# On the maximal multivariate spacing extension and convexity tests

Catherine Aaron, Cholaquidis Alejandro, Fraiman Ricardo

# A new test of the convexity of the density support

Catherine Aaron

May 29, 2012

**Abstract**

Given a sample of $n$ independant and identically distributed random vectors drawn from a density $f$ we study two ways of testing the convexity of the support of the density. The first one requires the hypothesis that $f$ is uniform on its support. Two statitistics are proposed, one to test non-convexity due to the boundary, the other one to detect non convexity due to the existance of a hole. p-values for each test can be bounded above but this upper bound depends on unknown parameters. An estimator for this upper bound is given, and it is proved to be "almost surely similar" (with a given convergence speed). It is also proved that there exists a consistent decision rule associated to this test. When the density is unknown the test is adapted via a density estimation with $k-$nearest neighbors. Other similar results are given. They are obviously weaker (convergence speed can not be given) but still give good results.

**key words:** density-support, convexity, nearest-neighbors.

# 1 Introduction, notation and hypotheses

## 1.1 introduction

Let $\mathcal{X}_n$ be a random sample in $\mathbb{R}^d$ drawn from an unknown density $f$. The aim of this paper is to test wether $S$ the support of the density is convex or not. Such a test has various possible applications. Obviously it can help to choose a dimension reduction method, since the density support is convex $PCA$ can be considered as the best way to solve the problem: it is the easiest and it is well adapted. Suppose we are interested in the following regression model: $Y = \phi(X) + \varepsilon$, $X$ a random sample in $\mathbb{R}^d$ that has a

convex support, $\varepsilon \in \mathbb{R}$ independant to $X$ and $\varepsilon$ having a density $f_\varepsilon$ such that: $\{f_\varepsilon > 0\}$ is an intervall. A test of convexity allows one to verify the hypothesis on $X$ and the convexity of $(X, \varepsilon)$ indicates that a linear model can solve the problem. Hernández, Delicado and Lugosi also mentioned in [10] an application to the parametrization of $\varepsilon-$isomap [5] method. This way to use the convexity test can be applied to every statistical method that requires the use of geodesic distance: usually the geodesic distance is computed via a local graph weighted by the euclidian distance. If the graph links two points whenever they are closer than $\varepsilon$ (as in $\varepsilon-$isomap) the $\varepsilon$ parameter has to be carefully chosen:

- small $\varepsilon$ may disconnect the graph, inducing a huge overestimation of the geodesic distance and a long computational time for the geodesic distance.
- large $\varepsilon$ may estimate the euclidian distance and not the geodesic one.

In [9] the authors proposed an a posteriori way to choose $\varepsilon$. But we think a convexity test can also provide an a priori way to choose.

In spite of all these possible applications iwe have only found two papers that mention a test for convexity.

- In [9] the idea is (very briefly) proposed to test the convexity by using as test statistic the measure of the symetric difference between an estimator of the density support (using the estimator originally introduced by Devroye and Wise in [8]) and the convex hull of the sample. According to [4] (for the latest results on the asymptotic behavior of the estimation of the density support) and [3], [14] and [15] (for asymptotic results on the convex hull), under the null hypothesis (the support is convex) this volume converges toward 0 with a known speed.
- In [10] Hernández, Delicado and Lugosi build a test based on:
  $S = \frac{1}{n(n-1)} \sum_{i<j} ||X_{i,j} - X_{k(i,j)}||$ with $X_{i,j} = 0.5(X_i + X_j)$ and $k(i,j) = argmin_k\{||X_k - X_{i,j}||\}$. They proved that when the support is convex $S \to 0$ and that if the support is not convex the statistic stays bounded. Two disadvantages of this test can be seen:
  - a slow convergence speed due to the average (see for instance Figure 3; for such a set a global and averaged statistic may not be useful decide the non-convexity),

– the necessity of calibration for every distribution and every size sample.

The aim of this paper is to propose another test that erases the two above-mentioned disavantages.

For that it is proposed to build our statistic with a max instead of an average and to look for the maximum value over the entire the convex hull (and not only the middle points of sample pairs). More precisely the following statistic is considered:

$$\delta(\mathcal{X}_n) = d(\mathcal{X}_n, \mathcal{H}(\mathcal{X}_n)) = \max_{x \in \mathcal{X}_n} \min_{y \in \mathcal{H}(\mathcal{X}_n)} (||x - y||).$$

It will be seen that such a statistic is useful to decide if there is a non-convexity due to a non convex boundary when the sample is assumed to be uniform (and the support to have a $\mathcal{C}^2$ boundary). Another statitistic to solve the case of non-convexity due to an interior hole under the uniform hypothesis is also built. The theoretical properties of a test based on these two statistics is detailed in Section 1. Section 2 is devoted to adapt the statistic and the theoretical results to the most useful case where the density is unknown.

Our test gives very good results but the ability to compute it is conditionnal upon the ability to compute the convex hull and the Delaunay complex. This requires quite small dimension. When the dimension becomes too high to compute our statistics, the test detailed in [10] can be used (as its computational time does not depend on the dimension).

