

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

Catherine Aaron

▶ **To cite this version:**

Catherine Aaron. Using the $k-$nearest neighbor restricted Delaunay polyhedron to estimate the density support and its topological properties. 2012. hal-00672705v2

HAL Id: hal-00672705
https://hal.science/hal-00672705v2

Preprint submitted on 3 Dec 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Catherine Aaron

December 3, 2012

Abstract

We consider random samples in $\mathbb{R}^d$ drawn from an unknown density. This paper is devoted to the study the properties of the Delaunay polyhedron restricted to nearest neighbors as an estimator of the density support preserving its topological properties. When the dimension of the support is d, we exhibit suitable value for the number of neighbors to be used. This value ensure that, when $d = 2$, our estimator is a.a.s. homeomorph to the support. Empirically, our estimator also preserves the topology for higher dimensions but it is not proved here. When f is Lipschitz continuous and the boundary of S is smooth the value depends on d and the size of the sample only. The convergence of the underlying estimator to the support is proved and a lower bound for the convergence rate is given. When the dimension of the support is less than d, another estimator is proposed.

Key Words: Delaunay complex, polyhedron, support estimation, topological data analysis, geometric inference.

1 Introduction

Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ be a random sample in $\mathbb{R}^d$ drawn from an unknown density f. 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, econometrics, medical diagnostics, and so on (see the discussion in [1]).

For instance, classification and clustering methods can be deduced from the support estimation. A natural way to cluster points is to look at the

connected components of the density support. For a classification problem an intuitive solution is given by the following algorithm: first one has to estimate the supports of each class and, after that, a new observation can be affected to a class if its belong to its support [1]. Of course, in practice it is not that simple, and it may be more realistic to work with level sets instead of the whole supports. Indeed, in the first case (clustering) one can easily imagine a situation where the support is connected, while the level sets has several connected components (the density is multi-modal) which correspond to natural clusters (see [8]). In the second case (classification), level sets can be separated, even when the supports are overlapping. However, in this work we concentrate on the density support estimation only. We think that generalization of our method to the level sets is possible, and so, this paper can be considered as a first step.

There exists various ways to estimate the density support (see, for example, [19], [17] and [24]). However, usually the statisticians study the asymptotic behavior of a distance between the support and its estimator, and are rarely interested in the recognition of its topological properties (we can, nevertheless, mention an informal discussion in [5]).

In our opinion, the recognition of the topological properties is very interesting and important from the statistical point of view. As an illustration of that we can stress the impact of the topological knowledge on the dimension reduction problem. When the intrinsic dimension of the support is $d' < d$ then:

- if S is convex then Principal Component Analysis will do the job;
- if S is homeomorph to a convex then non linear projection method such as isomap [30] will succeed;
- if S is not homeomorph to a convex then such methods as curvilinear distance analysis [25] may be preferred.

Other potential statistical applications can be found in [10] or [16].

Numerous results concerning the recognition of the topological properties based on a finite number of points can be found in the literature devoted to computational geometry (see, for example, [21], [20] and [6]). However, in this field, the statistical or probabilistic point of view is rarely discussed.

The aim of the present work is to unify this two approaches. More precisely we are interested in the construction of a density support estimator preserving the topological properties (the estimator is homeomorph to the support). A similar problem is considered in [13], [12] and [14], but the proposed estimator is radically different from ours.

1.1 Support estimation methods

When the aim is to estimate the density support, the estimated object is a set which is quite unusual in statistics. To study the convergence of the estimators, one have to choose a distance between sets. Usually, in support estimation problems, the distance given by the measure (typically the Lebesgue one) of the symmetric difference is considered.

Definition 1. *The symmetric difference of two sets A and B is*

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

Definition 2. *Let μ be a measure in $\mathbb{R}^d$, $A \subset \mathbb{R}^d$ and $B \subset \mathbb{R}^d$. The symmetric difference distance (with respect to μ) is given by $d_\mu(A, B) = \mu(A\Delta B)$.*

The most natural support estimator (and the most studied one) has been introduced in [15] and [19] as:

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

where $\overline{\mathcal{B}}(x, r)$ denotes the closed ball of radius r centered in x. Throughout this paper, this estimator will be called Devroye-Wise estimator.

The properties of this estimator have been extensively studied. Namely, an exact convergence rate and a central limit theorem (for the symmetric difference distance) were obtained in [4] and [3].

Other estimators can also be found in statistical literature (see, for example, [17] and [24]). However, as we will see later, the topology preserving methods are mainly based on balls unions, and so linked with the Devroye-Wise estimator.

Our idea is that focusing on the measure can not help to solve the topology preservation problem: one can easily imagine a sequence of sets that converges to a limit (for the symmetric difference distance) having very different topological properties.

This is illustrated by two examples in Figure 1, where

$$S_n = [0,1]^2 \setminus \left(\bigcap_{i=1}^{n} \left[\frac{i}{n+1} - \frac{1}{n^3}, \frac{i}{n+1} + \frac{1}{n^3} \right] \times [0,1] \right)$$

and

$$\mathcal{S}_n = [0,1]^2 \setminus \left(\bigcap_{i=1}^{n} \mathcal{B}(x_i, 1/n^2) \right)$$

with x_i drawn uniformly in the unit square.

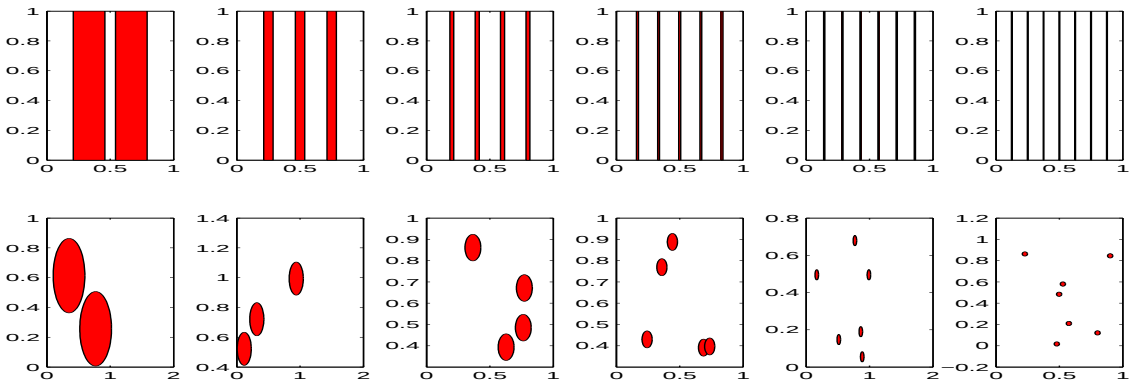


Figure 1: S_n and $\mathcal{S}_n$ for $n \in \{2, 3, 4, 5, 6\}$.

In each of the two examples, the measure (surface) of the symmetric difference with the unit square converges to 0 (with rate $1/n$), but the elements of the sequence are not homeomorph to the unit square.

As we will see later, a situation similar to the second example of Figure 1 can occur when using the Devroye-Wise estimator and so, the latter may not preserve the topology of the support.

The aim of the present paper is to propose an estimator of the density support preserving the topological properties (that is, homeomorph to it).

Topology preserving estimation is a new and challenging domain that has many application fields such as times series, data analysis, image processing and computer vision (see [10] for a review of applications of topological properties estimation and [26] or [16] for concrete applications).

1.2 Topology recognition methods

The problem of recognition of the topological properties of a given set based on a finite set of points is widely considered in computational geometry. By recognition of the topological properties we mean here: finding a set homeomorph to the given one, but having a "simple structure" permitting to "easily" compute some topological invariants such as homology groups or homotopy groups [7]. A natural way to guarantee the "simple structure" is to use polyhedron sets [7].

Definition 3 (Simplex, Sub-simplex)**.** *Let $x_1, \ldots, x_k$ be k affinely independent points (that is, points which are not in a k' hyperplane with $k' \leq k-2$). The associated simplex $\sigma = (x_1, \ldots, x_k)$ is the convex hull of the points.*

A simplex $\sigma' = (x'_1, \ldots, x'_{k'})$ *is a sub-simplex of* σ *if* $\{x'_1, \ldots, x'_{k'}\} \subset \{x_1, \ldots, x_k\}$.

Definition 4 (Simplicial Complex)**.** *A simplicial complex* $\mathcal{K}$ *is a collection of simplexes such that:*

- *if* σ *is a simplex of* $\mathcal{K}$*, then every sub-simplex of* σ *is a simplex of* $\mathcal{K}$*;*
- *if* σ_1 *and* σ_2 *are two simplexes of* $\mathcal{K}$ *and if* $\sigma = \sigma_1 \cap \sigma_2$ *is not empty, then* σ *is a sub-simplex of* σ_1 *and of* σ_2*.*

Definition 5 (Polyhedron)**.** *Let* $\mathcal{K}$ *be a simplicial complex. The set* $K = \bigcup_{\sigma \in \mathcal{K}} \sigma$ *is called the polyhedron of* $\mathcal{K}$*.*

Remark: the simplicial complex $\mathcal{K}$ is an abstract object that allows to compute some topological invariants of the concrete geometric object K. In this paper, we focus on the geometric object K. A set A homeomorph to a polyhedron K is told triangulable.

In this paper a specific complex: the Delaunay Complex is used. There is two equivalent definitions of the Delaunay Complex.

Definition 6 (Delaunay Complex via Voronoi Cells)**.** *Let* $\mathcal{X}_n = \{X_1, .., X_n\}$ *be a set of points in* $\mathbb{R}^d$*. The Voronoi cell of the point* X_i *denoted* $\mathrm{Vor}(X_i)$ *is the set* $\mathrm{Vor}(X_i) = \{x \in \mathbb{R}^d$ *such that, for all* $j \neq i, ||\overrightarrow{xX_i}|| \leq ||\overrightarrow{xX_j}||\}$.

A simplex $\sigma = (X_{i_1}, \ldots, X_{i_k})$ *belongs to the Delaunay Complex* $\mathcal{D}(\mathcal{X}_n)$ *if* $\bigcap_j \mathrm{Vor}(X_{i_j}) \neq \emptyset$.

Definition 7 (Delaunay Complex via circumscribed spheres)**.** *Let* $\mathcal{X}_n = \{X_1, .., X_n\}$ *be a set of points in* $\mathbb{R}^d$*.*

A $d-$*simplex* $\sigma = (X_{i_1}, \ldots, X_{i_{d+1}})$ *belongs to the Delaunay Complex* $\mathcal{D}(\mathcal{X}_n)$ *if* $\mathcal{B}(O, r) \cap \mathcal{X}_n = \emptyset$*, where the open ball* $\mathcal{B}(O, r)$ *has the same center and the same radius as the hypersphere* $\mathcal{S}(O, r)$ *circumscribed to* $\{X_{i_1}, \ldots, X_{i_{d+1}}\}$.

A simplex σ' *belongs to the Delaunay complex* $\mathcal{D}(\mathcal{X}_n)$ *if it is a sub-simplex of a* $d-$*simplex of the Delaunay Complex* $\mathcal{D}(\mathcal{X}_n)$*.*

The polyhedron $D(\mathcal{X}_n)$ associated to the Delaunay complex $\mathcal{D}(\mathcal{X}_n)$ is the convex hull of the set $\mathcal{X}_n$. To recognize the topology of a set using a Delaunay based polyhedron it is so needed to remove some simplexes. Based on that idea, in [21], Edelsbrunner built a restriction of the Delaunay polyhedron that is homeomorph to a set.

Definition 8 (Edelsbrunner's Restriction)**.** *Let* $\mathcal{X}_n = \{X_1, .., X_n\}$ *be a set of points in* $\mathbb{R}^d$*. We denote* $\mathrm{Vor}_{(S)}(X_i) = \mathrm{Vor}(X_i) \cap S$ *the Voronoi cell of* X_i

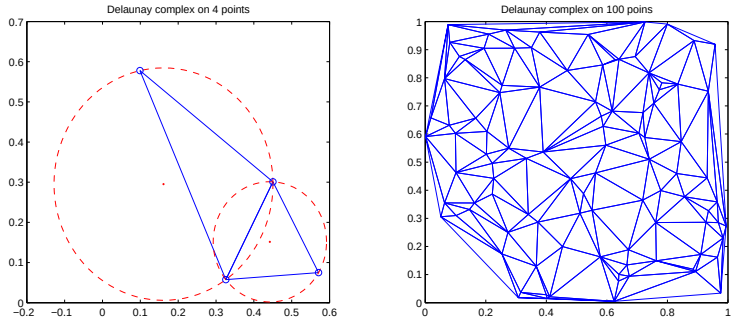


Figure 2: Example of Delaunay complexes built on 4 points (left) and 100 points (right). Points are drawn uniformly in the unit square. The interior of every triangle is a 2−dimensional simplex.

restricted to S. The Delaunay Complex restricted to S is denoted $\mathcal{D}_{(S)}(\mathcal{X}_n)$ and is defined as follows:

A simplex $\sigma = (X_{i_1}, ..., X_{i_k})$ belongs to the Delaunay Complex$\mathcal{D}_{(S)}(\mathcal{X}_n)$ if $\cap_j Vor_S(X_{i_j}) \neq \emptyset$.

The associated polyhedron is denoted $D_{(S)}(\mathcal{X}_n)$.

Definition 9 (Closed Ball Property)**.** *Let S be a compact $d'-$ manifold (with or without boundary ∂S). We say that $\mathcal{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^d$ has the closed ball property on S if for every $\{i_1, .., i_k\}$:*

$\bigcap Vor_{(S)}(X_{i_j})$ is empty or homeomorph to a $d' + 1 - k$ dimensional closed ball

$\bigcap Vor_{(\partial S)}(X_{i_j})$ is empty or homeomorph to a $d' - k$ dimensional closed ball

The fact that such an assumption is reasonable is discussed in [21] when S is a smooth manifold.

Theorem 1. *If $\mathcal{X}_n$ has the closed ball property on S then: $D_{(S)}(\mathcal{X}_n) \cong S$ ($D_{(S)}(\mathcal{X}_n)$ is homeomorph to S).*

Obviously for the statistician S is unknown and the Edelsbrunner's restriction of the Delaunay complex can not be computed.

1.3 Topology preserving methods when the support is unknown

1.3.1 The $\alpha-$shape method

The $\alpha-$shape method, introduced in [20] proposes to use the Edelsbrunner's Delaunay restriction. Here, $\mathcal{X}_n = \{X_1, \ldots, X_n\}$ is a random sample drawn from an unknown density that is supported by S. As $D_{(S)}(\mathcal{X}_n)$ can not be computed because S is unknown it is proposed to restrict the Delaunay polyhedron on an estimator of the support. The chosen way to estimate the support is the Devroye-Wise method. So the proposed polyhedron is: $D_{(\hat{S}_{n,r_n})}(\mathcal{X}_n)$. The name α in the name $\alpha-$shapes comes from the fact that the authors named the radius α (here r_n).

We think that there is two main problems with such a method. First, the choice of local radius may improve the global radius method when density is far from uniform. Second, an interesting phenomena occurs when we work with union of balls estimators: The optimal radius choice (according with the symmetric difference distance) may gives estimator that is not homeomorph to the support. More than that, the associated $\alpha-$shape does not correct the problem. The Figure 3 illustrates that phenomena. For an uniform sample of 1000 points in the unit disk, the optimal radius r for Devroye-Wise estimator has been estimated (via Monte Carlo method as here the support is known). The left graphic is the Devroye-Wise estimator which is not homeomorph to the disk because of the existence of various small holes. The right graphics is the Edelsbrunner's restriction of the Delaunay polyhedron to our support estimator. The small holes are strongly emphasized and the result is not homeomorph to the disk.

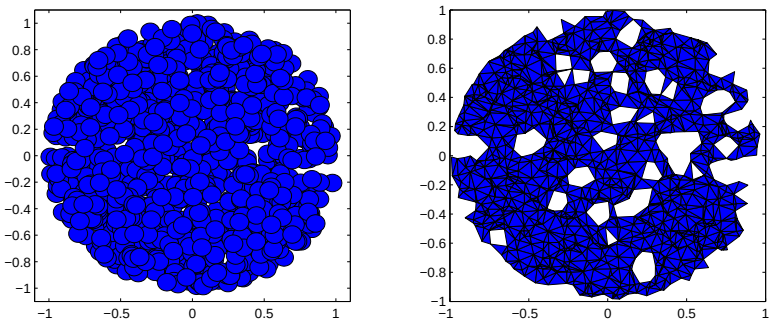


Figure 3: Devroye-Wise estimator for an optimal radius (left) and the associated $\alpha-$shape (right).

As a conclusion the choice of the radius for the topology preserving problem is not optimal for the support estimation problem. So both problem can not be optimally solves together. Practically a good radius choice is fundamental and this choice can not be done only according to the support estimation problem.

Exactly the same kind of problems occurs for the [13], [12] [14] method.

1.3.2 Persistent homology

Another method for the estimation of some topological invariants of the density support is the computation of persistent homology (see [9], [31] and [11]). Instead of the Delaunay complex the complex used is now the $\varepsilon-$Rips Complex.

