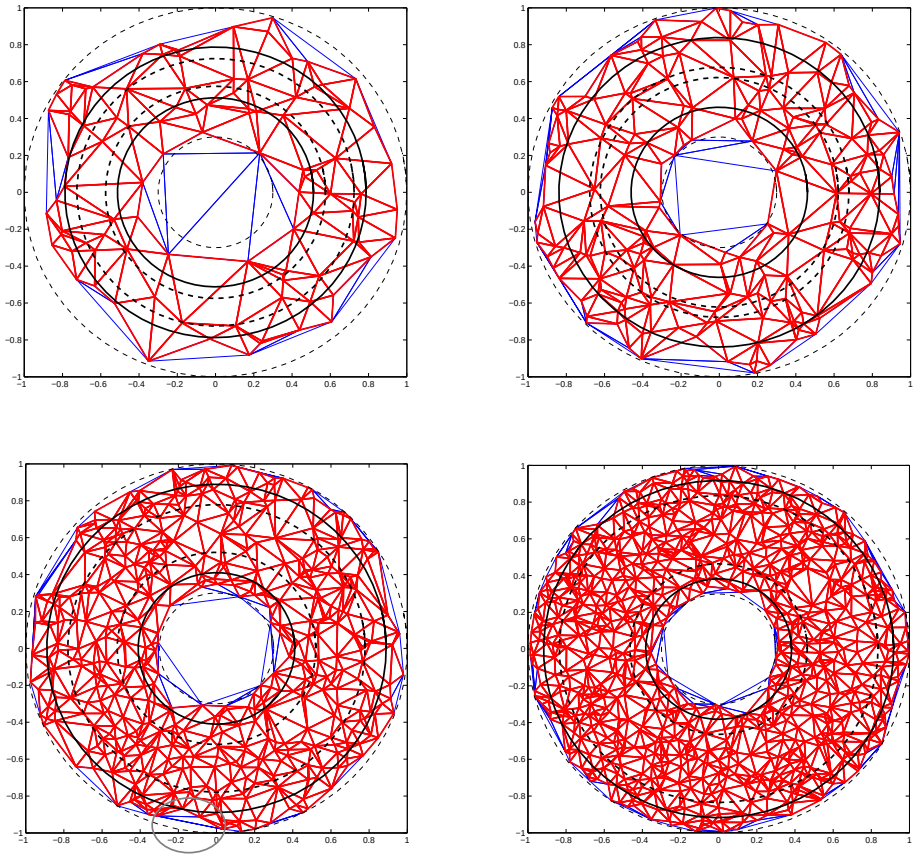

Figure 1: Examples of restricted Delaunay complexes

Such a procedure is expected to successively erase the unexpected cycles near the boundary (if they exist). It can also provide a lot of precious informations such as:

- $P^{(1)}(D_{k_n^0}) = \emptyset$ indicates that the intrinsic dimension is lower than d
- the last non empty $P^{(i)}(D_{k_n^0})$ can give a kind of reduced skeleton of the data support.

5 Conclusion and further works

The idea of estimating the support of the density and its topological properties via the Delaunay complex restricted to the k_n-nearest neighbors gives

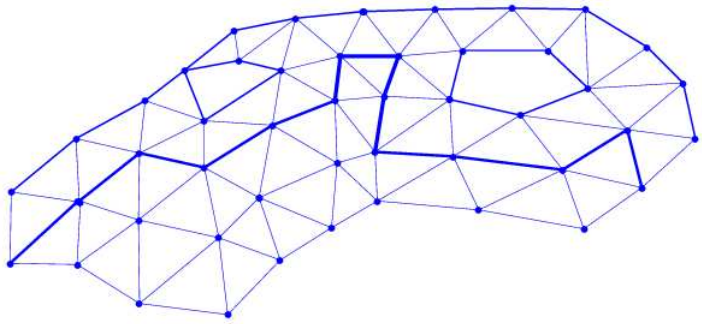

Figure 2: Effect of the peeling function on a complex. Not all the complex is plotted: little bold segments are those of $\delta(c)$ and hugely bold segment are those of $\delta(P(c))$. It can be seen that the peeling operation erases the two non retractile cycles located near the boundary.

very good results. First let us notice the fact that k_n is explicitly known since the density is uniformly continuous and the boundary of the support is a smooth manifold (which are not too strong hypotheses). Practical results are better than what is proved in this paper: in practice $D_{k_n^0}(X) \approx S$, and the proposed peeling procedure in the last section is not really needed. It would be great to be able to prove this. Some other theoretical extensions would be very interesting:

- The first would be to deal with densities that are null on the boundary of the support (i.e. the closeness of $\{x$ such that:$f(x) > 0\}$ is no longer assumed).

- The second (and maybe the more interesting point) would be: how to deal with a support of lower dimension ? (intrinsic dimension $d' < d$) The field of application field would be greater. It could be applied to dimension reduction problem linked with sparsity. An algorithm can already be easily proposed: to build Delaunay complex restricted to d' (via local PCA [2]) and to $k_n(d')$ neighbors for instance). Study of the theoretical properties of such an estimator is the main problem that has to be solve. It can yet be empirically said that the restricted

delaunay complex with a number of neigbors choosen as in this paper but replacing d by d' gives quite good results. Obviously there is no homeomorphism between S and its estimate but convergence of the estimate seems to appear. It can be observed in figures 3 and 4 (a cercle and a cylinder). First proof of this empiricall result has to be done and second an algorithm that does not start by the computation of the initial delaunay complex has to be found xhen d is large (practically when $d > 5$) to have reasonable computing time. An even more challenging problem is to deal with uncknown intrinsic dimension of the support.