## 1.2 Hypotheses

The support $S$ of the density is defined by $S = \overline{\{x, f(x) > 0\}}$. Throughout the paper we assume $\min_{x \in S} f(x) = f_m > 0$. This is a important and necessary hypothesis as mentioned in [10].

It also will we required that $\partial S$ ,the boundary of $S$, is a $\mathcal{C}^2$ manifold. This is an important hypothesis when testing non-convexity due to a boundary non-convexity but it can be removed when dealing with non-convexity due to an inside hole.

It is also supposed that the dimension of $S$ and the dimension of the space of the observation are the same. This is only a technical hypothesis and the important dimension is that of $S$. If $S$ is $d'$ dimensional convex set with $d' < d$, a $PCA$ can isometrically map $S \subset \mathbb{R}^d$ onto $S'\mathbb{R}^{d'}$ so the test can be applied on $S'$.

### 1.3 Notations

Throughout the paper $\theta_d$ is the volume of the $d-$dimensional unit ball.

For the sample $\mathcal{X}_n$, $Vor(X_i)$ denotes the Voronoi cell of $X_i$, i.e. the set of all points closer to $X_i$ than to any other observations.

Let $A$ be a $d-$dimensional set

- $V(A)$ denotes the volume of $A$,
- $\partial A$ denotes the boundary of $A$,
- $\mathcal{H}(A)$ denotes the convex hull of $A$,
- $\Omega(A)$ denotes the affine surface area of $A$ [13],
- $\nu(A, \varepsilon)$ is the interior covering number of $A$, i.e. the smallest number of balls centered in $A$ and with a radius $\varepsilon$ that covers $A$. We denote by $c(A)$ and $C(A)$ the two constants that can be defined by: $\nu(A, \varepsilon) \sim c(A)\varepsilon^{-d}$ and, for all $\varepsilon \leq 1$, $\nu(A, \varepsilon) \leq C(A)\varepsilon^{-d}$
- The function $\alpha_A$ defined as follows:

$$\alpha_A[r] = \inf_{x \in A, \rho \leq r} \frac{V(\mathcal{B}(x, \rho) \cap A)}{V(\mathcal{B}(x, \rho))},$$

plays an important role in the paper.

## 2 Test for convexity under uniform hypothesis

### 2.1 Introduction

In this part $\mathcal{X}_n = \{X_1, ..., X_n\}$ is supposed to be a random sample drawn from an uniform law on an unknown support $S$. We wish to be able to decide whether $S$ is convex or not. The underlying idea of the test is the following: if $S$ is a convex set, then $d(S, \mathcal{H}(S)) = 0$. Obviously, as $S$ is unknown, $d(S, \mathcal{H}(S))$ can not be measured and it is proposed here to evaluate:

$$\delta(\mathcal{X}_n) = d(\mathcal{X}_n, \mathcal{H}(\mathcal{X}_n)) = \max_{x \in \mathcal{X}_n} \min_{y \in \mathcal{H}(\mathcal{X}_n)} (||x - y||).$$

If one wants to test $H_0$ versus $H_1$ with:

- $H_0$: The support is a convex set with boundary, $\partial S$, of differentiability class $\mathcal{C}^2$

- $H_1$: The support is a non-convex set with boundary, $\partial S$, of differentiability class $\mathcal{C}^2$

then the test $T_1$ defined via its critical region

$$W_1 = \{\delta(\mathcal{X}_n) > a\}$$

can be realized and its properties are described by Theorem 1

**Theorem 1.** *When the density is assumed to be uniform then :*

- *(i) If $H_0$ is true and $\lambda > 4$ is a constant:*

$$\frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \leq \frac{4V(S)}{\theta_d} \text{ almost surely.}$$

$$\frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \geq \frac{\lambda V(\mathcal{H}(\mathcal{X}_n))}{\theta_d} \text{ for finitely many } n.$$

- *(ii) Given a fixed support that satisfies the hypthesis $H_1$ there exists a constant $d_S$ such that:*

$$\delta(\mathcal{X}_n) \leq d_S \text{ for finitely many } n \text{ and so } \frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \to +\infty.$$

- *(iii) Let us denote $\delta_{0,n} = (\frac{\log(n)\delta_0}{n})^{1/d}$ and $\varepsilon_{0,n} = \frac{2V(\mathcal{H}(\mathcal{X}_n))}{\theta_d n\delta^d(\mathcal{X}_n)}$. Then :*
  - *$P^{H_0}(\frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \geq \delta_0) \leq p_n(S, \delta_0)$ with:*

$$p_n(S, \delta_0) = \nu(S, \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d \alpha_S[\delta_{0,n}]\delta_{0,n}^d(1-\varepsilon_{0,n})^d}{V(S)}\right)^n$$

  - *$p_n(S, \delta_0)$ depends on unknown parameters but it can be approximated by $\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0)$ with:*

$$\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0) = \nu(\mathcal{H}(\mathcal{X}_n), \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d \delta_{0,n}^d(1-\varepsilon_{0,n})^d}{2V(\mathcal{H}(\mathcal{X}_n))}\right)^n$$

  *with the following asymptotical result: there exist $a_1$ and $a_2$ two constants such that:*

$$a_1 \leq \frac{n^{1/d}}{\log(n)^{1+1/d}}(\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0)/p_n(S, \delta_0) - 1) \leq a_2 \ a.s.$$

*Remark:* Points $(i)$ and $(ii)$ prove that there exist consistant decision rules for the test $T_1$.