Definition 10. *Let $\mathcal{X}_n = \{X_1, .., X_n\}$ be a set of points in $\mathbb{R}^d$. The $\varepsilon-$Rips Complex $\varepsilon - \mathcal{R}(\mathcal{X}_n)$ is defined as follows:*

A simplex $\sigma = (X_{i_1}, ..., X_{i_k})$ belongs to the $\varepsilon-$Rips Complex if $||\overrightarrow{X_{i_j} X_{i'_j}}|| \leq \varepsilon$

The practical choice of a suitable value for ε when the support is unknown is hard (and the existence of a value that ensures homeomorphism or, at least homotopy equivalence is not ensured). To skirt this problem, the authors propose to compute the invariants on a large set of ε values and to look for the persistent ones.

This method gives quite good results but the reading of the persistent invariants is not that easy and there is no associated support estimator.

To conclude this section it can be seen that the two main problems with the above exposed method are :

- The choice of suitable values for the radius is not an easy job.
- Moreover choosing of local values for the radius may improve the methods when the density is not uniform.

1.4 The proposed estimator

We based our estimator on the following ideas:

On one hand, if the support is convex then the convex hull of the observation (which is also their Delaunay polyhedron) can be seen as the best density support estimator and it also conserves the topology of the support. Some asymptotic properties of this estimator can be found for instance in [22], [2],

[28] or [29] and the convergence rate is $n^{-2/(d+1)}$. The convexity hypothesis is a strong assumption, and it can appear natural to try to generalize this result to non-convex sets estimating the support restriction of the Delaunay polyhedron.

On the other hand, Edelsbrunner proved that there exists a restriction of the Delaunay Polyhedron that is homeomorph to the support, but unfortunately the knowledge of the support is needed to ensure the homeomorphism. Using a restriction of the Delaunay polyhedron to a ball-based support estimator can give good result for topology recognition but with some problems previously exposed. So we are going to look to another way to restrict the Delaunay polyhedron.

The idea here is to use the well-known dual method for a fixed radius method, that is nearest neighbors method. More precisely, we propose to estimate the support with $D_{k_n}(\mathcal{X}_n)$, the Delaunay polyhedron restricted to the k_n-nearest neighbors.

Definition 11 (Delaunay polyhedron restricted to nearest neighbors)**.** *Let $\mathcal{X}_n = \{X_1, .., X_n\}$ be a set of points in $\mathbb{R}^d$.*

A simplex $\sigma = (X_{i_1}, ...X_{i_{d'}})$ belongs to $\mathcal{D}_{k_n}(\mathcal{X}_n)$ if:

1) σ is a simplex of $\mathcal{D}_{k_n}(\mathcal{X}_n)$, the Delaunay complex of $\mathcal{X}_n$

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

As usual, $D_{k_n}(\mathcal{X}_n)$ is the associated polyhedron.

In the Section 2 we study some properties this polyhedron-estimator. The aim is to find suitable sequences for k_n that allows to preserve the topology of S. According to Edelsbrunner's work we wish to find value for k_n such that $D_{(S)} = D_{k_n}$. Unfortunately such a result can not be obtained but we find explicit values for k_n, such that: $k_n/n \to 0$ and $D_{(S)} \subset D_{k_n}$ a.a.s. (asymptotically almost surely). The second property ($D_{(S)} \subset D_{k_n}$) ensures that not too many simplexes are removed by the restriction. The first point ($k_n/n \to 0$) ensures that local phenomena are recognized (kept simplex are all included in small balls). So both points are a first step towards the existence of a homeomorphism. In the specific case of the dimension 2 it can be shown that $D_{k_n} \cong S$ a.a.s. but for higher dimension such a result could not be obtained.

We also carry out a quick study of the properties of D_{k_n} as an estimator of S. Note first that we chose to work with the Hausdorff distance between sets.

Definition 12 (Hausdorff distance)**.** *Let A an B be two subsets of $\mathbb{R}^d$. We denote $A + \varepsilon\overline{\mathcal{B}}$ the set $\bigcup_{a\in A} \overline{\mathcal{B}}(a, \varepsilon)$*

$$d_H(A, B) = \inf\{r > 0, A \subset B + \varepsilon\overline{\mathcal{B}} \text{ and } B \subset A + \varepsilon\overline{\mathcal{B}}\}.$$

In literature, the Hausdorff distance is the second usual choice after the symmetric difference distance. When the dimension of S is the dimension of the ambient space the two choices are almost equivalent.

We will show that for our value of k_n, the sequence $(\ln n/n)^{1/d} d_H(D_{k_n}, S)$ is e.a.s. bounded. Such a rate of convergence is similar to the one founded in [18] for Devroye-Wise estimator. This is disappointing with regard to the expected rate ($n^{-2/(d+1)}$ according to [2]). Empirical simulations show that our theoretical result underestimates the real one and that a better result can be obtained.

In Section 3, the method is adapted to the most interesting case when the dimension of the support is less than the dimension of the ambient space. However, it will be assumed that the dimension d' of the support is known. A new estimator is proposed. This estimator is also a nearest-neighbor type restriction of the Delaunay complex but it is done according to the dimension of the support. Suitable values for the number of neighbors are exhibited when S is with or without boundary. The properties of the new polyhedron as a support estimator are also studied and, in this case, the choice of the Hausdorff distance is meaningful (while the symmetric difference distance is not anymore).

1.5 Definitions and Notations

Throughout the paper we will use the following notations:

$\mathcal{B}(x, r)$ (resp. $\overline{\mathcal{B}}(x, r)$) denotes the open (resp. closed) ball of radius r, centered in x.

$\mathcal{B}_k$ (resp. $\overline{\mathcal{B}}_k$) denotes the open (resp. closed) $k-$dimensional unit ball (if there is no dimension ambiguity we write simply $\mathcal{B}$ or $\overline{\mathcal{B}}$).

For a set A in $\mathbb{R}^d$, $V(A)$ denotes the volume (according to the Lebesgue measure μ). We denote $\theta_d = V(\mathcal{B}_d)$.

A set A is said to be a $k-$dimensional manifold, if each point x of A has a neighborhood homeomorph to $\mathcal{B}_k$ or to $\overline{\mathcal{B}}_k$. A manifold is told to be a manifold with boundary if there exists at least one x of A that does not admit any neighborhood homeomorph to $\mathcal{B}_k$. If $\mathring{A}$ denotes the interior of A and $\overline{A}$ the closure of A, the boundary ∂A of A is defined by $\partial A = \overline{A} \setminus \mathring{A}$.

A compact set A is said to be $\alpha-$standard, if there exists ε_0 such that: for all $x \in S$ and for all $\varepsilon \leq \varepsilon_0$ we have $V(\mathcal{B}(x, \varepsilon) \cap A) \geq \alpha\varepsilon^d\theta_d)$ (see **??**).

The functions $\varphi(x)$ and $\psi(x)$ sometimes used in the paper are respectively defined by $\varphi(x) = (x+1)\ln(x+1) - x$ and $\psi(x) = x - \ln(1+x)x$.

To simplify the notations it will now be denoted D_{k_n} for the Delaunay polyhedron restricted to the k_n-nearest neighbors (resp. $D_{(S)}$ for the Delaunay polyhedron restricted to the support as in [21]) instead of $D_{k_n}(\mathcal{X}_n)$ (resp. $D_{(S)}(\mathcal{X}_n)$)

2 Case when the dimension of the support is the dimension of the observation space

2.1 Hypotheses

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

Throughout this section S will be supposed to verify the two following hypotheses:

- **H1:** S is a compact $d-$dimensional manifold. It is $d-$dimensional as the ambient space. Remark: a compact $d-$dimensional manifold included in $\mathbb{R}^d$ has a boundary.
- **H2:** S is $\alpha-$standard. This covers a huge range of possible manifolds such as polyhedron or manifold with a $\mathcal{C}^2$ boundary. For a $d-$dimensional manifold with a $\mathcal{C}^2$ boundary, S is $\alpha-$standard for all $\alpha < 0.5$. For a $d-$dimensional manifold with a piecewise $\mathcal{C}^2$ boundary, α can be seen as a function of the most acute solid angle of S.

The density f will be supposed to verify the two following hypotheses:

- **H3:** $\inf\{f(x), x \in S\} = f_0 > 0$. We also denote $\overline{f}_0 = \inf\{f(x), x \in \partial S\}$ (obviously, as $f_0 > 0$ we have $\overline{f}_0 > 0$) and $f_1 = \max\{f(x), x \in S\}$

Two additional hypotheses will sometimes be used

- **H4:** S is a manifold with a $\mathcal{C}^2$ boundary.
- **H5:** $f|_S$ (f restricted to the support) is Lipschitz continuous. That is there exists a real K_f such that for all $x, y \in S^2$, $|f(x) - f(y)| \leq K_f ||\overrightarrow{xy}||$

2.2 Preservation of the topological properties

Let us recall that when the density support S is known, Edelsbrunner showed in [21], that the simplicial complex $D_{(S)}$ is homeomorph to S (under a reasonable hypotheses). So, it is natural to consider the following problem: can we find a sequence k_n that ensures that:

- D_S is (asymptotically) included in D_{k_n} ?
- D_{k_n} is (asymptotically) included in $D_{(S)}$?

The first question will be answered in Section 2.2.1 where the corresponding values of k_n will be found.

Unfortunately, we will see in Section 2.2.3 that it is not possible to find values of k_n that ensure both inclusions. However, a "homeomorphism theorem" will be established in the two-dimensional case ($d = 2$).

2.2.1 The first inclusion

The first inclusion ($D_{(S)} \subset D_{k_n}$) is, for us, the most important because it ensures that not too many simplexes are removed.

Property 1. *Let us put:* $\tilde{k}_n(a, \lambda) = \lceil \lambda a \ln n \rceil$

If hypotheses ***H1*** *($\mathcal{X}_n \subset \mathbb{R}^d$ and S is $d-$dimensional compact manifold),* ***H2*** *(S is $\alpha-$standard) and* ***H3*** *($f_0 = \max_S(f) > 0$) are satisfied then, there exists an explicit value a such that:*

- *For all $\lambda > 1$ $D_{(S)} \subset D_{\tilde{k}_n(a,\lambda)}$ asymptotically almost surely;*
- *For all $\lambda > 2$ $D_{(S)} \subset D_{\tilde{k}_n(a,\lambda)}$ eventually almost surely.*

If, additionally, the hypotheses ***H4*** *(Lipschitz Continuity of f) and/or* ***H5*** *(smoothness of the boundary) are satisfied, the explicit constant a can be adapted. The values of a corresponding to different sets of hypotheses are given in the following table:*

hypothesis	*notation*	a
H1, ***H2*** *and* ***H3***	$\tilde{a}_d(f, S)$	$\frac{f_1}{f_0}\frac{1}{\alpha}2^d(1 + \varphi^{-1}(2^{-d-1}))$
H1, ***H2***, ***H3*** *and* ***H4***	$\tilde{a}_d(f, \cdot)$	$\frac{f_1}{f_0}2^{d+1}(1 + \varphi^{-1}(2^{-d-1}))$
H1, ***H2***, ***H3*** *and* ***H5***	$\tilde{a}_d(\cdot, S)$	$\frac{1}{\alpha}2^d(1 + \varphi^{-1}(2^{-d-1}))$
H1, ***H2*** , ***H3***, ***H4*** *and* ***H5***	$\tilde{a}_d$	$2^{d+1}(1 + \varphi^{-1}(2^{-d-1}))$

	$d=1$	$d=2$	$d=3$	$d=4$	$d=5$	$d=6$	$d=7$
$\tilde{a}_d$	7.15	12.32	21.98	40.33	75.64	144.33	278.96
n^{min}	23	48	102	217	465	997	2140

Table 1: Constants for the number of neighbors and minimal number of points required according to the dimension.

with $f_1 = \max_S(f)$ *and* $\varphi(x) = (x+1)\ln(x+1) - x$

As the last case is the most useful in practice, we give in Table 1 some numerical values. Values of $\tilde{a}_d$ are given for $d \in \{1,2,3,4,5,6,7\}$. The last line of the table gives values for the smallest value of n (n^{min}), such that $\tilde{k}_n(\tilde{a}_d, 1) < n$. This is the smallest number of observations that allows to deal with the given dimension.

2.2.2 Choice of the parameter λ

In practice we have to choose a value for λ to compute the Delaunay restriction. In his section we show that, even if it is not proved by Property 1, a choice $\lambda = 1$ gives satisfying results. Here we have chosen to draw uniform sample in $d-$dimensional unit balls (even if it can seems a very restrictive example the manifold hypotheses on the support ensure that it is locally relevant). For every values for d ($d \in \{1,2,3,4\}$) and for every value for n ($n \in \{100, 200, ..., 1000\}$) 1000 sample are drawn. For each sample $\mathcal{X}_n$ it is possible (because the support is known) to compute $D_{(S)}$ and to extract the minimum value for k ($k_n^*(\mathcal{X}_n)$) such that the graph associated to $D_{(S)}$ is included in the $k-$nearest neighbor graph. In Figure 4 we plot different percentiles for $k_n^*(\mathcal{X}_n)/\ln n$ and compare them to $\tilde{a}_d$. The following percentiles are computed: 90%, 95%, 99%, 99.5% and the max.

It seems that $k_n(\mathcal{X}_n)/\ln n$ is asymptotically inferior to $\tilde{a}_d$. This justify that in almost every presented examples in the paper, it will be chosen to work with $k_n(\tilde{a}_d, 1)$ neighbors.

2.2.3 Topological properties

Values for k_n such that $D_{(S)}$ can be reasonably supposed to be included in D_{k_n} are now founded. Next, the two following points have to studied:

- To have $D_{(S)} \cong S$, The closed ball property of $\mathcal{X}_n$ on S needs to be verified. Edelsbrunner argued about the reasonability of such an assumption in when S is smooth. Can that purpose be probabilistically specified?

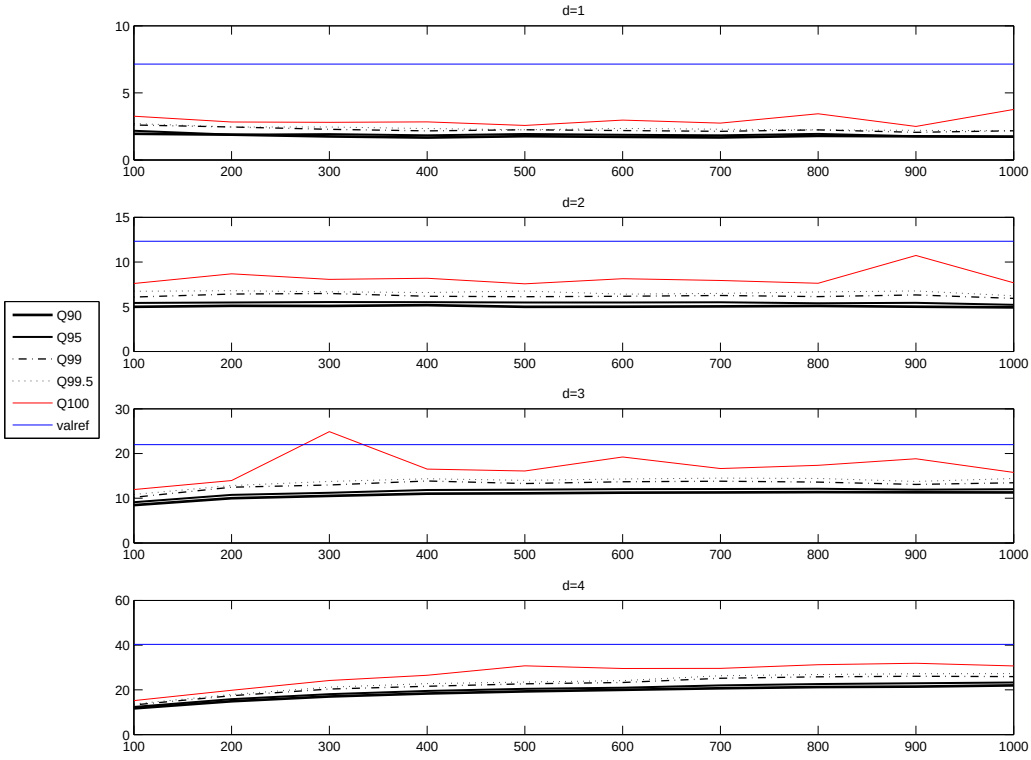


Figure 4: Choice of the parameter

- Unfortunately we can not expect that $D_{k_n} = D_{(S)}$. In Figure 5 we can see that nothing prevents existence of small simplices having the centers of their circumscribed spheres outside of S (the red triangles). Nevertheless can it be established that $D_{k_n} \cong D_{(S)}$?

The two-dimensional case. The specificity of the case $d = 2$ is that the above points can be theoretically answered by Lemma 1 that ensures that the closed ball property is satisfied a.a.s and Theorem 2 that establishes the existence of an homeomorphism a.a.s.