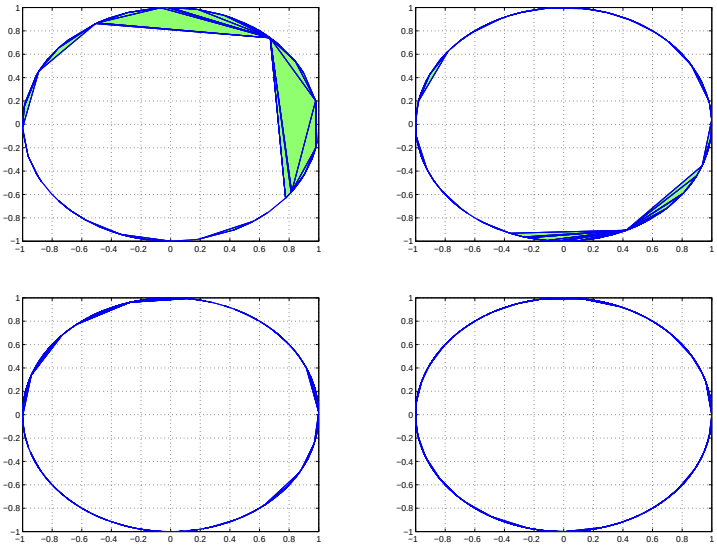

Figure 3: Examples of restricted Delaunay complexes with points drawn on a circle : $d = 2$, $d' = 1$ (100, 200, 500 and 1000 points)

References

[1] A. Cuevas A. Baillo and A. Justel. Set estimation and nonparametric detection. *The Canadian Journal of Statistics*, 28:765–782, 2000.

[2] J.D. Boissonnat and A. Ghosh. Manifold reconstruction using tangential delaunay complexes. In *In Proc. 26th Annual Symposium on Computational Geometry,*, 2010.

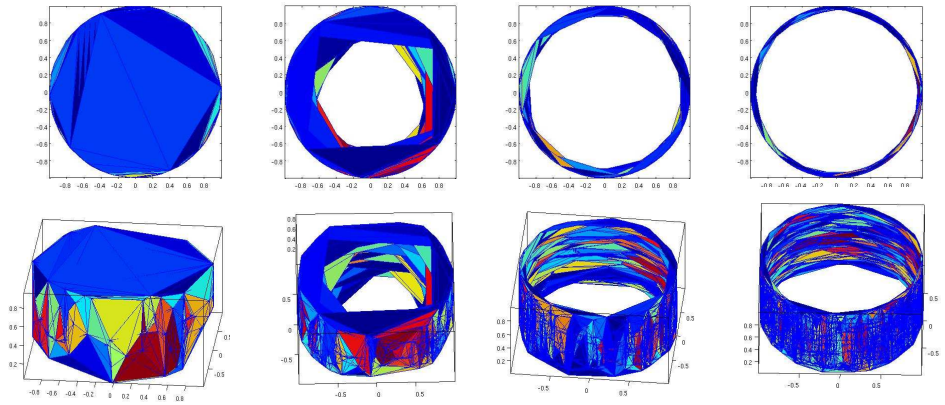

Figure 4: Examples of restricted Delaunay complexes with points drawn on a cylinder: $d = 3$, $d' = 2$ (100, 200, 500 and 1000 points). Each times two different views are presented

[3] G. Carlsson. Persistent homology and the analysis of high dimensional data. In *Symposium on the Geometry of Very Large Data Sets , Fields Institute for Research in Mathematical Sciences*, 2005.

[4] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46:255–308, 2009.

[5] J. Chevalier. Estimation du support et du contoir du support d'une loi de probabilit. *Ann. Inst. H. Poincar sec. B*, 12:339–364, 1976.

[6] A. Cuevas and R. Fraiman. A plug-in approach to support estimation. *The Annals of Statistics*, 25:2300–2312, 1997.

[7] H. Edelsbrunner and N. R. Shah. Triangulating topological spaces. *Internat. J. Comput. Geom. Appl.*, 7:365–378, 1997.

[8] B. Efron. The convex hull of a random set of points. *biometrika*, 15:331–343, 1965.

[9] B. Cadre G. Biau and B. Pelletier. Exact rates in density support estimation. *Journal of Multivariate Analysis*, 99:2185–2207, 2008.

[10] D.M. Mason G. Biau, B. Cadre and B. Pelletier. Asymptotic normality in density support estimation. *Electronic Journal of Probability*, 14:2617–2635, 2009.

[11] A. Collins G. Carlsson, A. Zomorodian and L. Guibas. Persistence barcodes for shapes. *International Journal of Shape Modeling*, 11:149–187, 2005.

[12] D. G. Kirkpatrick H. Edelsbrunner and R. Seidel.rect. On the shape of a set of points in the plane. *IEEE Trans. Inform. Theory*, 29:551–559, 1983.

[13] M. W. Hirsch. *Differential Topology*. Springer Verlag, 1976.

[14] P. Fu P. V. Sudharkar J. Liang, H. Edelsbrunner and 18-29a S. Subramaniam. P (1998). Analytic shape computation of macromolecules ii: inaccessible cavities in proteins. *roteins: Structure, Function, and Genetics*, 33:18–29, 1998.

[15] J. Klemel. Complexity penalized support estimation. *Journal of Multivariate Analysis*, 88:274–297, 2004.

[16] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete and Computational Geometry*, 33:247–274, 2005.