Such a test gives good results when non-convexity is induced by the boundary ($\partial(\mathcal{H}(S)) \neq \partial(S)$) but empirically it seems a little weak when the non convexity is induced the existence of a hole in $S$ ($\partial(\mathcal{H}(S)) = \partial(S)$ and $(\mathcal{H}(S)) \neq (S)$). To understand why and correct that default let us notice that $\delta_n$ can also be written:

$$\delta_n(\mathcal{X}_n) = \sup_r\{r \text{ such that: } \exists x \in \mathcal{H}(\mathcal{X}_n) \text{ such that: } \forall i, X_i \notin \mathcal{B}(x, r)\},$$

i.e. $\delta_n$ is the maximum radius for a ball centerd in $x \in S$ that does not contain any observation.

The $p-$value computation is thus the evaluation of the probability, under $H_0$, that there exists a ball $B_0$ centered in $S$ that does not contains any observation. Such a probability is obviously very dependent on the volume of the intersection between the ball $B_0$ and $S$. When there is no knowledge of the location of the center, the worst situation has to be considered, i.e. the center is located near the boundary and the volume of the intersection is approximately the volume of the half ball (because of the diffentiability class of $\partial S$).

To correct that effect we also define:

$$\delta_n^{int}(\mathcal{X}_n) = \sup_r\{r \text{ such that: } \exists x \ , \mathcal{B}(x, r) \subset \mathcal{H}(\mathcal{X}_n) \text{ and } \forall i, X_i \notin \mathcal{B}(x, r)\}$$

If one wants to test $H_0$ versus $H_1$ with:

- $H_0$: The support is a convex set,
- $H_1$: The support is a non-convex set,

the test $T_2$ defined via its critical region

$$W_2 = \{\delta_n^{int}(\mathcal{X}_n) > a'\}$$

can be realized and its properties are described by Theorem 2.

**Theorem 2.** *When the density is assumed to be uniform then :*

- *(i) If $H_0$ is true and $\lambda > 2$ is a constant then:*

$$\frac{n\delta^{int}(\mathcal{X}_n)^d}{\log(n)} \leq \frac{2V(S)}{\theta_d} \text{ almost surely.}$$

$$\frac{n\delta^{int}(\mathcal{X}_n)^d}{\log(n)} \geq \frac{\lambda V(\mathcal{H}(\mathcal{X}_n))}{\theta_d} \text{ for finitely many } n.$$

- *(ii) Given a fixed support that satisfies the hypthesis $H_1$ there exists a constant $d'_S$ such that:*

$$\delta^{int}(\mathcal{X}_n) \leq d'_S \text{ for finitely many } n \text{ and so } \frac{n\delta^{int}(\mathcal{X}_n)^d}{\log(n)} \to +\infty.$$

- *(iii) Let us denote $\delta_{0,n} = (\frac{\log(n)\delta_0}{n})^{1/d}$ and $\varepsilon_{0,n} = \frac{V(\mathcal{H}(\mathcal{X}_n))}{\theta_d n\delta^d(\mathcal{X}_n)}$ then:*
  - *$P^{H_0}(\frac{n\delta^{int}(\mathcal{X}_n)^d}{\log(n)} \geq \delta_0) \leq q_n(S,\delta_0)$ with:*

$$q_n(S,\delta_0) = \nu(S, \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{V(S)}\right)^n$$

  - *$q_n(S,\delta_0)$ depends on unknown parameters but it can be approximated by $\hat{q}_n(\mathcal{H}(\mathcal{X}_n),\delta_0)$ with:*

$$\hat{q}_n(\mathcal{H}(\mathcal{X}_n),\delta_0) = \nu(\mathcal{H}(\mathcal{X}_n), \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{V(\mathcal{H}(\mathcal{X}_n))}\right)^n$$

  *with the following asymptotical result: there exists $b_1$ and $b_2$ two constants such that :*

$$b_1 \leq \frac{n^{1/d}}{\log(n)^{1+1/d}}(\hat{p}_n(\mathcal{H}(\mathcal{X}_n),\delta_0)/p_n(S,\delta_0) - 1) \leq b_2 \text{ a.s.}$$

To conclude, the following strategy to test the convexity of the support under the uniform hypothesis is chosen:

Compute the two estimated $p-values$ $\hat{p}$ and $\hat{q}$, and if the smallest one is small enough reject the null hypothesis.

Proof of Theorem 1 can be found in the appendix, the proof of theorem 2 is very similar (even a little easier), and is left to the reader. The next section is dedicated to the computational aspects of the estimated $p-$value and it is followed by the study of some results.