Lemma 1. *If $d = 2$ and if hypotheses **H1** ($\mathcal{X}_n \subset \mathbb{R}^d$ and S is $d-$dimensional compact manifold), **H2** (S is $\alpha-$standard), **H3** ($f_0 = \max_S(f) > 0$) and **H4** (here ∂S is a $\mathcal{C}^2$ $1-$manifold) are satisfied then, $\mathcal{X}_n$ has the closed ball property a.a.s.*

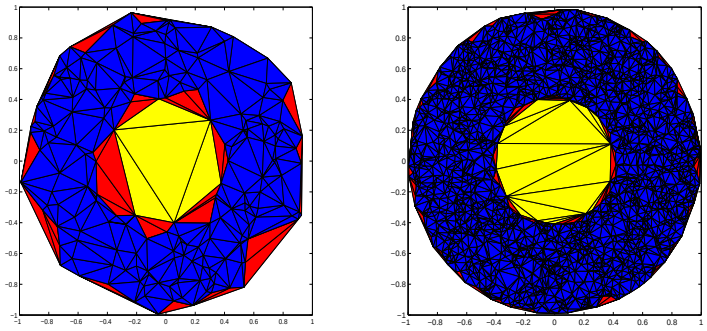


Figure 5: Delaunay polyhedron (yellow), Delaunay polyhedron restricted to its nearest neighbors (red) and Delaunay polyhedron restricted to the support (blue) on an example.

Theorem 2. *Let us put:* $\tilde{k}_n(a,\lambda) = \lceil \lambda a \ln n \rceil$. *If* $d = 2$ *and if hypotheses* ***H1*** *(*$\mathcal{X}_n \subset \mathbb{R}^d$ *and* S *is* $d-$*dimensional compact manifold),* ***H2*** *(*S *is* $\alpha-$*standard),* ***H3*** *(*$f_0 = \max_S(f) > 0$*),* ***H4*** *(here* ∂S *is a* $\mathcal{C}^2$ $1-$*manifold) and* ***H5*** *(Lipschitz continuity of* f*) are satisfied then, for all* $\lambda > 1$, $D_{\tilde{k}_n(\tilde{a}_2,\lambda)}$ *is homeomorph to* S *a.a.s. with* $\tilde{a}_2 = 8(1+\varphi^{-1}(8^{-1}))$

Of course a similar result can be obtained without the hypothesis **H5** replacing $\tilde{a}_2$ by $\tilde{a}_2(.,f)$

Discussion for other dimensions. For the the case $d = 2$, the proofs are relatively easy because all the phenomena that could contradict the closed ball property can be classified and studied. The second point is that it can be clearly seen that in this case the restriction D_{k_n} is homeomorph to D_S (if $D_{(S)} \subset D_{k_n}$ and k_n/n is small enough).

For dimensions higher than 2, the generalization of Lemma 1 and Theorem 2 is still an open problem, but it can be reasonably assumed that the asymptotic behavior of the Delaunay polyhedron restricted to the k_n-nearest neighbor is still good.

First, for the closed ball property we can refer to the discussion done in [21], where Edelsbrunner pointed out that it is a very reasonable hypothesis when the manifold is $\mathcal{C}^2$.

Let us now discuss the existence of a homeomorphism between $D_{(S)}$ and D_{k_n}. First, let us remark that the suitable sequences k_n exhibited in Property 1 satisfy the usual conditions for nearest neighbors statistics: $k_n \to \infty$ and $k_n/n \to 0$. The second condition ensures that the kept simplexes are small

$n =$	100	110	120	130	140	150
homeomorph to $\overline{\mathcal{B}}$	79.8%	36.6%	11.8%	3.6%	0.5%	0%
homeomorph to S	20.2%	63.4%	88.2%	96.4%	99.5%	100%

Table 2: Results for simulations on $\mathcal{B}_2(0,1) \setminus \mathcal{B}_2(0,0.5)$. 1000 samples for each value for n.

(their edge's length converges to 0) and so, unsuitable edges (as those that fill in holes for instance) are removed. The remaining problem is the following: can the restriction to the nearest neighbor create undesirable phenomena near the boundary ?

Let us remark that choosing $k'_{d,n}(a,\lambda) = \lceil \lambda\alpha a \ln n \rceil$ (with a in the set of suitable values exhibited in Property 1 i.e. $\{\tilde{a}_d(f,S), \tilde{a}_d(.,S), \tilde{a}_d(f,.), \tilde{a}_d\}$ according to the supposed hypotheses) avoid the creation of undesirable phenomena far from the boundary and gives to every points far from the boundary a neighborhood homeomorph to a ball. The use of $\tilde{k}_{d,n}(a,\lambda) = \lceil \lambda a \ln n \rceil$ take the boundary effect into account and it is reasonable that every points near the boundary has a suitable neighborhood (that contains a set homeomorph to a close ball). An empirical validation of this discussion can be found in Table 3.

2.2.4 Some simulations

In Figure 2.2.4 we present a simulated restricted Delaunay polyhedron 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). Even if it can be consider as a very easy toy example this geometric figures allows to observe almost all the local cases that can occurs in case when $d = 2$ and the boundary is smooth.

Here we have chosen to use $\tilde{k}_n(\tilde{a}_d, 1)-$nearest neighbors restriction.

To illustrate the convergence towards 1 of the probability that $D_{\tilde{k}_n(\tilde{a}_d,1)}(\mathcal{X}) \cong S$, we present in Table 2 (resp. Table 3) the results of 1000 samples of size n uniformly drawn in $\mathcal{B}_2(0,1) \setminus \mathcal{B}_2(0,0.5)$ (resp. $\mathcal{B}_3(0,1) \setminus \mathcal{B}_3(0,0.5)$). We have only observed two different behaviors of the restricted Delaunay polyhedron: first, the inside hole is unfortunately filled in and the polyhedron is homeomorph to a ball; second, the polyhedron is homeomorph to S. No other cases (creation of unexpected phenomena near the boundary) had been observed. For the two experiences, the convergence of the probability to be homeomorph to the support towards 1 (which is proved when the dimension is 2 and not proved when the dimension is 3) can be observed .

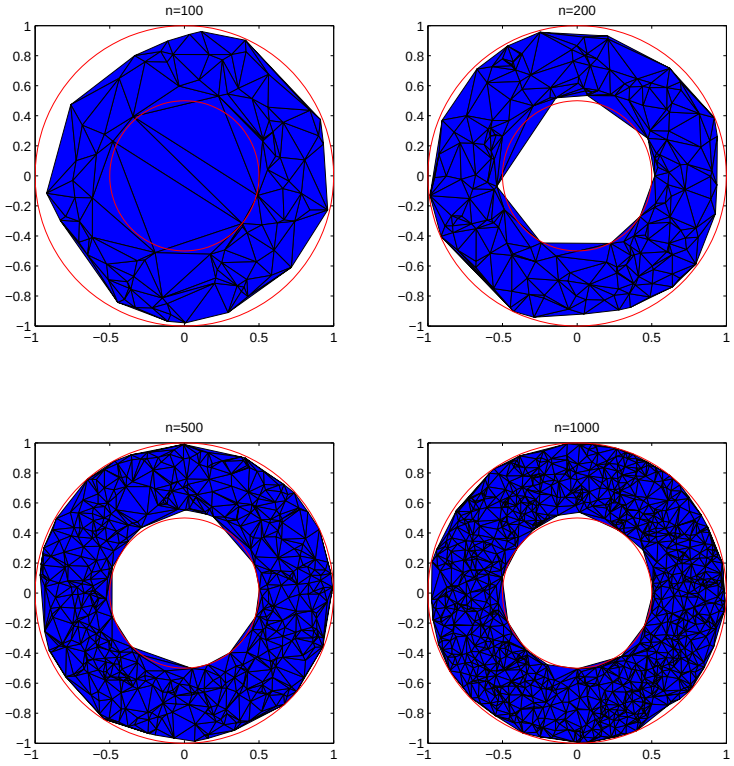


Figure 6: Restricted Delaunay polyhedron for the CD-Rom example.

2.3 Estimation of the support

Beyond the topological aspect, it can be interesting to observe the properties of the restricted Delaunay polyhedron as an estimator of the density support. The different lemmas needed in the proofs of the results of the previous section allow to quickly obtain a first result which will be given in Theorem 3. The convergence rate seems to be good (the same that for Devroye-Wise estimator in [18] where an analogous idea is developed : we wish $D_{(S)} \subset D_{k_n}$ while, in [18] authors wish $S \subset \hat{S}_n$). But we will see on simulation results that our Theorem 3 seem to be highly improvable to reach a convergence rate that may not be so different that the rate of the convex hull as an estimator of the support under the convexity hypothesis.

As will also be interested in the case where the dimension of the support is smaller than the dimension of the ambient space, we need first to choose

$n =$	200	220	240	260	280	300
homeomorph to $\overline{\mathcal{B}}$	87.9%	50.1%	21.2%	7.5%	1.8%	0.3%
homeomorph to S	12.1%	49.9%	78.8%	92.5%	98.2%	99.70%

Table 3: Results for simulations on $\mathcal{B}_3(0,1) \setminus \mathcal{B}_3(0,0.5)$. 1000 samples for each value for n.

carefully a distance between sets. This necessity will be illustrated in section 3.4.1.

2.3.1 Choice of a distance between sets

Usually to measure the distance between the support and its estimator the symmetric difference distance is used.

If such a distance measure is suitable when the dimension of the support is the dimension of the ambient space, it is no longer the case when the dimension of the support is smaller than the dimension of the ambient space.

As in [18] we propose here to use the Hausdorff distance between sets instead of the symmetric difference distance.

When the dimension of the support, the dimension of the ambient space and the dimension of the estimator of the support are equal there is a strong link between the Hausdorff distance and the symmetric difference distance via the Steiner Minkowski formula:

Definition 13 (Steiner-Minkowski formula)**.** *let $d \geq 2$, and $A \subsetneq \mathbb{R}^n$ be a $d-$dimensional manifold with a $\mathcal{C}^2$ boundary. Let μ be the Lebesgue measure. Then the boundary of S is measurable and its measure $\lambda(\partial A)$ is given by the following formula (Minkowski-Steiner):*

$$\lambda(\partial A) = \liminf_{r \to 0} \frac{\mu(A + \overline{\mathcal{B}}) - \mu(A)}{r},$$

This formula seems intuitive but the proof is not so simple [23]. But a consequence is that the Hausdorff convergence implies the convergence according to the symmetric difference distance via the following property:

Property 2. *If $d_H(S, \hat{S}_n) = \varepsilon_n \to 0$, then $\mu(S \Delta \hat{S}_n) = O(\varepsilon_n)$.*

- $\hat{S}_n \subset S + \varepsilon_n \overline{\mathcal{B}} \Rightarrow V(\hat{S}_n \setminus S) \leq V(S + \varepsilon_n \overline{\mathcal{B}} \setminus S) \sim \varepsilon_n \lambda(\partial S)$
- $S \subset \hat{S}_n + \varepsilon_n \overline{\mathcal{B}} \Rightarrow V(S \setminus \hat{S}_n) \leq V(\hat{S}_n + \varepsilon_n \overline{\mathcal{B}} \setminus \hat{S}_n) \sim \varepsilon_n \lambda(\partial S)$

2.3.2 Consistency of the support estimator

The consistency of the density support estimated by the restricted Delaunay polyhedron is given by theorem 3.

Theorem 3. *Let us put:* $\tilde{k}_n(a,\lambda) = \lceil \lambda a \ln n \rceil$.

For each density f *and support* S*, satisfying hypotheses* ***H1*** *(*$\mathcal{X}_n \subset \mathbb{R}^d$ *and* S *is* $d-$*dimensional compact manifold),* ***H2*** *(*S *is* $\alpha-$*standard) and* ***H3*** *(*$f_0 = \max_S(f) > 0$*), there exists an explicit constant* $M(S,f)$ *such that for all* $\lambda > 2$ *and a suitable values as in Property 1.*

$$d_H(D_{\tilde{k}_n(a,\lambda)}, S) \left(\frac{n}{\ln n}\right)^{1/d} \leq M(S,f) \ \ e.a.s.$$

If, additionally, the hypotheses ***H4*** *(Lipschitz Continuity of* f*) and/or* ***H5*** *(smoothness of the boundary) are satisfied, the explicit constant* $M(S,f)$ *can be adapted. The values of* $M(S,f)$ *corresponding to different sets of hypotheses are given in the following table:*

hypothesis	a	$M(S,f)$
H1, ***H2*** *and* ***H3***	$\tilde{a}_d(f,S)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-d}))}{\theta_d \alpha f_0}, \frac{2}{\theta_d \alpha f_0}\right)^{1/d}$
H1, ***H2***, ***H3*** *and* ***H4***	$\tilde{a}_d(f,.)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-d}))}{\theta_d \alpha f_0}, \frac{\max\left(1, 2(1-\frac{1}{d})\frac{f_0}{\overline{f}_0}\right)}{\theta_d f_0}\right)^{1/d}$
H1, ***H2***, ***H3*** *and* ***H5***	$\tilde{a}_d(.,S)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-d}))}{\theta_d \alpha f_0}, \frac{2}{\theta_d \alpha f_0}\right)^{1/d}$
H1, ***H2*** , ***H3***, ***H4*** *and* ***H5***	$\tilde{a}_d$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-d}))}{\theta_d \alpha f_0}, \frac{\max\left(1, 2(1-\frac{1}{d})\frac{f_0}{\overline{f}_0}\right)}{\theta_d f_0}\right)^{1/d}$

with $f_1 = max_S(f)$, $\overline{f}_0 = \min_{\partial S}(f)$, $\varphi(x) = (x+1)\ln(x+1) - x$ *and* $\psi(x) = x - \ln(1+x)$.

2.3.3 Some simulations

In this section (and only here), we use the symmetric difference distance. This choice has been done in order to compare the restricted Delaunay polyhedron with the Devroye-Wise estimator using the most usual distance in the literature. Each time we have use uniform samples to avoid any discussion about the fact that local radius may over-perform the fixed one for the Devroye-Wise estimator.

In Figure 7 we present the results for the uniform samples on a 2 dimensional ball and on a 2 dimensional "CD-Rom". On the first line of each example we represent the restricted Delaunay polyhedron (with $k_n = \tilde{k}_n(\tilde{a}_2, 1)$).

On the second line of each example, we represent the Devroye-Wise estimator with an optimal choice of the radius (here the real support is known, and so we can choose the radius that minimize the symmetric difference distance). In the ball example, the performance of the two estimators seems to be comparable. However it will be clear, from Figure 8 that the restricted Delaunay estimator over-perform the Devroye-Wise estimator. In the "CD-Rom" example, the restricted Delaunay is worse than the Devroye-Wise estimator for small values of n, but seems to be comparable when $n \geq 150$. As in the ball example, the restricted Delaunay estimator becomes in fact quickly better than the Devroye-Wise estimator.

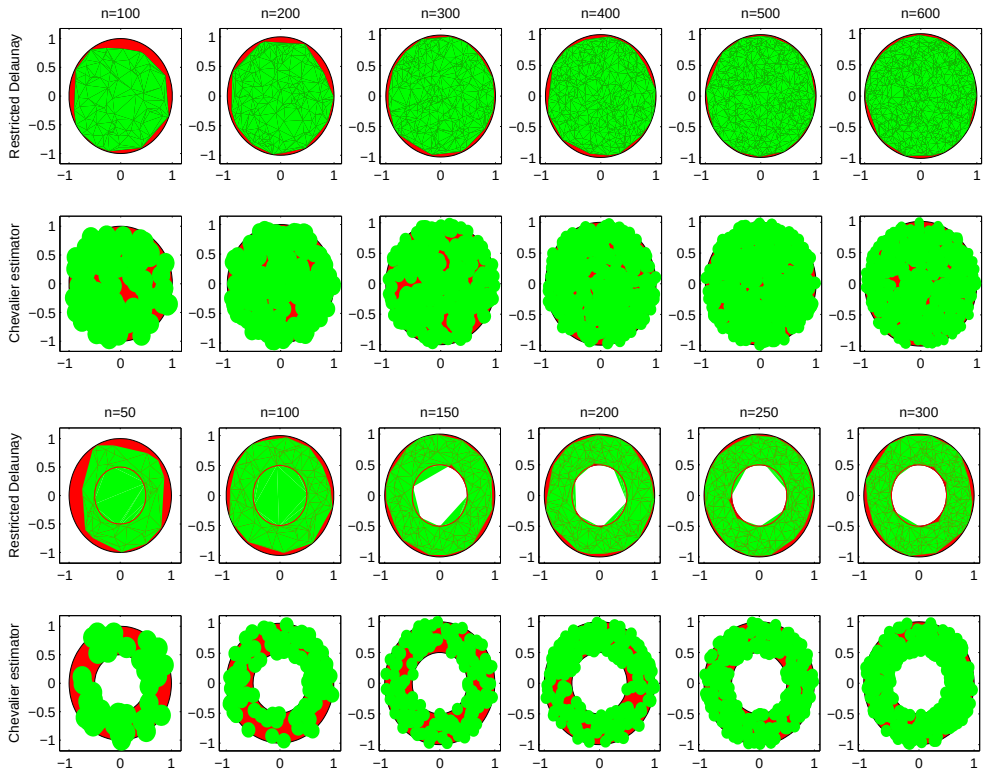


Figure 7: Comparison between restricted Delaunay polyhedron and Devroye-Wise estimator on two examples: sample in a ball (first line the restricted polyhedron and second line the best Devroye-Wise estimator) and in a CD-Rom (third line the restricted polyhedron and fourth line the best Devroye-Wise estimator).

In Figure 8 we plot the histogram of the density estimation of the symmetric difference distance for different estimator of the support: the restricted

Delaunay polyhedron, the Devroye-Wise estimator (the best one as previously) and the convex hull (that is the Delaunay polyhedron). The chosen shape is ball (for dimension 2 and 3). This choice allows to compute easily the volume of the symmetric difference distance using Monte-Carlo integration. It also allows to point out that the Restricted Delaunay polyhedron is very similar to the convex hull estimator (which can be assumed to be the best one in the convex case). As previously, even if it can appear as very restricted, this choice of ball shape is justified by the manifold hypotheses. For each $n \in \{50, 100, 200, 500, 1000\}$ we compute 200 uniform samples. For each sample the optimal (according to the symmetric difference distance) radius for the Devroye-Wise estimator is computed (Monte Carlo) and the symmetric difference distance for the three proposed estimators are computed.

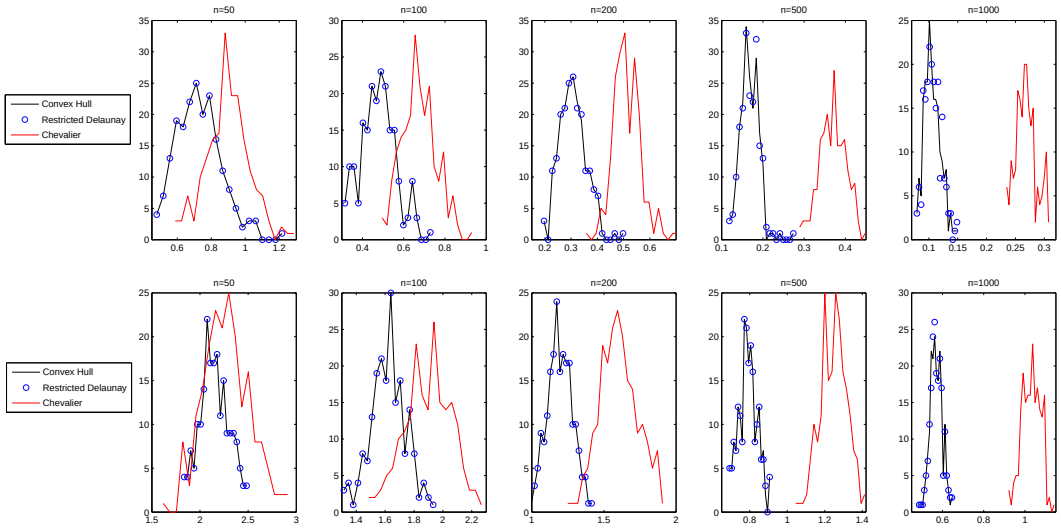


Figure 8: Comparison between the convex hull, restricted Delaunay polyhedron and Devroye-Wise estimator: histograms for 200 uniform samples of size n in a ball for dimension 2 (first line) and dimension 3 (second line).

3 Case when the dimension of the support is smaller than the dimension of the observation space

In this section we focus on a more interesting case when the dimension of the support (d') is smaller than the dimension of the ambient observation space

(d). In this case, the definitions of the support and of the density are a little more tricky. If we denote $\mathcal{E}$ the set of closed sets $E \subset \mathbb{R}^d$ such that $P(E) = 1$ the support can be defined by: $S = \bigcap_{E \in \mathcal{E}} E$. Once the support is defined, a finite measure μ_S associated to the support can be constructed with the use of the Lebesgue measure on $\mathbb{R}^d$ μ as follows. For all $A \subset S$:

$$\mu_S(A) = \lim_{x \to 0} \frac{\theta_{d'} \mu(A + x\mathcal{B}_d)}{\theta_d x^{d-d'}}$$

So, finally, the density f w.r.t. to the measure μ_S can be defined.

In the following section we will set up the needed hypothesis on S and f. Then we will find suitable values k_n that ensure that $D_{(S)} \subset D_{k_n}$ a.a.s. (and e.a.s.). A procedure to remove simplexes of dimension higher than d' is presented after. The final algorithm is tested on several data sets.

3.1 Hypotheses

Throughout this section S will be supposed to have the two following properties:

- **H'1**: S is a compact $\mathcal{C}^2$, $d'-$dimensional manifold with $d' < d$.
- **H'2**: S is $d'-$dimensional $\alpha-$standard, that is, there exists ε_0 such that: for all $x \in S$ and for all $\varepsilon \leq \varepsilon_0$, we have $\mu_S(\mathcal{B}(x, \varepsilon) \cap S) \geq \alpha \theta_{d'} \varepsilon^{d'}$.

The density f has to satisfy the hypothesis:

- **H3**: $\inf\{f(x), x \in S\} = f_0 > 0$. We will also denote $\overline{f}_0 = \inf\{f(x), x \in \partial S\}$ (obviously, as $f_0 > 0$, we have $\overline{f}_0 > 0$ when $\partial S \neq \emptyset$).

Three additional hypotheses are also studied

- **H4**: S is a manifold with a $\mathcal{C}^2$ boundary.
- **H'4**: S is a manifold without boundary.
- **H5**: $f|_S$ is Lipschitz continuous.

3.2 Number of neighbors

To preserve the topology we need to find a polyhedron made of $d'-$simplexes. So, we will first restrict the Delaunay polyhedron to nearest neighbors, then propose an algorithm that is expected to keep only $d'-$dimensional simplexes. As in the previous part, we want to choose a number of neighbors that ensures

that the Delaunay polyhedron restricted to to nearest neighbors contains the Delaunay polyhedron restricted to the support. When S is a $d'-$manifold, the analogue of Property 1 is given by property 3.

Property 3. *Let us put:* $\tilde{k}_n(a,\lambda) = \lceil \lambda a \ln n \rceil$

If hypotheses ***H'1*** *(S is a $d'-$dimensional $\mathcal{C}^2$ manifold),* ***H'2*** *(S is $\alpha-$standard) and* ***H3*** *($f_0 = \min_S(f) > 0$) are satisfied then, there exists an explicit value for a such that:*

- *for all $\lambda > 1$, $D_{(S)} \subset D_{\tilde{k}_n(a,\lambda)}$ asymptotically almost surely;*
- *for all $\lambda > 2$, $D_{(S)} \subset D_{\tilde{k}_n(a,\lambda)}$ eventually almost surely.*

If, additionally, the hypotheses ***H4*** *(smoothness of the boundary) or* ***H'4*** *(absence of boundary) and/or* ***H5*** *(Lipschitz continuity of f) are satisfied, the explicit constant a can be adapted. The values of a corresponding to different sets of hypotheses are given in the following table:*

hypothesis	*notation*	a
H'1, ***H'2*** *and* ***H3***	$\tilde{a}_{d'}(f,S)$	$\frac{f_1}{f_0}\frac{1}{\alpha}2^{d'}(1+\varphi^{-1}(2^{-d'-1}))$
H'1, ***H'2***, ***H3*** *and* ***H4***	$\tilde{a}_{d'}(f,\cdot)$	$\frac{f_1}{f_0}2^{d'+1}(1+\varphi^{-1}(2^{-d'-1}))$
H'1, ***H'2***, ***H3*** *and* ***H5***	$\tilde{a}_{d'}(\cdot,S)$	$\frac{1}{\alpha}2^{d'}(1+\varphi^{-1}(2^{-d'-1}))$
H'1, ***H'2*** , ***H3***, ***H4*** *and* ***H5***	$\tilde{a}_{d'}$	$2^{d'+1}(1+\varphi^{-1}(2^{-d'-1}))$
H'1, ***H'2***, ***H3*** *and* ***H'4***	$0.5\tilde{a}_{d'}(f,\cdot)$	$\frac{f_1}{f_0}2^{d'}(1+\varphi^{-1}(2^{-d'-1}))$
H'1, ***H'2*** , ***H3***, ***H'4*** *and* ***H5***	$0.5\tilde{a}_{d'}$	$2^{d'}(1+\varphi^{-1}(2^{-d'-1}))$

with $f_1 = \max_S(f)$ and $\varphi(x) = (x+1)\ln(x+1) - x$.

3.3 Estimation of the support

3.3.1 The choice of the Hausdorff distance

When the dimension of S is different from the dimension of the ambient space, there is no more relation between the symmetric difference distance and the Hausdorff one. The choice of the Hausdorff distance between sets is still meaningful. That is not the case of the symmetrical difference distance. Indeed, the estimator G_{d',k_n} and S are two $d'-$dimensional sets. So, if μ is the Lebesgue measure on $\mathbb{R}^d$, $\mu(G_{d',k_n}\Delta S) = 0$ (even if the support and its

estimator are very different). If we choose instead of μ , a $d'-$dimensional "measure" μ'there is high chance to have a measure 1 (even if the support and its estimator are very similar).

3.3.2 Consistency of the estimator

The following analogue of Theorem 3 holds:

Theorem 4. *Let us put:* $\tilde{k}_n(a,\lambda) = \lceil \lambda a \ln n \rceil$

*For each density f and support S, satisfying hypotheses **H'1** (S is a $d'-$dimensional $\mathcal{C}^2$ manifold), **H'2** (S is $\alpha-$standard) and **H3** ($f_0 = \min_S(f) > 0$), if a denotes the explicit constant of Property 3, there exists an explicit constant $M(S,f)$ such that*

$$d_H(D_{\tilde{k}_n(a,\lambda)}, S)\left(\frac{n}{\ln n}\right)^{1/d'} \leq M(S,f) \ \ e.a.s.$$

*If, additionally, the hypotheses **H4** (smoothness of the boundary) or **H'4** (absence of boundary) and/or **H5** (Lipschitz continuity of f) are satisfied, the explicit constant $M(S,f)$ can be adapted. The values of $M(S,f)$ corresponding to different sets of hypotheses are given in the following table:*

hypothesis	a	$M(S,f)$
H1**, **H2** and **H3	$\tilde{a}_{d'}(f,S)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{2}{\theta_d \alpha f_0}\right)^{1/d}$
H1**, **H2**, **H3** and **H4	$\tilde{a}_{d'}(f,.)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{\max\left(1, 2(1-\frac{1}{d})\frac{f_0}{\bar{f}_0}\right)}{\theta_d f_0}\right)^{1/d}$
H1**, **H2**, **H3** and **H5	$\tilde{a}_{d'}(.,S)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{2}{\theta_d \alpha f_0}\right)^{1/d}$
H1**, **H2** , **H3**, **H4** and **H5	$\tilde{a}_{d'}$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{\max\left(1, 2(1-\frac{1}{d})\frac{f_0}{\bar{f}_0}\right)}{\theta_d f_0}\right)^{1/d}$
H'1**, **H'2**, **H3** and **H'4	$0.5\tilde{a}_{d'}(f,.)$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{2}{\theta_d \alpha f_0}\right)^{1/d}$
H'1**, **H'2** , **H3**, **H'4** and **H5	$0.5\tilde{a}_{d'}$	$\max\left(\frac{\lambda 2a(1+\psi^{-1}(2^{-2}))}{\theta_d \alpha f_0}, \frac{\max\left(1, 2(1-\frac{1}{d})\frac{f_0}{\bar{f}_0}\right)}{\theta_d f_0}\right)^{1/d}$

with $f_1 = \max_S(f)$, $\varphi(x) = (x+1)\ln(x+1) - x$ *and* $\psi(x) = x - \ln(1+x)$.

3.4 Proposed algorithm to recover the topology

To recover the topological properties, we propose to compute a new polyhedron: the $d'-$dimensional and k_n-nearest-neighbor restriction of D, $G_{k_n,d'} = G_{k_n,d'}(\mathcal{X}_n)$ which is constructed as follows.

Definition 14. *The $d'-$dimensional and k_n-nearest-neighbor restriction of the Delaunay polyhedron, $G_{k_n,d'}$ can be algorithmically defined by:*

1) First compute D_{k_n} and extract the polyhedron $D_{k_n,d'}$ defined as follows: a simplex σ belongs to $D_{k_n,d'}$ if it belongs to D_{k_n} and if it is a $d''-$dimensional simplex (with $d'' \leq d'$).

2) A $d'-$simplex σ belongs to $G_{k_n,d'}$ if σ belongs to $D_{(D_{k_n,d'})}$.

3) When $d'' < d'$, a $d''-$simplex σ belongs to $G_{k_n,d'}$ if it is a sub-simplex of a $d'-$simplex.

4) For $d'' = d' + 1$ to d: A $d''-$simplex σ belongs to $G_{k_n,d'}$ if all its faces ($d'' - 1-$sub-simplexes) belongs to $G_{k_n,d'}$.

The idea that lead us to the construction of $G_{k_n,d'}$ is the Following:

- First we want to find all the "good" $d'-$simplexes. It is so natural to look for the $d'-$simplexes of D_{k_n} which are those of $D_{k_n,d'}$.
- As D_{k_n} is a consistent estimator of S which is $d'-$dimensional, we can reasonably expect $D_{k_n,d'}$ to be a consistent estimator of S. In this polyhedron, there exists some non-retractable undesirable "holes" in place of $d''-$simplexes of D_{k_n} (with $d'' > d'$). The idea is to keep only simplexes that are in $D_{(D_{k_n,d'})}$ (the Edelsbrunner's restriction). This step allows to remove most of this phenomenas. To finish we apply the rules 3) and 4) to achieve the construction.

3.5 Choice of the dimension

For a chosen dimension d^* we can compute G_{k_n,d^*}. Our intuitive idea is that if d^* underestimates the true intrinsic dimension d', then there exists many $(d^* + 1)-$simplexes in G_{k_n,d^*}. So, for each d^*, we can compute $\rho(d^*)$ that represents the part of the d^*-simplexes that are sub-simplexes of a $d^* + 1-$simplex. If we have no preliminary knowledge of the intrinsic dimension then we propose to compute $G_{\tilde{k}_n(\tilde{a}_1,1),1}, \ldots, G_{\tilde{k}_n(\tilde{a}_{d^*},1),d^*}$ and to stop when $\rho(d^*)$ can be consider small enough.

3.6 Examples

3.6.1 Simulated $1-$dimensional manifold

First we present in Figure 9 simulation results for recognition of $1-$dimensional sets. The first one is a manifold. 100 points are drawn on a spiral in $\mathbb{R}^3$ (with

uniform law for the angle and the height). The spiral form is recognized and no undesired 2 or 3−dimensional simplexes are conserved. The second and third examples are drawn on an "X" form. This is not a manifold. The general form is recognize but some undesired simplexes are conserved in the center, where the manifold hypothesis is not satisfied.

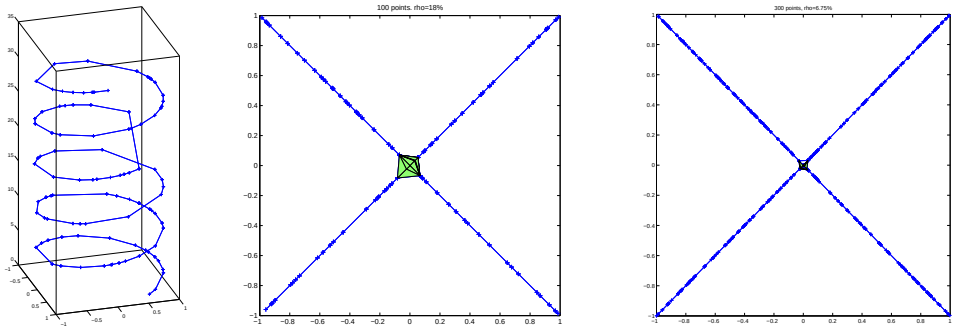


Figure 9: From left to right: 100 points on a $3d-$spiral ($\rho(1) = 0$); 100 points on an "X' ($\rho(1) = 18\%$); 300 points on an "X" ($\rho(1) = 6.75\%$).

3.6.2 Simulated $2-$dimensional manifold

We present in Figure 10 simulation results for $2-$dimensional manifolds in $\mathbb{R}^3$. 300 points are uniformly drawn on a cylinder. When the dimension 1 is tested, $\rho(1) = 0.94$ that clearly indicates an underestimation of the dimension. For the dimension 2, $\rho(2) = 0.016$ that means that there exists some residual $3-$dimensional simplex but they are few.

3.6.3 Simulated $3-$dimensional manifold

We present in Figure 11 simulation results for a $3-$dimensional manifold in $\mathbb{R}^3$. 400 points are drawn in a thicken cylinder. $\rho(1) = 0.97$ and $\rho(2) = 0.65$ that indicates that the dimension is 3. So the used method is the restricted Delaunay polyhedron.