## 2.2 Computation of the estimated p-value

An interesting fact is that each parameter of the estimated $p-$value can be quite easily computed since the Delaunay Complex and the Voronoi cells can be computed. The only limitation is the ability to obtain the Delaunay complex (and the Voronoi cells) within a reasonable computational time. This limitation becomes significant once the dimension is stricly superior to 4. For such dimensions we will give a way to compute the test more quickly in the next section.

### 2.2.1 Computation of $\delta_n$ and $\delta_n^{int}$

Since the Voronoi cells and the convex hull are computed $\delta_n$ and $\delta_n^{int}$ can be quite easily computed.

Let us first focus on $\delta_n$ and look for $x \in \mathcal{H}(\mathcal{X}_n)$ such that $\mathcal{B}(x, \delta_n)$ does not contains any observation. It is easy to see that:

- if $x \in \mathring{\mathcal{H}}(\mathcal{X}_n)$ (the interior of the convex hull) then $x$ is in $d+1$ different Voronoi cells
- if $x \in \delta\mathcal{H}(\mathcal{X}_n)$ then $x$ is in $d$ differents Voronoi cells

So to find $x$ it is sufficient to consider only a finite set of points and their associated radius:

- Points $y_k$ that are the intersection of $d+1$ different Voronoi cells and that are in $\mathcal{H}(\mathcal{X}_n)$. Here the associated radius is the radius of the circumscript sphere of the associated $X_i$.
- Intersection $z_j$ of "faces" of the boundary of $\mathcal{H}(\mathcal{X}_n)$ (dimension $d-1$) with the intersections of $d$ different Voronoi cells (dimension 1). Here the associated radius is the minimum distance between $x$ and $\mathcal{X}_n$.

Finally $\delta_n = \min\{\min_k\{r(y_k)\}, \min_j\{r(z_j)\}\}$

Let us now focus on $\delta_n^{int}$. It is clear that the associated - $x$: $x \in \mathcal{H}(\mathcal{X}_n)$ such that $\mathcal{B}(x, \delta_n)$ does not contains any observation and such that $\mathcal{B}(x, \delta_n^{int}) \subset \mathcal{H}(\mathcal{X}_n)$- is located on one of the previous $y_k$. The only thing that changes is that the associated radius is now $r^{int}(y_k) = \min\{r(y_k), d(y_k, \delta(\mathcal{H}(\mathcal{X}_n)))\}$

### 2.2.2 Upper bound of $\nu(S, \varepsilon)$

Let us recall that $\nu(S, \varepsilon)$ is the minimum number of balls of radius $\varepsilon$ that are needed to cover $S$. Since a number $\nu^*(S, \varepsilon)$ of balls of radius $\varepsilon$ that cover $S$ is known $\nu(S, \varepsilon) \leq \nu^*(S, \varepsilon)$. It is proposed here to compute a $\nu^*$ as follows. For a simplex $\sigma = \{y_1, .., y_{d+1}\}$, $g(\sigma)$ is the barycentre of the simplex and $R(\sigma) = \max_i(d(y_i, g(\sigma)))$. For all $\varepsilon \geq R(\sigma)$ $\nu(\sigma, \varepsilon) \leq 1$. As $\sigma$ can be divided into $2^d$ simplexes, all isometric we have that for all $\varepsilon \geq R(\sigma)/2$, $\nu(\sigma, \varepsilon) \leq 2^d$ and, iteratively, for all $\varepsilon \geq R(\sigma)/2^k$ , $\nu(\sigma, \varepsilon) \leq 2^{kd}$.

Let us first compute $\mathcal{H}(\mathcal{X}_n)$ and its Delaunay complex $\mathcal{H}(\mathcal{X}_n) = \cup_j \sigma_j^*$ Under $H_0$, $\nu(S, \varepsilon)$ can be estimated by $\nu(S, \mathcal{H}(\mathcal{X}_n)) \leq \sum_j \nu(\sigma_j^* \varepsilon)$.

## 2.3 Some results

Here we present two simulated examples.

Each figure is drawn the same way:

- Blue points represent the data set.
- The yellow complex is the support estimation via the Delaunay complex restricted to nearest neighbors as in [2]. An interesting point is that the proposed test sometimes detects the non-convexity even when the support estimation is convex, and that can be a starting point to improve the density support method.
- The red point is the center of the empty ball. It is surrounded by a black circle when the test $T_2$ has been choosen and by a Red dashed circle when $T_1$ has been choosen.
- The background of the picture is green when the decision is the correct one (accept convexity when support is convex and reject convexity when the support is not convex). The decision has been made according to a comparison between the estimated $p-$value and $5.10^{-2}$.

In the first example, points $(x, y) = (r\cos(\theta), r\sin(\theta))$ are simultated in the part of the disk where the radius $r$ is in $[0.5, 1]$ and the angle $\theta$ in $[0, \theta_0]$. Simulated examples correspond to $\theta_0 \in \{\pi/6, \pi/3, \pi/2, 2\pi/3, 5\pi/6, \pi\}$ (each line of Figure 1 is associated to a value of $\theta_0$) and $n \in \{50, 100, 200, 500\}$ (each column of Figure 1 is associated to a value of $n$). When $\theta_0 \leq \pi/3$ non-convexity is never found. For the first line ($\theta_0 = \pi/6$) it is a decision similar to the "human eye" point of view (the simulated data seem to be convex). Since $\theta_0 \geq \pi/2$ non-convexity is observed for an increasing number of values for $n$, and each time this non-covexity is detected a little after the human eye.