3.6.4 The Standford Bunny

The Standford Bunny data set (that can be find in the Stanford 3D Scanning Repository project data-base `http://graphics.stanford.edu/data/`

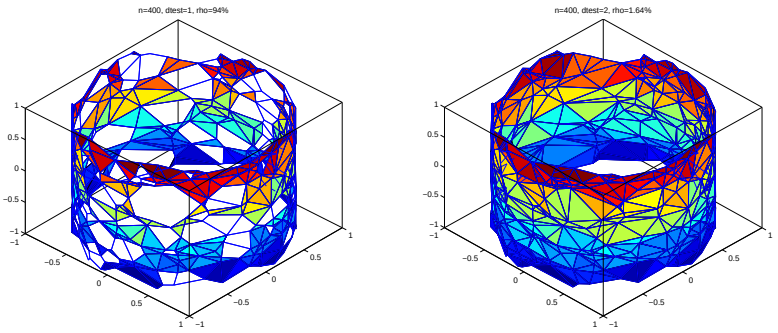


Figure 10: 300 points on a cylinder. Left: the result when the dimension is supposed to be 1. Right the result when the dimension is supposed to be 2.

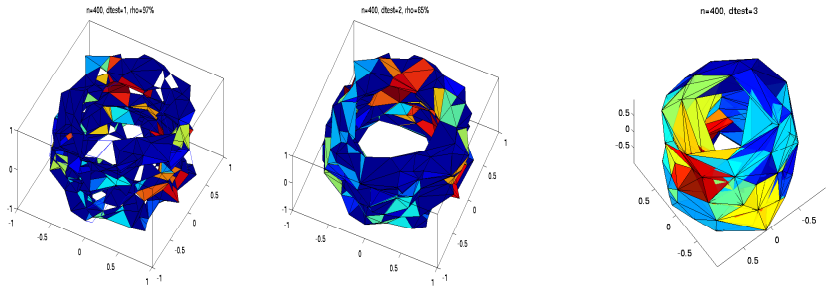


Figure 11: 400 points on a thicken cylinder, 1−dimensional, 2−dimensional and 3−dimensional recognition.

`3Dscanrep/#bunny`) contains 35947 points and 69451 triangles obtained after having scanned a ceramic figurine of a rabbit. Note that the given triangulation on the Bunny data set is obtained with the knowledge of the data construction (physics of the scanner). Here we will treat the data as a sample with no a a priori knowledge on its construction. We will only assume that the dimension 2 is known. We randomly choose $n \in \{200, 300, 500, 1000, 1500, 3000\}$ points in the data base and reconstruct the polyhedron. Table 4 presents the values of $\rho(1)$ and $\rho(2)$ for different values of n. The expected convergence of $\rho(1)$ toward 1 and of $\rho(2)$ toward 0 can be empirically observed but it seems very slow.

Figure 12 presents the different polyhedron estimators, the convergence toward the Bunny surface can be observed. When $n \geq 1000$, the bunny shape is recognized. Figure 13 presents the residual 3−simplexes in the

n	200	300	500	1000	1500	3000
$\rho(1)$	0.8695	0.9015	0.9160	0.9202	0.9251	0.9314
$\rho(2)$	0.4073	0.3872	0.3053	0.2473	0.2577	0.2106

Table 4: ρ values for the Standford Bunny

Bunny recognition. It can be observed on this figure that they seems to have a vanishing importance.

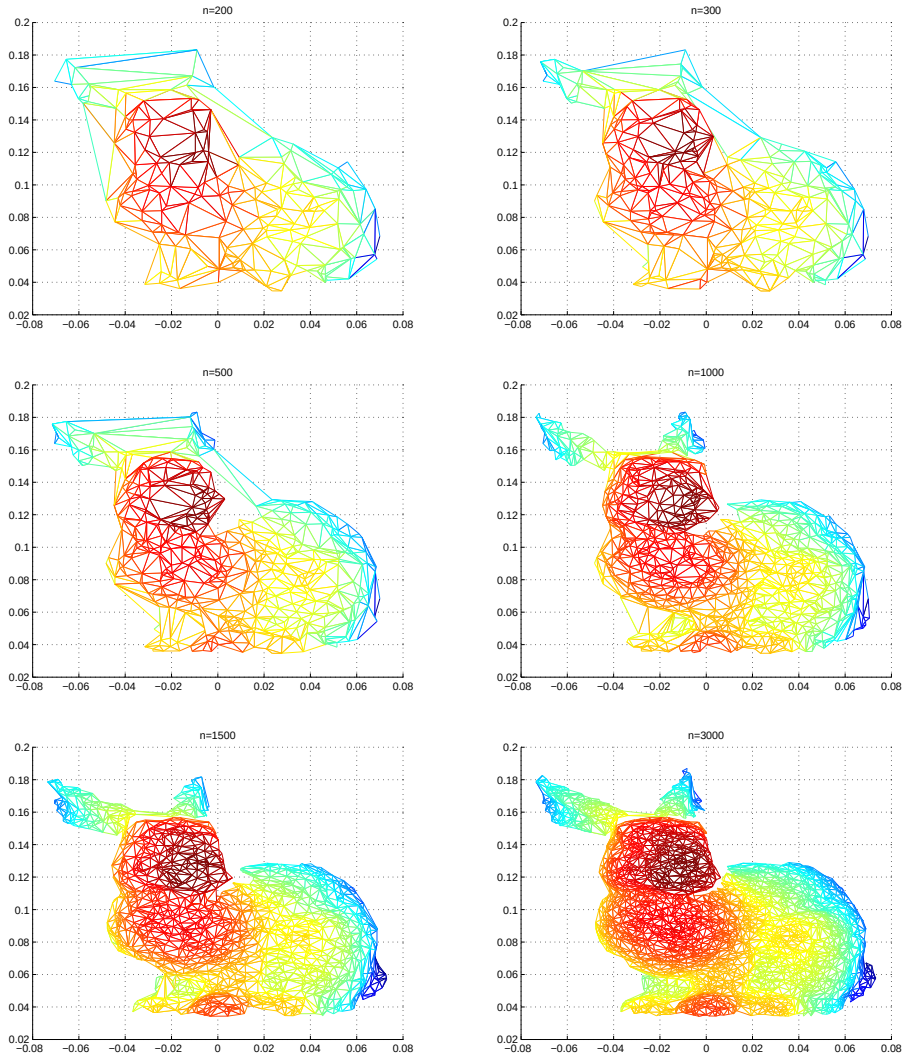


Figure 12: Bunny recognition.

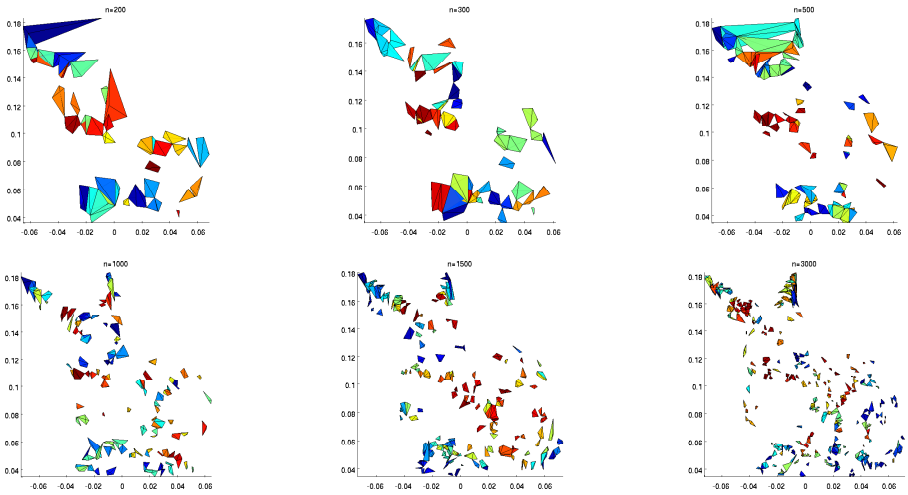


Figure 13: Residual 3−simplexes in the Bunny recognition.

3.6.5 Face data base

We present here the result of our algorithm for the face data set used in the presentation of isomap [30]. The data set contains $n = 698$ pictures (64×64 pixel each, $d = 1024$) of a face with 3 varying parameters (horizontal rotation angle, vertical rotation angle and light direction). The intrinsic dimension is at most 3. A first dimension reduction via isomap is done (with the use of the 40−nearest neighbor graph to compute the geodesic distance) and the chosen dimension is 4 according to the eigenvalue values of the Multi Dimensional Scaling.

When testing the different dimensions between 1 and 3 we have : $\rho(1) = 0.9652$, $\rho(2) = 0.9208$ and $\rho(3) = 0.5524$ and so the dimension 3 is chosen. An

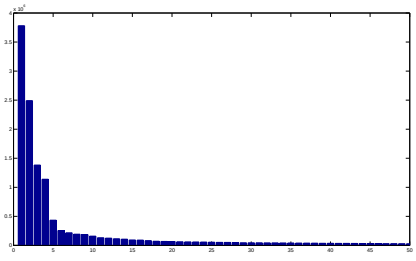


Figure 14: Eigen values for the multi dimensional scaling

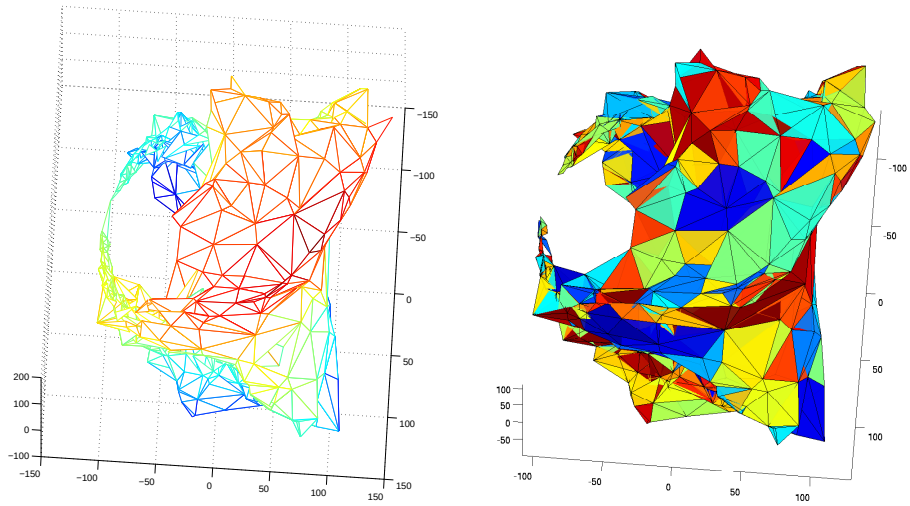


Figure 15: Boundary of $G_{k_n,2}$ and $G_{k_n,3}$ (that can be seen as the interior of $G_{k_n,2}$) (Projection in the 3 first axes of isomap)

interesting phenomena appears here: every 3−simplex of $G_{k_n,3}$ is a 3−simplex of $G_{k_n,2}$. That is why we presents both $G_{k_n,2}$ and $G_{k_n,3}$ in Figure 15.

Another interesting phenomena is the existence of a non retractable cycle in $G_{k_n,2}$ (that does not exist anymore in $G_{k_n,3}$). The presence or not of such a cycle is not an evidence. A non retractable cycle of $G_{k_n,2}$ is illustrated in Figure 16 the faces that are separated in $G_{k_n,3}$ are indicated by the "slash".

4 Conclusion and perspectives

Let us first note the major limitation of our method which is the need of preliminary computation of the Delaunay complex. This requires d to be small enough to have a reasonable computational time. This also requires the data sets to be numerical. Obviously, for large dimension sets, union of balls method such as in ([13], [12] or [14]) may be preferred. When data sets are not numerical, one can choose to use persistent homology [9].

When it can be practically applied, the idea of estimating the support of the density preserving topological properties by the Delaunay complex restricted to the k_n−nearest neighbors gives very good results. First, let us notice the fact that k_n is explicitly known when the density is Lipschitz continuous and the boundary of the support is a $\mathcal{C}^2$ manifold (which are not too

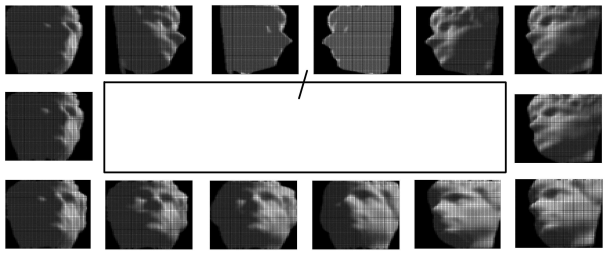


Figure 16: A non retractable cycle of $G_{k_n,2}$.

strong hypotheses). Simulation results are better than what is proved in this paper. Empirically, we have observed the convergence of the probability to have a homeomorphism between the support and its estimator to 1. We have also observed that the convergence rate of the Hausdorff distance between the support and its estimator is higher than the established one. So, there is still some theoretical improvement to be done.

We have an asymptotic result concerning the right choice of the number of neighbors but, as it can be seen in Figure 4, our bound overestimates sufficient values. So when the number of observations is quite small, some holes may be filled by our estimator. Can we find a practical way of choosing the good number of neighbors ?

Last but not least, let us note that several improvements are necessary to make this work applicable to "real-life" problems (as classification).

We need to adapt our method to the estimation of the level sets of the density instead of the whole density support, in order to have a really useful tool for classification and clustering for instance but also for more specific applications such as in [16].

When the support is a $d'-$manifold that can be decomposed into two parts: a principal $d''-$manifold ($d'' < d'$) and small-noise for other directions (as in the face example), can we re-built the principal structure (instead of the whole support)?

A Some useful lemmas

A.1 Elementary lemmas

Lemma 2. *If S is a $d-$dimensional compact manifold, then there exists constants $c_0(S)$ and $r_0(S)$ such that: for all $r \in \mathbb{R}$, $r < r_0(S)$ one can choose*

deterministic $x_1, \ldots, x_\nu$ *such that* $S \subset \bigcup \mathcal{B}(x_i, r)$ *and* $\nu \leq c_0(S) r^{-d}$.

Such a constant ν is usually called the covering number from the inside.

Lemma 3. *If* $S \subset \mathbb{R}^d$ *is a* $d-$*dimensional compact manifold (with boundary* ∂S*), then there exists constants* $c_1(\partial S)$, $\rho_1(S)$ *and* $r_1(S) > 0$ *such that: for all* $\rho \in \mathbb{R}$, $\rho < \rho_1(S)$ *and* $r \in \mathbb{R}$, $r < r_1(S)$ *one can choose deterministic* $x_1, \ldots, x_\nu$ *such that* $\partial S + \rho\overline{\mathcal{B}} \subset \bigcup \mathcal{B}(x_i, r)$ *and* $\nu \leq c_1(\partial S)\rho r^{-d}$.

Proof for Lemmas 2 and 3 are not presented here, They are consequences of covering number properties and are direct corollaries of Lemmas 2.1 and 2.2 in [27]

Lemma 4. *If* S *is a* $d-$*dimensional compact manifold with a* $\mathcal{C}^2$ $(d-1)-$*dimensional boundary, then there exist,* $c_3(S) \geq 0$ *and* $r_3(S) > 0$ *such that: for all* $x \in S$ *and all* $r \leq r_3(S)$, *we have* $V(\mathcal{B}(x, r) \cap S) \geq 0.5(1 - c_3(S)r)\theta_d r^d$.

See Lemma 2.4 in [27].

Lemma 5. *If* f *is a Lipschitz continuous function on a compact* S *and* $\min f = f_0 > 0$, *then* $f^{-1/d}$ *is a Lipschitz continuous function on* S, *and so there exists a constant* $K_{f^{-1/d}}$ *such that* $|f^{-1/d}(x) - f^{-1/d}(y)| \leq K_{f^{-1/d}} ||\overrightarrow{xy}||$.

Lemma 6. *Let* S *be a compact subset of* $\mathbb{R}^2$ *with a* $\mathcal{C}^2$ *boundary. There exists constants* $r_4 > 0$ *and* A_S *such that: for all* X *and* Y *such that* $||\overrightarrow{XY}|| \leq r_4$, *and* $[X, Y] \cap \partial S \neq \emptyset$, *we have* $d_{min}([X, Y], \partial S) \leq A_S ||\overrightarrow{XY}||^2$.

It is a well known result about linear interpolation.

Lemma 7. *Let* S *be a compact subset of* $\mathbb{R}^2$ *with a* $\mathcal{C}^2$ *boundary. For all* $x \in \partial S$, *let us denote* $\overrightarrow{u}_x$ *a vector normal to* ∂S *at the point* x *and that has an unit norm.*

There exists constants $r_S > 0$ *such that: For all* $x \in \partial S$ $\mathcal{B}(x + r_S \overrightarrow{u}_x, r_S) \cap S = \emptyset$ *or* $\mathcal{B}(x - r_S \overrightarrow{u}_x, r_S) \cap S = \emptyset$

Every r_S inferior to the the minimum radius of curvature of ∂S is suitable.

A.2 Other useful lemmas

Lemma 8. *Let* $r(x)$ *be a Lipschitz continuous function on* S. *Suppose that* $\{X_1, \ldots, X_n\}$ *is a sample on* S *drawn with some density such that there exists constants* a_0 *and* b_0 *satisfying: for all* $x \in S$ *we have* $P(X_i \in \mathcal{B}(x, r(x)\rho)) \geq a_0(1 - b_0\rho)\rho^d$. *The following holds:*

if $\lambda > 1$, *then "*$\mathcal{B}(x, r(x)(\lambda \ln n/(a_0 n))^{1/d}) \cap \mathcal{X}_n \neq \emptyset$ *for all* $x \in S$*" a.a.s.;*

if $\lambda > 2$, then "$\mathcal{B}(x, r(x)(\lambda \ln n/(a_0 n))^{1/d}) \cap \mathcal{X}_n \neq \emptyset$ for all $x \in S$" e.a.s.

Proof. First, let us cover S with small deterministic balls of radius $\varepsilon_n(\lambda \ln n/(a_0 n)^{1/d}$ centered in $x_1^*, \ldots, x_\nu^*$ ($\nu \leq c_0(S)\varepsilon_n^{-d} n/\ln n$).

Let us suppose that there exist an x such that $\mathcal{B}(x, r(x)(\ln n/n)^{1/d})$ does not contain any observation. There exits an i such that $x \in \mathcal{B}(x_i^*, \varepsilon_n(\lambda \ln n/(a_0 n))^{1/d})$, and so $\mathcal{B}(x_i^*, (r(x) - \varepsilon_n)(\lambda \ln n/(a_0 n))^{1/d})$ does not contain any observation. The Lipschitz hypothesis implies that $\mathcal{B}(x_i^*, r(x_i^*)(1 - K_r\varepsilon_n)(\lambda \ln n/(a_0 n))^{1/d})$ does not contain any observation neither.

On the other hand, for a given (deterministic) x:

$$P\left(\mathcal{B}\left(x, r(x)(1 - K_r\varepsilon_n)\left(\frac{\lambda \ln(n)}{a_0 n}\right)^{1/d}\right) \cap \mathcal{X}_n = \emptyset\right) \leq \left(1 - \lambda(1 + o(1))\frac{\ln n}{n}\right)^n.$$

So, the probability q_n that there exists a x_i^* such that $\mathcal{B}(x_i^*, r(x_i^*)(1 - K_r\varepsilon_n)(\lambda \ln n/(a_0 n))^{1/d})$ does not contains any observation satisfies

$$q_n = O\left(\frac{\varepsilon_n^{-d} n^{1-\lambda+o(1)}}{\ln n}\right).$$

Choosing $\varepsilon_n = \ln n^{-1/d}$ we obtain the desired. □

Lemma 9. *Let a_n be a sequence that satisfies $a_n > 0$ and $a_n \to a > 0$ and let $x > 0$ be some constant.*

Let us now define the sequences

$$u_k^n = C_n^k \left(a_n \frac{\ln n}{n}\right)^k \left(1 - a_n \frac{\ln n}{n}\right)^{n-k}, k_n^* = \lceil (1+x) a \ln n \rceil, \text{ and } v_n = \sum_{k=k_n^*}^{n} u_k^n.$$

Then

$$v_n \leq \frac{\sqrt{2}}{x\sqrt{\pi k_n^*}} n^{-a(1+x)\ln(1+x)-x}(1 + o(1)).$$

Proof. Let us first remark that:

$$u_{k+1}^n = \frac{n-k-1}{k+1}\left(a_n \frac{\ln n}{n}\right)\left(1 - a_n \frac{\ln n}{n}\right)^{-1} u_k^n.$$

and so:

$$u_{k+1}^n = \frac{n-k-1}{n}\left(\frac{a_n \ln n}{\lceil (1+x)a \ln n \rceil}\right)\left(\frac{k_n^*}{k+1}\right)\left(1 - a_n \frac{\ln n}{n}\right)^{-1} u_k^n$$

So it is possible to choose N_0 such that: for all $n \geq N_0$ and all $k \geq k_n^*$ we have

$$u_{k+1}^n < (1 - x/2)u_k^n.$$

So, for all $n \geq N_0$ we have $v_n \leq u_{k_n^*}^n (2/x)$.

Applying the Stirling formula we get

$$u_{k_n^*}^n \sim \frac{1}{\sqrt{2\pi k_n^*}} n^{-a_n[(1+x)\ln(1+x)-x]},$$

and so

$$v_n \leq \frac{\sqrt{2}}{x\sqrt{\pi k_n^*}} n^{-a[(1+x)\ln(1+x)-x]}(1 + o(1)).$$

The lemma is proved. □

Corollary 1. *Let $r(x)$ be a Lipschitz continuous function on S. Suppose that $\{X_1, \ldots, X_n\}$ is a sample on S drawn with some density such that there exists constants a_1 and b_1 satisfying for all $x \in S$ $P(X_i \in \mathcal{B}(x, r(x)\rho)) \leq a_1(1 + b_1\rho)\rho^d$.*

Then, for all $\lambda' > 0$ the following holds:

if $y > \varphi^{-1}((\lambda' a_1)^{-1})$, then "for all $x \in S$ the ball $\mathcal{B}(x, r(x)(\lambda' \ln n/n)^{1/d})$ contains less than $k_n = \lceil a_1(1+y)\lambda' \ln n \rceil$ points" a.a.s.

if $y > \varphi^{-1}(2(\lambda' a_1)^{-1})$, then "for all $x \in S$ the ball $\mathcal{B}(x, r(x)(\lambda' \ln n/n)^{1/d})$ contains less than $k_n = \lceil a_1(1+y)\lambda' \ln n \rceil$ points" e.a.s.

Proof. First, let us cover S with small deterministic balls of radius $\varepsilon_n(\lambda' \ln n/(n))^{1/d}$ centered in $x_1^*, \ldots, x_\nu^*$ ($\nu \leq c_0(S)\varepsilon_n^{-d} n/\ln n$).

Let us suppose that there exist an x such that $\mathcal{B}(x, r(x)(\lambda' \ln n/n)^{1/d})$ contains more than k observations. There exits an i such that $x \in \mathcal{B}(x_i^*, \varepsilon_n(\lambda' \ln n/n)^{1/d})$, and so $\mathcal{B}(x_i^*, (r(x) + \varepsilon_n)(\lambda \ln n/(n))^{1/d})$ contains more than k observations. The Lipschitz hypothesis implies that $\mathcal{B}(x_i^*, r(x_i^*)(1 + K_r\varepsilon_n)(\lambda \ln n/(n))^{1/d}))$ also contains more than k observations.

On the other hand, for a given (deterministic) x let us denote $q_n(x)$ the probability that $\mathcal{B}(x, r(x)(1 + K_r\varepsilon_n)(\lambda \ln n/(n))^{1/d})$ contains more than $k_n = \lceil a_1\lambda'(1 + y) \ln n\rceil$ observations. Applying Lemma 9 we get:

$$q_n(x) = O\left(\frac{\sqrt{2}}{y\sqrt{\pi k_n}} n^{-a_1\lambda'[(1+y)\ln(1+y)-y]}\right).$$

So, the probability q_n that there exists an x_i^* such that $\mathcal{B}(x_i^*, r(x_i^*)(1 + K_r\varepsilon_n)(\lambda \ln n/(n))^{1/d})$ contains more than k_n observations satisfies

$$q_n = O\left(\frac{\varepsilon_n^{-d}}{y \ln n^{3/2}} n^{1-a_1\lambda'[(1+y)\ln(1+y)-y]}\right),$$

that is,

$$q_n = O\left(\frac{\varepsilon_n^{-d}}{y \ln n^{3/2}} n^{1-a_1\varphi(y)}\right)$$

choosing $\varepsilon_n = \ln n^{-3/2d}$ we obtain the desired. □

Lemma 10. *Let a_n be a sequence that satisfies $a_n > 0$ and $a_n \to a > 0$ and let $x > 0$ be a constant.*

Let us now define the sequences

$$u_k^n = C_n^k \left(a_n(1 + x)\frac{\ln n}{n}\right)^k \left(1 - a_n(1 + x)\frac{\ln n}{n}\right)^{n-k}, k_n^* = \lceil a \ln n\rceil \text{ and } v_n = \sum_{k=0}^{k_n^*} u_k^n.$$

Then

$$v_n \leq \frac{\sqrt{2}}{x\sqrt{\pi k_n^*}} n^{a(\ln(1+x)-x+(1-a_n/a)(1+x)+\ln(an/a))}(1 + o(1)),$$

that is

$$v_n \leq \frac{\sqrt{2}}{x\sqrt{\pi k_n^*}} n^{-a\psi(x)+o(1)}.$$

The proof is very similar to that of Lemma 9 and is omitted.

Corollary 2. *Let $r(x)$ be a Lipschitz continuous function defined on S. Suppose that $\{X_1, \ldots, X_n\}$ is a sample on S drawn with some density such that there exists constants a_0 and b_0 satisfying for all $x \in S$ we have $P(X_i \in \mathcal{B}(x, r(x)\rho)) \geq a_0(1 - b_0\rho)\rho^d$.*

Then, for all $\lambda' > 0$, the following holds

if $y > \psi^{-1}((\lambda' a_1)^{-1})$ *then : "for all* $x \in S$ *the ball* $\mathcal{B}(x, r(x)(\lambda' \ln n/n)^{1/d})$ *contains less than* $k_n = \lceil a_0 \lambda' \ln n \rceil$ *points" a.a.s.*

if $y > \psi^{-1}(2(\lambda' a_1)^{-1})$ *then : "for all* $x \in S$ *the ball* $\mathcal{B}(x, r(x)(\lambda' \ln n/n)^{1/d})$ *contains less than* $k_n = \lceil a_0 \lambda' \ln n \rceil$ *points" e.a.s.*

The proof is very similar of that of Corollary 1 an is omitted.

Lemma 11. *Let* a_n *be some sequence and* $\rho > 1$ *be some constant. We put* $p_n = \left(a_n \frac{\ln n}{n}\right)^{\rho}$.

Let us now define the sequences

$$u_k^n = C_n^k (p_n)^k (1 - p_n)^{n-k} \quad \textit{and } v_n = \sum_{k=l}^{n} u_k^n.$$

Then

$$v_n \leq a_n^l \frac{1}{l!} \frac{(\ln n)^{\rho l}}{n^{(\rho-1)l}} e^{a_n \ln n / n^{\rho-1}}.$$

Proof. First let us note that

$$u_k^n \leq \frac{1}{k!} \left(\frac{a_n \ln n^{\rho}}{n^{\rho-1}} \right)^k.$$

Now it is sufficient to use the inequality

$$\sum_{k=l}^{\infty} \frac{x^k}{k!} \leq \frac{x^l}{l!} e^x$$

in order to conclude the proof of the Lemma. □

B Proof of Property 1

Let $r(x)$ be a Lipschitz continuous function on S.

Suppose that for every deterministic ball centers in $x \in S$ and with a small enough radius ρ there exists constants a_0, b_0, a_1 and b_1 such that

$$a_0 \rho^d (1 - b_0 \rho) \leq P(X_i \in \mathcal{B}(x, r(x)\rho)) \leq a_1 \rho^d (1 + b_1 \rho).$$

Then:

i) Let $\lambda > 1$ (resp. $\lambda > 2$). Then, Lemma 8 implies that for every edge $[X_i, X_j]$ belonging to $D_{(S)}$ we have $||\overrightarrow{X_i X_j}|| < 2(\lambda \ln n/(a_0 n))^{1/d})$ a.a.s. (resp. e.a.s.).

Proof. For every edge $[X_i, X_j]$ belonging to $D_{(S)}$, there exists a point O in $\mathrm{Vor}(X_i) \cap \mathrm{Vor}(X_j) \cap S$ such that $\mathcal{B}(O, ||\overrightarrow{OX_i}||)$ does not contain any observation. Lemma 8 implies that $||\overrightarrow{OX_i}|| \leq (\lambda \ln n/(a_0 n))^{1/d}$ a.a.s. (resp. e.a.s.). Of course, the same argument gives $||\overrightarrow{OX_j}|| \leq (\lambda \ln n/(a_0 n))^{1/d}$ a.a.s. (resp. e.a.s.). So $||\overrightarrow{X_i X_j}|| < 2(\lambda \ln n/(a_0 n))^{1/d}$ a.a.s. (resp. e.a.s.). □

ii) Corollary 2 implies that for all $y > \varphi^{-1}(2^{-d} a_0/(\lambda a_1))$ (resp. $y > \varphi^{-1}(2^{-d+1} a_0/(\lambda a_1))$, every edge $[X_i, X_j]$ belonging to $D_{(S)}$ satisfy: $\mathcal{B}(X_i, ||\overrightarrow{X_i X_j}||)$ contains less than $\lceil 2^d \lambda \frac{a_1}{a_0}(1+y)\ln n \rceil$ points.

To conclude the proof, it remains to apply the above with $r(x)$, a_0, a_1, b_0 and b_1 chosen according to the following table.

hypotheses	$r(x)$	a_0	a_1	b_0	b_0
H1, **H2** and **H3**	1	$\alpha\theta_d f_0$	$\theta_d f_1$	0	0
H1, **H2**, **H3** and **H4**	1	$0.5\theta_d f_0$	$\theta_d f_1$	$c_b(S)$	0
H1, **H2**, **H3** and **H5**	$f(x)^{-1/d}$	$\alpha\theta_d$	θ_d	$K_{f^{d-1}}$	$K_{f^{d-1}}$
H1, **H2**, **H3**, **H4** and **H3**	$f(x)^{-1/d}$	$0.5\theta_d$	θ_d	$K_{f^{d-1}} + c_b(S)$	$K_{f^{d-1}}$

C Proof of Theorem 3

Theorem 3 is a direct corollary of the two following lemmas.

Lemma 12. *Let us denote:* $d_H^*(A, B) = \inf\{r, A \subset B + r\overline{\mathcal{B}}\}$. *If* $y > \psi^{-1}(2^{-d})$ *(with* $\psi(y) = \ln(1+y) - y$*) and* $a \in \{\tilde{a}(f, S), \tilde{a}(., S), \tilde{a}(f, .), \tilde{a}\}$ *,we have*

$$d_H^*(D_{\lceil a \ln n \rceil}, S) \leq \left(\frac{a(1+y)\ln n}{\theta_d \alpha f_0 n} \right)^{1/d} \quad e.a.s.$$

Proof. Let us suppose that $D_{k_n} \not\subseteq S + r_n \overline{\mathcal{B}}$. Then there exists a simplex σ of D_{k_n} (to simplify notations let $\sigma = (X_1, \ldots, X_{d+1})$) and $x \in \sigma$ with $d_{\min}(x, S) > r_n$, and so $||\overrightarrow{xX_i}|| > r_n$ for all $i \in \{1, \ldots, d+1\}$.

As $||\overrightarrow{xX_1}|| \leq \max_{i,j \in \{1,\ldots,d+1\}} ||\overrightarrow{X_i X_j}||$, there exists i and j in $\{1, \ldots, d+1\}$ such that $||\overrightarrow{X_i X_j}|| > r_n$ and and X_i and X_j are k_n nearest neighbor one of each other.

So, there exists a point $x \in S$ (X_i for instance) such that $\mathcal{B}(x, r_n)$ contains less than k_n points. However, Corollary 2 ensures that such a situation can not occur e.a.s. □

Lemma 13.

$$d_H^* \left(S, D_{\lceil a \ln n \rceil}\right) \left(\frac{n}{\ln n}\right)^{1/d} \text{ is bounded e.a.s.}$$

The bound is:

- $\max\left(\theta_d^{-1} f_0^{-1}, \theta_d^{-1} 2(1-1/d)(\min_{\partial S} f)^{-1}\right)^{1/d}$ *when* ∂S *is smooth;*
- $(b/(f_0 \theta_d \alpha))^{1/d}$ *for every* $b > 2$ *in the general case.*

Proof. Let us suppose that x satisfies $x \in S$ and $x \notin D_{k_n} + r\overline{\mathcal{B}}$. Then $\min_i ||\overrightarrow{xX_i}|| > r$ and so $\mathcal{B}(x,r) \cap \mathcal{X}_n = \emptyset$.

Let us note that in [27] Penroses already proved the result for the case when the boundary of S is $\mathcal{C}^2$:

$$d_H^* \left(S, D_{\lceil a \ln n \rceil}\right) \leq \left(\max\left(f_0^{-1}, 2(1-1/d)(\min_{\partial S} f)^{-1}\right)^{1/d} \frac{\ln n}{n\theta_d}\right)^{1/d} \text{ e.a.s.}$$

Lemma 8 gives a weaker version with a higher constant and the same rate for all $b > 2$

$$d_H^* \left(S, D_{\lceil a \ln n \rceil}\right) \leq \left(\frac{b \ln n}{n f_0 \alpha \theta_d}\right)^{1/d} \text{ e.a.s}$$

The Lemma is proved. □

D Proofs of the results for the case $d = 2$

In this section we suppose that $d = 2$ and that ∂S is $\mathcal{C}^2$.