In the second example, points $(x, y) = (r\cos(\theta), r\sin(\theta))$ are simultated in the part of the disk where the radius $r$ is in $[r_0, 1]$ and the angle $\theta$ in $[0, 2\pi]$. Simulated examples correspond to $r_0 \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ (each line of Figure 3 is associated to a value of $\theta_0$) and $n \in \{50, 100, 200, 500\}$ (each column of Figure 3 is associated to a value of $n$).

# 3 Test for convexity without uniform hypothesis

## 3.1 Adaptation of the previous test

In this section it is no longer assumed that $\mathcal{X}_n$ is uniformly randomized on $S$. The sample is now drawn from an unknown density $f$ that has to satisfy some properties:

- $f$ is bounded on $S$ with a lower bound $f_m > 0$. Let us also denote by $f_M$ the maximum value of $f$ on $S$.
- $f$ is continuously differentiable on $S$. this (combined with the previous hypothesis) implies the existence of $K_1$ and $K_{-1/d}$ such that, for all $x$ and $y$ in $S$, $|f(x) - f(y)| \leq K_1||x - y||$ and $|f(x)^{-1/d} - f(y)^{-1/d}| \leq K_{-1/d}||x-y||$ (in fact the existence of such constants is the only needed hypothesis).

The idea is now to build a test for the convexity of $S$ that generalizes the previous test and adapt it to a more general case. Working with the same statitistics $\delta(\mathcal{X}_n)$ and $\delta^{int}(\mathcal{X}_n)$ is a convergent method but may not be very convenient because it does not take into account the local effect of the non-uniformity. The problem in choosing of such a statistic is illustrated in figure 3. To realize this figure 100 points have been uniformly drawn in $\mathcal{B}(0, 1) \setminus \mathcal{B}(0, 0.1)$. Then 100 more points are added, uniformly drawn in

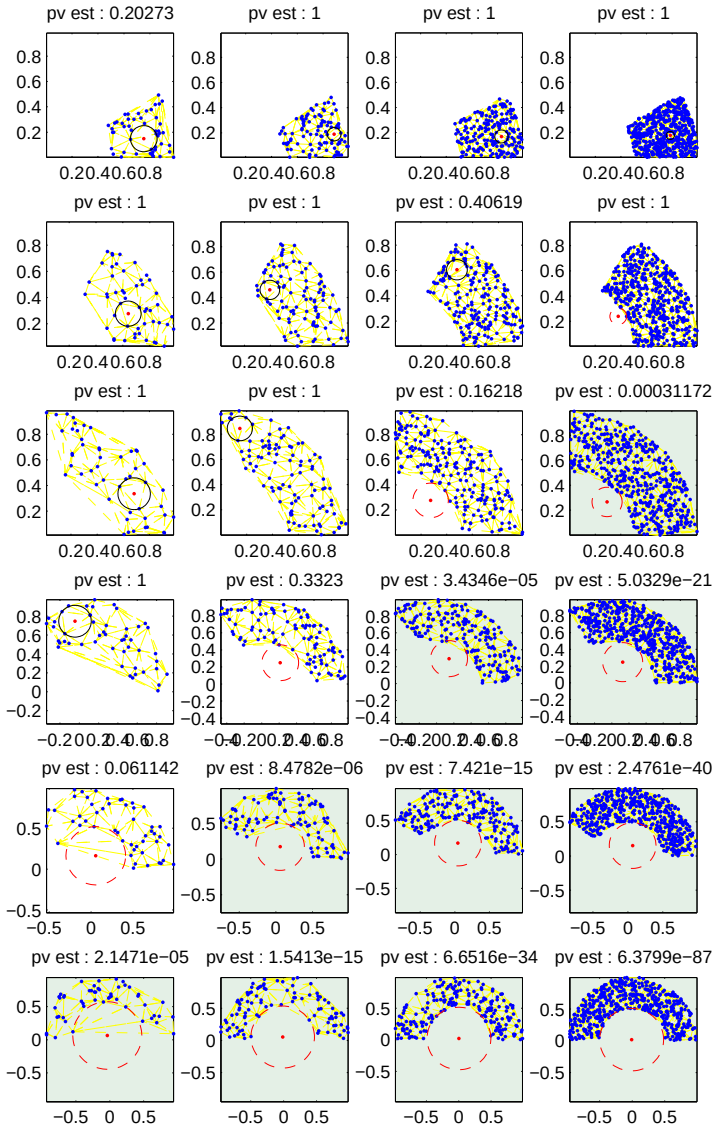


Figure 1: Some results for a uniform sample on an arc.