D.1 The Closed Ball Property

The specificity of the dimension 2 is that the different cases contradicting the closed ball property can be easily enumerated. The four possible cases are:

i) there exits X_i and X_j such that $\text{Vor}(X_i) \cap \text{Vor}(X_j) \cap \partial S \neq \emptyset$ and $\text{Vor}(X_i) \cap \text{Vor}(X_j) \cap \partial S \not\cong \overline{\mathcal{B}}_0$;

ii) there exits X_i and X_j such that $\text{Vor}(X_i) \cap \text{Vor}(X_j) \cap S \neq \emptyset$ and $\text{Vor}(X_i) \cap \text{Vor}(X_j) \cap S \not\cong \overline{\mathcal{B}}_1$;

iii) there exits X_i such that $\text{Vor}(X_i) \cap \partial S \neq \emptyset$ and $\text{Vor}(X_i) \cap \partial S \not\cong \overline{\mathcal{B}}_1$;

iv) there exits X_i such that $\text{Vor}(X_i) \cap S \neq \emptyset$ and $\text{Vor}(X_i) \cap S \not\cong \overline{\mathcal{B}}_2$.

Firs we prove that the situation iv) can not happen e.a.s.

Proof. Let us denote $r(x) = \sup_r\{\text{for all } U \cong \overline{\mathcal{B}}_2 \text{ such that } x \in U \text{ and } U \subset \mathcal{B}(x,r), \text{ we have } U \cap S \cong \mathcal{B}_2 \text{ or } U \cap S \cong \overline{\mathcal{B}}_2\}$. Let us now denote $r_0 = \inf_x r(x)$. We will first proof that $r_0 > 0$. Let us suppose the contrary. There exist a sequence $r_n \to 0$, a sequence of points x_n and a sequence of sets U_n ($x_n \in U_n$ and $U_n \subset \mathcal{B}(x_n, r_n)$) with $U_n \cap S \ncong \mathcal{B}_2$ and $U_n \cap S \ncong \overline{\mathcal{B}}_2$. The compactness of S allow to exhibit a subsequence of x_n that converges toward $x \in S$. Our supposition contradict the fact that x admit a neighborhood homeomorph to a ball (closed or open) that is that S is a manifold. The fact that for all X_i we have $\text{Vor}(X_i) \cap S \subset \mathcal{B}(X_i, (3/f_0\alpha\theta_d)(\ln n/n)^{1/2})$ e.a.s. concludes the proof. □

Remark: The proof for the case iv) is also valid in other dimensions.

Let us now suppose that iv) does not happen, the three other cases cases can be enumerated in a different way First we can note that i) or ii) implies that there exists $[x_1, x_2] = \text{Vor}(X_i) \cap \text{Vor}(X_j)$ and y_1, y_2, $y_1 \neq y_2$ two points in $[x_1, x_2] \cap \partial S$

a) is i) or ii) with $x_1 \in S$ or $x_2 \in S$.

b) i) or ii) with $x_1 \notin S$ and $x_2 \notin S$.

c) iii) but neither i) nor ii) and $\partial\,\text{Vor}(X_i) \cap \partial S \neq \emptyset$

d) iii) but neither a) nor b) nor c) and $\partial\,\text{Vor}(X_i) \cap \partial S = \emptyset$

Figure 17 illustrates this four possible cases

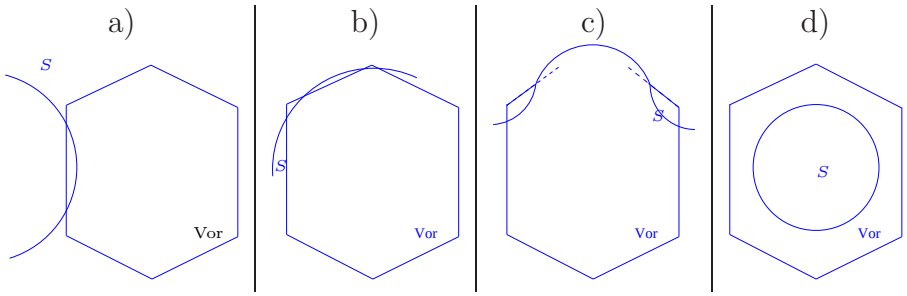


Figure 17: The different cases that contradict the closed ball property, the hexagon represents the Voronoi cell and the curve the boundary of S.

The idea of the proof is the following:

- On one hand, a), b) and c) implie the existence of constants $A > 0$ and $B > 0$ and of a point $x^* \in \partial S + A(\ln n/n)\overline{\mathcal{B}}$ such that $\mathcal{B}(x^*, B \ln n/n)$ contains at least 2 observations. Lemmas 3 and 11 allows thus to conclude that this situations can not occur a.a.s.

- On the other hand, d) can not occur a.a.s. because of the manifold hypothesis on ∂S and the fact that the Voronoi cell size converges to 0 (with a reasonment very similar to the proof that iv) can not occur (a.a.s).

Let us first prove that a) implies the existence of constants $A_1 > 0$ and $B_1 > 0$ and of a point $x^* \in \partial S + A_1(\ln n/n)\overline{\mathcal{B}}$ such that $\mathcal{B}(x^*, B_1 \ln n/n)$ contains at least 2 observations.

Proof. Suppose that the situation a) occurs with $x_1 \in S$.

According to the Rolle theorem there exists a point O in ∂S such that the tangent Δ of S at the point O is parallel to (x_1, x_2).

There exists O_r with $||\overrightarrow{OO_r}|| = r_S$, such that $\mathcal{B}(O_r, r_S) \cap S = \emptyset$ (Lemma 7).

Let us note that $[x_1, x_2] = \text{Vor}(X_i) \cap \text{Vor}(X_j)$ implies the existence of X_i and X_j with a midpoint x^* belonging to (x_1, x_2) and that are in a ball $\mathcal{B}(O, a^*(\ln n/n)^{1/2})$ e.a.s. ($a^* = 3/f_0\alpha\theta_d$ can be used according to Lemma 8).

Looking at Figure 18 it becomes clear that

- $d(x^*, \partial S) \leq d_{min}(z, O_r) - r_S \leq (a^{*2}/2r_S) \ln n/n$ e.a.s.;
- $||\overrightarrow{X_i x^*}|| \leq ||z' - y^*|| \leq ||\overrightarrow{y^* z}|| \sim (a^{*2}/2r_S) \ln n/n$ e.a.s.

□

Let us now quickly note that: b) and c) also imply the existence of constants $A_2 > 0$ and $B_2 > 0$ and of a point $x^* \in \partial S + A_2 \ln n/n\overline{\mathcal{B}}$ such that $\mathcal{B}(x^*, B_2 \ln n/n)$ contains at least 2 observations. See Figure 19 to be convinced that we are in a classical linear interpolation case and so Lemma 6 is sufficient to conclude.

We now use Lemmas 3 and 11 to prove that for all $x \in \partial S + A \ln n/n\overline{\mathcal{B}}$, the ball $\mathcal{B}(x^*, B \ln n/n)$ contains at most 1 observation a.a.s.

Proof. Let us first cover $\partial S + A \ln n/n\overline{\mathcal{B}}$ with deterministic balls of radius $\varepsilon_n \ln n/n$ centered in $x_1, \ldots, x_\nu$ with $\nu \leq c_1(\partial S) A \varepsilon_n^{-2}(n/\ln n)$ as in Lemma 3.

Let us now suppose that there exists $x^* \in \partial S + A \ln n/n\overline{\mathcal{B}}$, such that $\mathcal{B}(x^*, B \ln n/n)$ contains at least 2 observations. There exists an x_i such that $x^* \in \mathcal{B}(x_i, \varepsilon_n \ln n/n)$, and so $\mathcal{B}(x_i, (\varepsilon_n + B) \ln n/n)$ contains at least 2 points.

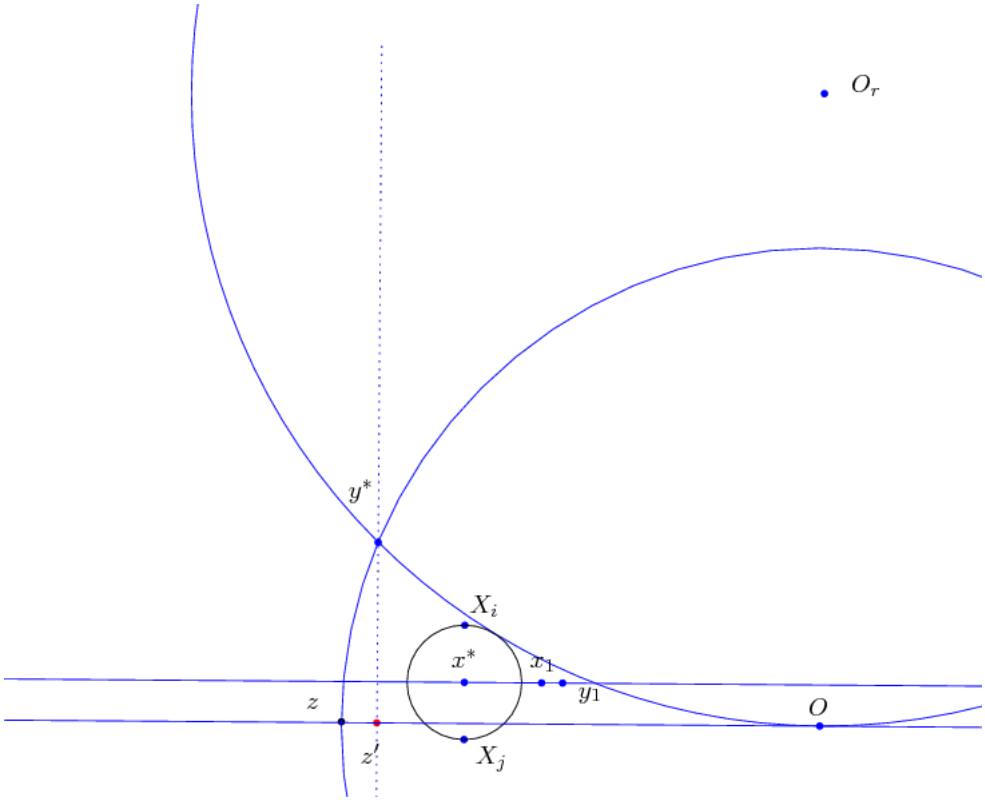
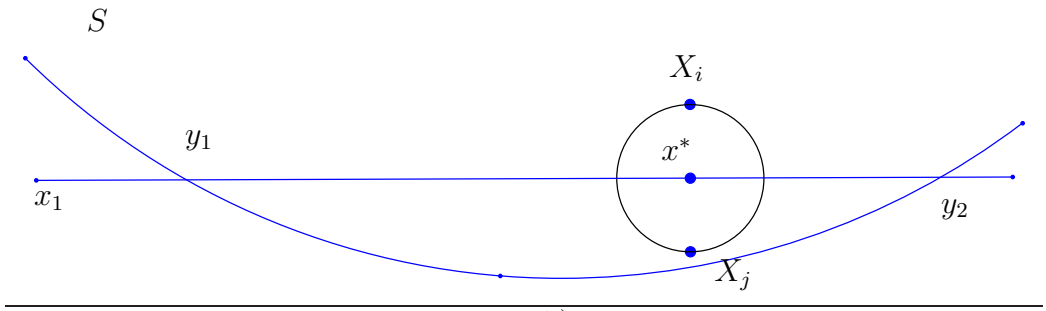
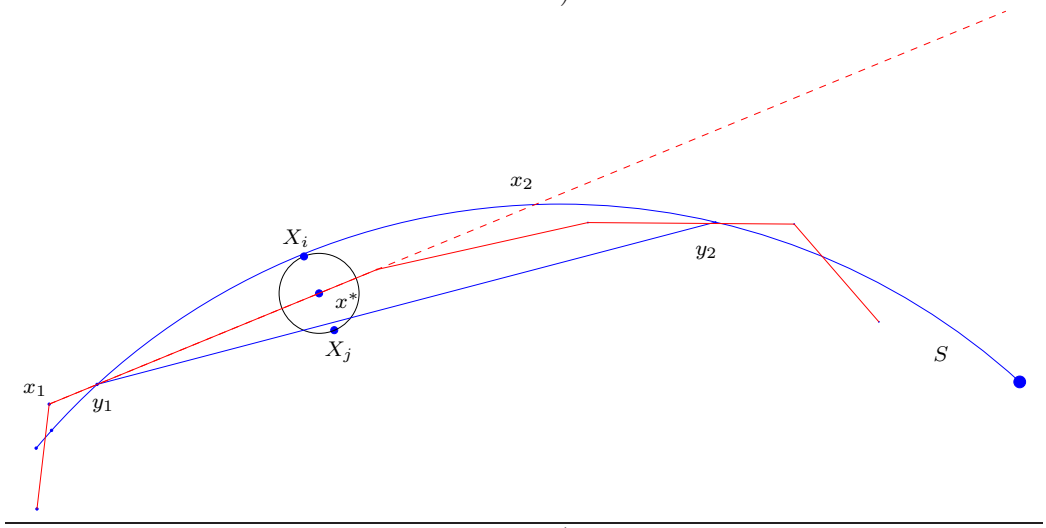


Figure 18: Construction for case a).



case b)



case c)

Figure 19: Construction for the cases b) and c).

On the other hand, Lemma 11 ensures that for a deterministic x the probability $q_n(x)$ that $\mathcal{B}(x, (\varepsilon_n + B)\ln n/n)$ contains at least 2 points satisfies: $q_n(x) = O\left(\frac{\ln n^4}{n^2}\right)$. So, if q_n denotes the probability that there exists an x_i such that $\mathcal{B}(x_i, (\varepsilon_n + B)\ln n/n)$ contains at least 2 points, $q_n(x) = O\left(\varepsilon_n^{-2}\frac{\ln n^3}{n}\right)$. Choosing, for instance, $\varepsilon = \ln n$ allows to conclude the proof. □

D.2 Proof of Theorem 2

Let us first remark that all X_i belong to $D_{(S)}$ (and so also to D_{k_n} e.a.s.).

Let us suppose that $D_{(S)} \subset D_{k_n}$, that $\mathcal{X}_n$ has the closed ball property on S, and that $D_{k_n} \subset S + \varepsilon_n\overline{\mathcal{B}}$ with $S + \varepsilon_n\overline{\mathcal{B}} \cong S \cong D_{(S)}$ (which is a.a.s. true).

There is three different ways to have $D_{k_n} \not\cong D_{(S)}$:

i) D_{k_n} "connects" different components of $D_{(S)}$;

ii) D_{k_n} "fills a hole" of $D_{(S)}$;

iii) D_{k_n} "adds a hole" to $D_{(S)}$.

These 3 different phenomena are illustrated in Figure 20.

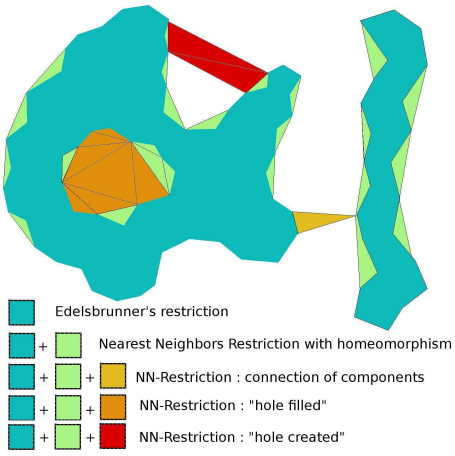


Figure 20: Different ways to have $D_{k_n} \not\cong D_{(S)}$

Let us first note that the two first points (i and ii) contradict the fact that $D_{k_n} \subset S + \varepsilon_n\overline{\mathcal{B}}$ with $S + \varepsilon_n\overline{\mathcal{B}} \cong S \cong D_{(S)}$.

We can thus focus on the last point (iii).

If there exists an unexpected "hole" in D_{k_n} , then there exists a simplex $\sigma = (X_1, X_2, X_4)$ that satisfies: $\sigma \in D_{k_n}$, $\sigma \notin D_{(S)}$ and a simplex $\sigma' = (X_1, X_2, X_3)$ that satisfies: $\sigma \notin D_{k_n}$, $\sigma \notin D_{(S)}$ (as in Figure 21).

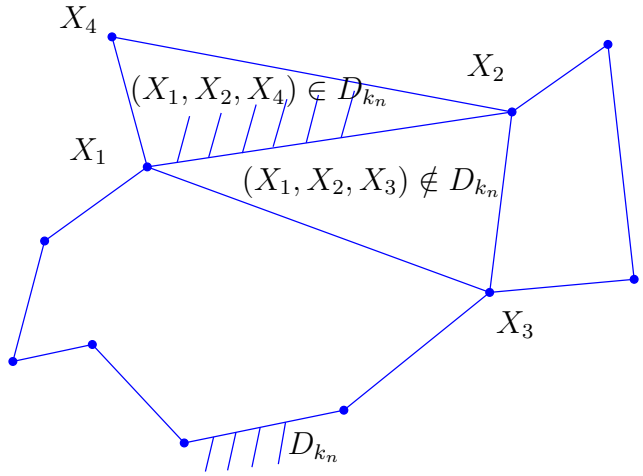


Figure 21: existence of a hole

Let us now focus on the two simplexes σ and σ'. The center O (resp. O') of the circumscribed circles of σ (resp σ') belongs to Δ (the mid-perpendicular of $[X_1, X_2]$) and is not in S. In this section we will denote $\mathcal{C}$ (resp. $\mathcal{C}'$ resp. $\mathcal{C}''$) the circle circumscribed to σ (resp. circumscribed to σ' resp. of radius $||\overrightarrow{X_1X_2}||$ and centered in X_1) and $\mathcal{B}$ (resp. $\mathcal{B}'$ resp. $\mathcal{B}''$) is the associated balls.

Let us first distinguish two cases :

a) O and O' "are on the same side" ($\overrightarrow{MO}.\overrightarrow{MO'} \geq 0$) where M is the middle point of $[X_1, X_2]$.

b) O and O' "are on opposite side" ($\overrightarrow{MO}.\overrightarrow{MO'} < 0$) where M is the middle point of $[X_1, X_2]$.

It quickly becomes clear that the case a) is not possible. The simplex (X_1, X_2, X_3) belongs to the Delaunay complex but not to the Delaunay complex restricted to D_{k_n}, so X_3 is not a k_n-nearest neighbor of X_1 or not a k_n-nearest neighbor of X_2. Let us suppose that X_3 is not a k_n-nearest neighbor of X_1. It is easy to see that $X_3 \in (\mathcal{C}' \setminus \mathcal{B}) \setminus \mathcal{B}''$ (because of the properties of the Delaunay complex and because X_3 is not a k_n-nearest neighbor of X_1, while X_2 is a nearest neighbor of X_1). Case a) implies that

$(\mathcal{C}' \setminus \mathcal{B}) \setminus \mathcal{B}'' = \emptyset$ (see Figure 22).

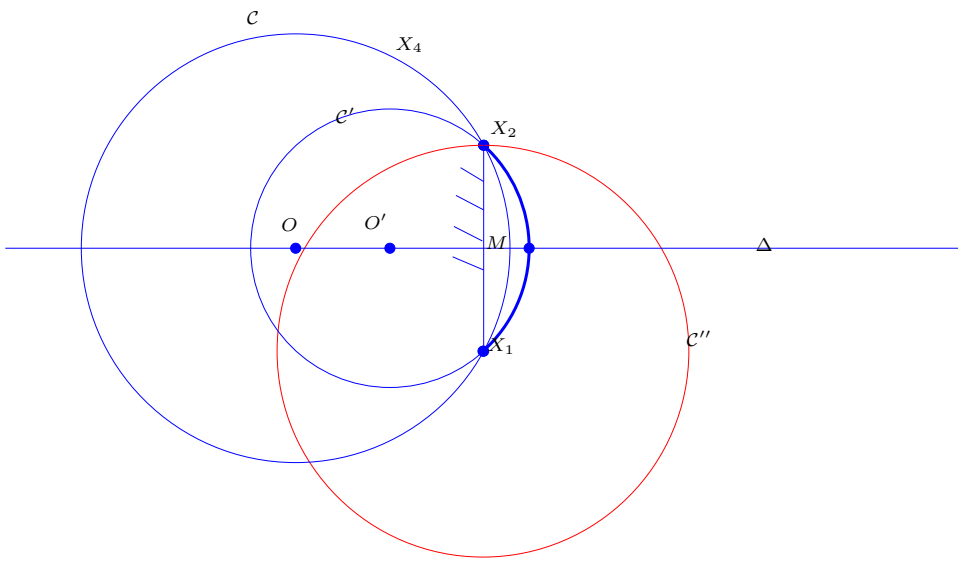


Figure 22: case a) (impossible a.a.s.)

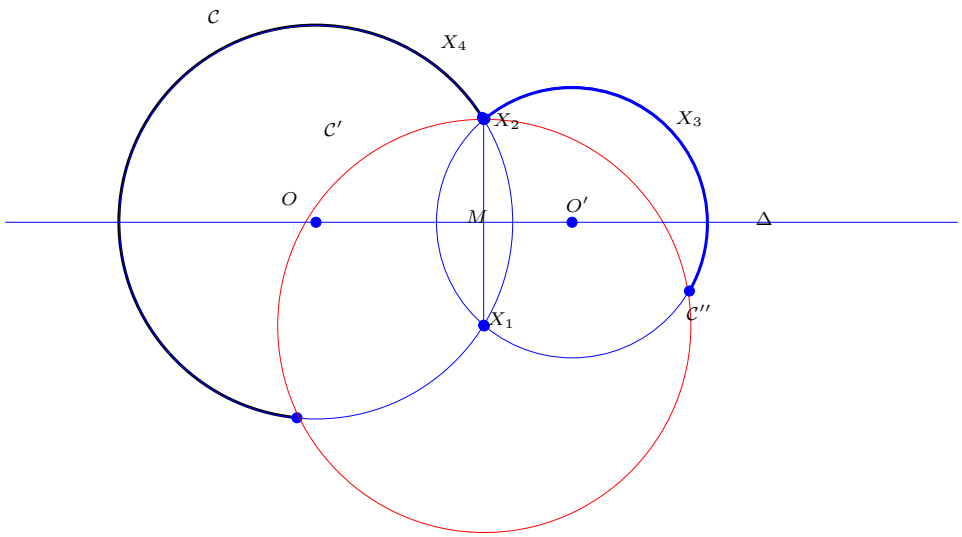


Figure 23: case b

Let us now remark that $\partial S \cap \Delta \cap \mathcal{B}'' \neq \emptyset$ e.a.s.

Proof. When n is large enough, all $\mathcal{B}(X_i, ||\overrightarrow{X_jX_i}||)$ where X_j is at most a k_n-nearest-neighbors of X_i satisfy : $\overline{\mathcal{B}}(X_i, ||\overrightarrow{X_jX_i}||) \cap S$ is homeomorph to a close ball. This implies that $\mathcal{B}'' \cap S$ is path connected, so there exists a continuous path in $\mathcal{B}'' \cap S$ that links X_1 to X_2 and so crosses Δ. So : $S \cap \Delta \cap \mathcal{B}'' \neq \emptyset$. As O and O' are in Δ and not in S we have $S^c \cap \Delta \cap \mathcal{B}'' \neq \emptyset$, and so $\partial S \cap \Delta \cap \mathcal{B}'' \neq \emptyset$. □

Let us first consider the case: $\partial S \cap \Delta \cap \mathcal{B}'' = \{x_0\}$.

The first sub-case is: $\overrightarrow{MO'}.\overrightarrow{Mx_0} > 0$. The point M does not belong to S (otherwise there exists another point in $\partial S \cap \Delta \cap \mathcal{B}''$). We can define two points $x_0^+ = \text{argmin}\{||\overrightarrow{Mx}||, x \in S, \overrightarrow{Mx} = t\overrightarrow{MX_2}, t > 0\}$ and $x_0^- = \text{argmin}\{||\overrightarrow{Mx}||, x \in S, \overrightarrow{Mx} = t\overrightarrow{MX_2}, t < 0\}$ both belonging to ∂S.

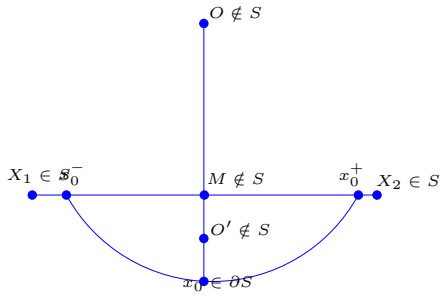


Figure 24: $\partial S \cap \Delta \cap \mathcal{B}'' = \{x_0\}$ and the intersection "is under O'".

By Lemma 6 $||\overrightarrow{Mx_0}|| \leq c||\overrightarrow{x_0^+x_0^-}||^2$. So, $||\overrightarrow{Mx_0}|| \leq c||\overrightarrow{X_1X_2}||^2$. Let us denote α the angle $\widehat{O'X_1M}$. $\tan(\alpha) = ||\overrightarrow{O'M}||/||\overrightarrow{MX_1}|| \leq 2c||\overrightarrow{X_1X_2}||$. Let us denote x^* the intersection of $\mathcal{C}''$ and (X_1, O') and $\rho = ||\overrightarrow{x^*X_2}||$. See Figure 25 for the construction. First, $\rho = \sin(\alpha)||\overrightarrow{X1X_2}|| \leq 2c||\overrightarrow{X_1X_2}||^2$. Second, $\mathcal{C}' \setminus \mathcal{B}'' \subset \mathcal{B}(x^*, \rho)$. Third, $||\overrightarrow{x^*x_0}|| \leq ||\overrightarrow{X_1X_2}||$. This three points imply a.a.s. the existence of a point x^*, and of constants A and B such that: $d_{min}(x^*, \partial S) \leq A(ln(n)/n)^{1/2}$ and there is at least two points in $\mathcal{B}(x^*, B\ln n/n)$. An argument similar to the one used in the previous section allows to conclude that it is impossible a.a.s. (the probability of occurrence is a $O(\varepsilon_n^{-2}\frac{\ln n^{5/2}}{n^{1/2}})$ for any $\varepsilon_n \to 0$, and so we can choose $\varepsilon_n = \ln n$, for instance, to conclude).

Let us now consider the second sub-case: $\overrightarrow{MO'}.\overrightarrow{Mx_0} < 0$. See Figure 26 to be convinced that we are back in an interpolation problem and that $d_{min}(X_2, \partial S) \leq A\ln n/n$ and $d||\overrightarrow{X_2X_3}|| \leq B\ln n/n$ which is impossible a.a.s. according to Lemma 6.

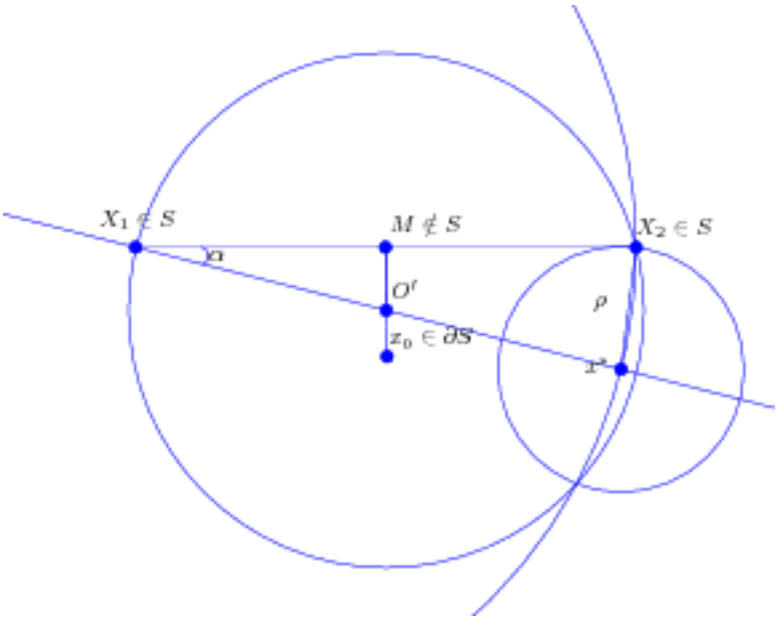


Figure 25: conclusion for $\partial S \cap \Delta \cap \mathcal{B}'' = \{x_0\}$ and the intersection "is under O'".

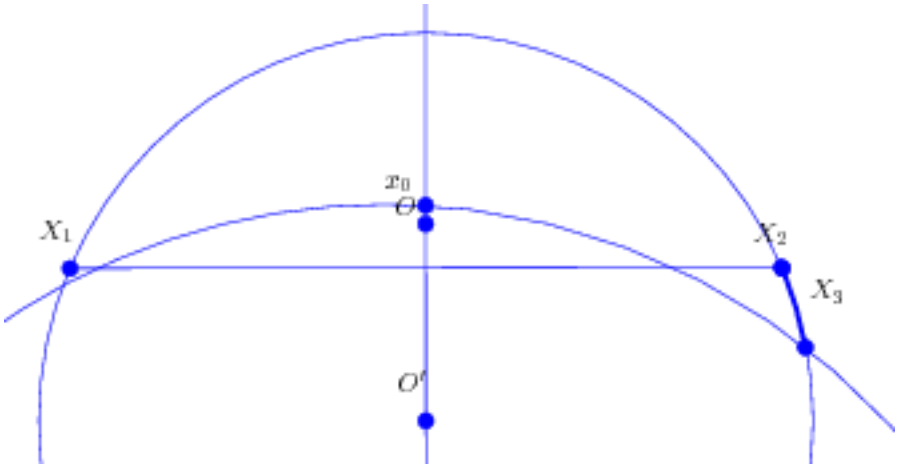


Figure 26: $\partial S \cap \Delta \cap \mathcal{B}'' = \{x_0\}$ and the intersection "is over O".

Let us now suppose that $\partial S \cap \Delta \cap \mathcal{B}''$ contains at least two points x_0 and x_1. This case is easy: let us denote $\rho = ||\overrightarrow{X_1X_2}|| > 0$. Lemma 6 ensures that $||\overrightarrow{x_0x_1}|| \leq A\rho^2$ and a new application of Lemma 6 ensures that $\rho \leq A^4\rho^4$ that is impossible a.a.s.

D.3 Discussion when $d' < d$

The proofs of Property 3 and of Theorem 4 in the case $d' < d$ are exactly the same than those of Property 1 and of Theorem 3. The only difference is the use of new inequalities for the probability to have an observation that falls into balls given by Lemmas 14, 15 and 16.

Lemma 14. *If S is a $\mathcal{C}^2$ manifold there exists constants $r_0 > 0$ and B_S^+ such that for all $r \leq r_0$: $\mu_S(\mathcal{B}(x,r) \cap S) \leq \theta_{d'} r^{d'}(1 + B_S^+ r)$*

Lemma 15. *If S is a $\mathcal{C}^2$ manifold with a $\mathcal{C}^2$ boundary there exists constants $r_0 > 0$, B_S^- and B_S^+ such that for all $r \leq r_0$: $\theta_{d'} r^{d'}(0.5 + B_S^- r) \leq \mu_S(\mathcal{B}(x,r) \cap S) \leq \theta_{d'} r^{d'}(1 + B_S^+ r)$*

Lemma 16. *If S is a $\mathcal{C}^2$ manifold without boundary there exists constants $r_0 > 0$, B_S^- and B_S^+ such that for all $r \leq r_0$: $\theta_{d'} r^{d'}(1 + B_S^- r) \leq \mu_S(\mathcal{B}(x,r) \cap S) \leq \theta_{d'} r^{d'}(1 + B_S^+ r)$*

References

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

[2] I. Bárány. Random polytopes in smooth convex bodie. *Mathematika*, 39:81–92, 1982.

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

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

[5] G. Biau, F. Chazal, D. Cohen-Steiner, L. Devroye, and C. Rodríguez. A weighted k-nearest neighbor density estimate for geometric inference. *electronic journal of statistics*, 5:204–237, 2011.

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

[7] G.e. Bredon. *Topology and geometry*. Graduate Texts in Mathematics, 1991.

[8] B. Cadre. Kernel estimation of density level sets. *Journal of Multivariate Analysis*, pages 999–1023, 2006.

[9] 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.

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

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

[12] F. Chazal, D. Cohen-Steiner, and A. Lieutier. Normal cone approximation and offset shape isotopy. *Computational Geometry : Theory and Applications*, 42:566–581, 2009.

[13] F. Chazal, D. Cohen-Steiner, and A. Lieutier. A sampling theory for compact sets in euclidean spaces. *Discrete and Computational Geometry*, 41:461–479, 2009.

[14] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for probability measures,. *Journal on Foundations of computational Mathematics*, 11:733–751, 2011.

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

[16] M.K. Chung, P. Bubenik, and P.T. Kim. *Information Processing in Medical Imaging 2009*, chapter Persistence Diagrams of Cortical Surface Data., pages 386–397. Lecture Notes in Computer Science, 2009.

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

[18] A. Cuevas and A. Rodríguez-Casal. On boundary estimation. *Advanced in Applied Probability*, 36:340–354, 2004.

[19] L. Devroye and G.L. Wise. Detection of abnormal behavior via nonparametric estimation of the support. *SIAM Journal of Applied Mathematics*, 38:480–488, 1980.

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

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

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

[23] H. Federer. *Geometric Measure Theory*. 1996.

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

[25] J.A. Lee, A. Lendasse, and M. Verleysen. Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis. *Neurocomputing*, 57:49–76, 2004.

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

[27] M.D. Penrose. A strong law for the largest nearest-neighbour link between random points. *Journal of the London Mathematical Society*, 60:951–960, 1999.

[28] M. Reitzner. Random polytopes and the efronstein jackknife inequality,. *Ann. Probab.,*, 31:21362166., 2003.

[29] C. Schütt. Random polytopes and affine surface area. *Math. Nachr.*, 170:227–249, 1994.

[30] J.B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

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