$\mathcal{B}(0, 0.2) \setminus \mathcal{B}(0, 0.1)$. In this case the support of the density is obviously not convex. Application of the $\delta$ statistics fails to recognize non-convexity due

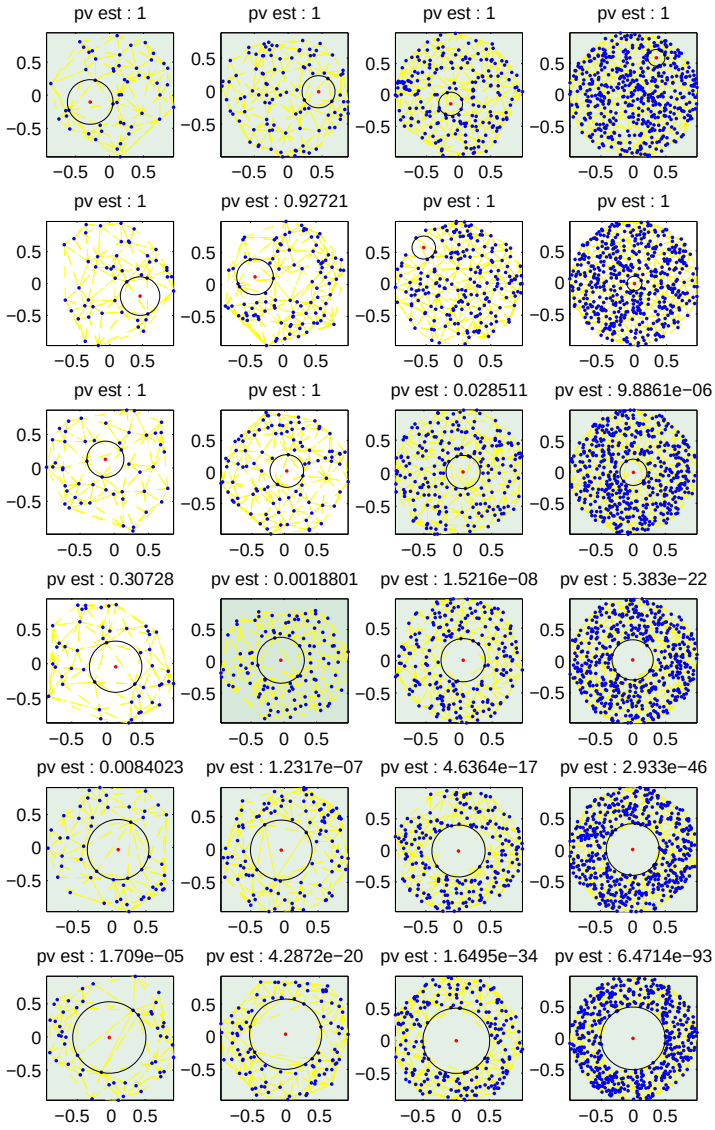


Figure 2: Some results for a uniform sample on an cd.

to the location of the empty ball ($\mathcal{B}(0, 0.1)$ should have been expected) for the estimated $p-$value.

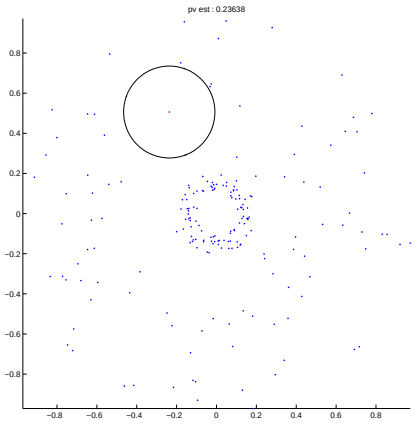


Figure 3: $\delta(\mathcal{X}_n)$ does not take into account the local effects.

The proposed way to deal with the non-uniformity of the density is to consider the volume (weighted by the density) instead of the radius of empty balls. As the density is unknown it leads us to consider the new statistics:

$$\Delta(\mathcal{X}_n) = \sup_r\{r \text{ such that: } \exists x \in \mathcal{H}(\mathcal{X}_n) \text{ such that: } \forall i, X_i \notin \mathcal{B}(x, r/\hat{f}^{1/d}(x))\}$$

and

$$\Delta^{int}(\mathcal{X}_n) = \sup_r\{r \text{ such that: } \exists x, \mathcal{B}(x, r/\hat{f}_n^{1/d}(x)) \subset \mathcal{H}(\mathcal{X}_n) \forall i, X_i \notin \mathcal{B}(x, r/\hat{f}^{1/d}(x))\}$$

The principal other idea is to base the test on a density estimation that over-estimates the density when the support is not convex. It is thus proposed here to work with:

$$\hat{f}_{k_n}(x) = \min_{x \in Vor(X_i)} \tilde{f}_{k_n}(X_i)$$

with $\tilde{f}_{k_n}$ the usual $k_n$ nearest neighbor density estimation ([16]). The choice of such an estimator (instead of the usual nearest neighbor one) is detailed in the next section.

A way to test $H_0$ versus $H_1$ with:

- $H_0$: The support is a convex set with boundary, $\partial S$, of differentiability class $\mathcal{C}^2$.

- $H_1$: The support is a non-convex set with boundary, $\partial S$, of differentiability class $\mathcal{C}^2$.

can be described by Theorem 3:

**Theorem 3.** *If* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $\frac{k_n}{\log(n)^{d+1}} \to +\infty$*:*

- *(i) If* $H_0$ *is true and* $\lambda > 8$ *is a constant then:*

$$\frac{n\Delta(\mathcal{X}_n)^d}{\log(n)} \leq \frac{8}{\theta_d} \text{ almost surely.}$$

$$\frac{n\Delta(\mathcal{X}_n)^d}{\log(n)} \geq \frac{\lambda}{\theta_d} \text{ for finitely many } n.$$

- *(ii) Given a fixed support that satisfies the hypthesis* $H_1$ *there exists a constant* $d_S^*$ *such that:*

$$\Delta(\mathcal{X}_n) \leq d_S^* \text{ for finitely many } n \text{ and so } \frac{n\Delta(\mathcal{X}_n)^d}{\log(n)} \to +\infty.$$

- *(iii) Let us denote* $\delta_{0,n} = (\frac{\log(n)\delta_0}{n})^{1/d}$ *and* $\varepsilon_{0,n} = \frac{2}{\theta_d n \delta^d(\mathcal{X}_n)}$ *Then there exists a function* $p_n^*(S, \delta_0, k_n)$ *that depends on unkown parameters that satisfies*

  - $P^{H_0}(\frac{n\Delta(\mathcal{X}_n)^d}{\log(n)} \geq \delta_0) \leq p_n^*(S, \delta_0, k_n)$
  - $\hat{p}_n^-(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) \leq p_n^*(S, r_0, r_{0,n}\varepsilon_n) \leq \hat{p}_n^+(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n)$ *almost surely*

  *with :*

  - $\hat{p}_n^+(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - 2^{-1}\theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n$
  - $\hat{p}_n^-(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - \theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n$

Now using $\Delta^{int}(\mathcal{X}_n)$ to test $H_0$ versus $H_1$ with:

- $H_0$: The support is a convex set

- $H_1$: The support is a non-convex set

can be done according to Theorem 4

**Theorem 4.** *If* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $\frac{k_n}{\log(n)^{d+1}} \to +\infty$ *then :*

- *(i) If* $H_0$ *is true and* $\lambda > 4$ *is a constant then:*

$$\frac{n\Delta^{int}(\mathcal{X}_n)^d}{\log(n)} \leq \frac{4}{\theta_d} \text{ almost surely.}$$

$$\frac{n\Delta^{int}(\mathcal{X}_n)^d}{\log(n)} \geq \frac{\lambda}{\theta_d} \text{ for finitely many } n.$$

- *(ii) Given a fixed support that satisfies the hypthesis* $H_1$ *there exist a constant* $d_S'^*$ *such that*

$$\Delta^{int}(\mathcal{X}_n) \leq d_S'^* \text{ for finitely many } n \text{ and so: } \frac{n\Delta^{int}(\mathcal{X}_n)^d}{\log(n)} \to +\infty.$$

- *(iii) Let us denote* $\delta_{0,n} = (\frac{\log(n)\delta_0}{n})^{1/d}$ *and* $\varepsilon_{0,n} = \frac{1}{\theta_d n \delta^d(\mathcal{X}_n)}$. *Then there exists a function* $q_n^*(S, \delta_0, k_n)$ *that depends on unkown parameters that satisfies*

  - $P^{H_0}(\frac{n\Delta^{int}(\mathcal{X}_n)^d}{\log(n)} \geq \delta_0) \leq q_n^*(S, \delta_0, k_n)$
  - $\hat{q}_n^-(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) \leq q_n^*(S, r_0, r_{0,n}\varepsilon_n) \leq \hat{q}_n^+(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n)$ *almost surely*

  *with :*

  - $\hat{q}_n^+(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - \theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n$
  - $\hat{q}_n^-(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - 2\theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n$

As in the uniform case the proofs Theorem 3 can be found in the appendix. The following subsection is dedicated to discussing the unusual choice of the density estimation method. Following that, as for the uniform case, we will focus on the computational aspects and present some results.

## 3.2 Remark about the choice of the density estimation method

The choice of the simplest nearest neighbor density estimation has been done for simplicity but it may be interesting to explore if other ways to estimate density the can be made.

The choice of a density estimation that is constant on the Voronoi cells is important for the computational aspects: such a choice allows us to keep the method to look for the maximum descibed in 2.2.1. It is for that reason the coice has been made, but it also has another advantage linked to the existence of a consistent decision rule. Let us first remark that the existence of a consistent decision rule is a corollary of points $(i)$ and $(ii)$ of the theorems.

The choice of this density estimator is linked to point $(ii)$. Let us now imagine that the support is not convex. There exist $x_0 \in \mathcal{H}(S)$ and $d(x_0, S) = d_0 > 0$. Thus $\Delta(\mathcal{X}_n)$ is expected to be higher than $d_0 \hat{f}_{k_n}(x_0)^{1/n}$.

- With our estimation method the estimated density is a $\tilde{f}(X_i)$ and as $X_i \in S$, $\tilde{f}(X_i)$ is expected to be higher than $f_m$ (in fact it will be $2^{-1/d} f_m$ but it is still a positive constant);
- With a classical nearest neighbor density estimation $\hat{f}_{k_n}(x_0)$ will converge towards 0.

## 3.3 Computation of the estimated $p - value$

### 3.3.1 Reasonable values for $d$ and $n$

When $d$ and $n$ are reasonable enough values to allow the computation of the convex hull, the Delaunay complex and the Voronoi cells, the computation of every part of the statitistic can be computed exactly as in the uniform case.

### 3.3.2 High dimension

When the dimension is too high to allow the computation of the convex hull, the Delaunay complex and the Voronoi cells, two problems appear:

- The computation of an upper bound for $\nu(\mathcal{H}(\mathcal{X}_n), \varepsilon)$,
- The computation of the statistic.

For the computation of an upper bound for $\nu(S, \varepsilon)$ it can be proposed to include $\mathcal{H}(\mathcal{X}_n)$ in a box (i.e. a set isometric to a $\prod[0, b_d]$) and to compute the convering number of the box.

For the computation of the statistics, two different cases have to be imagined:

- If $n$ is "reasonable", we can look for an upper bound on every $X_{i,j} = (X_i + X_j)/2$ (idea based on the [10]).
- When $n$ is so fast that the previously proposed algorithm is not fast enough, a quicker maximization algorithm (such as a genetic algorithm) has to be considered.

## 3.4 Some Results

As for the uniform, case results are better in practice than the poor convergence speed seems to indicate. We will not present here a detailed series of examples, but we make the choice to study the impact of the choice for $k$ the number of neighbors. When working with $k-$nearest neighbors an asymptotical result for convergence is established but it does not gives accurate values for $k$ for specific studies.

We first recall that in all applications convexity is prefered to non-convexity. Let us first observe the behaviour of the estimated $p-$value for a simulation on a support non-convex support. Here we have computed $N = 100$ samples of size $n = 200$ with $d = 2$ and we have used every value for $k$ from 1 to 199. Computed $p-$values are presented in Figure 4, and it can be observed that every value from 5 to 199 can be chosen. It can also be observed that when the number of neighbors is too small there is instability and the estimated $p-$value can be too small.

We now test our method for a sample of size 200 and dimension 2, realized as follows :

- $r$ follows a normal law of mean 0 and variance 1 conditioned to have its values in $[a_0, 2]$ (with $a_0 \in \{0.2, 0.25, 0.3, 0.35, 0.4, 0.45\}$),
- $\theta$ follows a uniform law on $[0, 2\pi]$,
- $X = (r\cos(\theta), r\sin(\theta))$.

Results are presented in Figure 5. To see the effect of the introduction of the estimated density, the first column of the figure presents the result for the first test based on the uniform hypothesis. The second column presents the result for the test without the uniform hypothesis. The indicated $p-$value has been automatically computed as the smallest for the tested values of $k$. This $p-$value has to be compared to the last column which presents the estimated $p-$value as a function of the number of neighbors. For instance, for

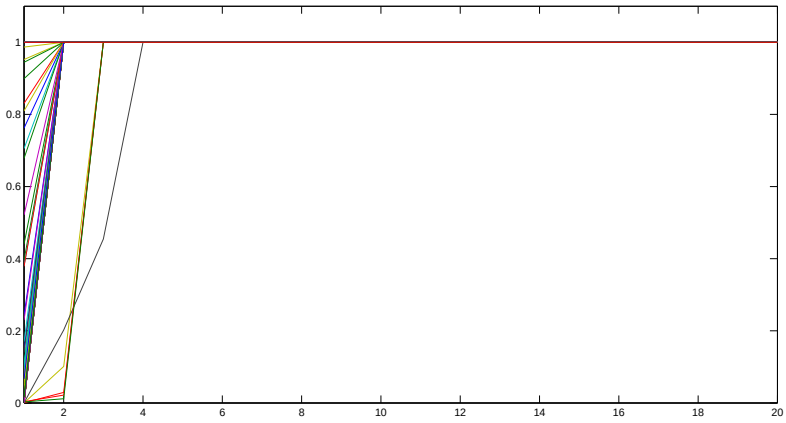


Figure 4: Estimation of the $p-$value for $N = 100$ samples of size $n = 200$ with $d = 2$ uniformly randomized on a $d-$dimensional ball for $k$ from 1 to 20 (for higher $k$ the estimated $p-$value is always 1)

the third line the minimum value is very small but the minimum is realized for $k = 5$. According to the previous remark, when looking at the result on a known convex support we should not trust this small value but rather trust the 0.06 value observed for $k = 84$. Let us remark that the second test always correctly locates the hole in the support even if it does not always recognize the non-convexity of the support. The values $a_0 \in [0.3, 0.35]$ have really been observed as a critical zone where the second test starts to sometimes recognize the non convexity.

When computing examples the same behaviour has always been observed for the estimated $p-$value as a function of the number of neighbors:

- For small $k$ values there is sometimes a highly irregular zone
- The $p-$value decreases then seems quite stable around a minimum value before increasing again.

It is proposed to consider the stable minimum zone to give an approximation of the $p-$value.

# 4 Conclusion

The proposed method to test the convexity of the density support gives quite good results but some theoretical improvement may be done :

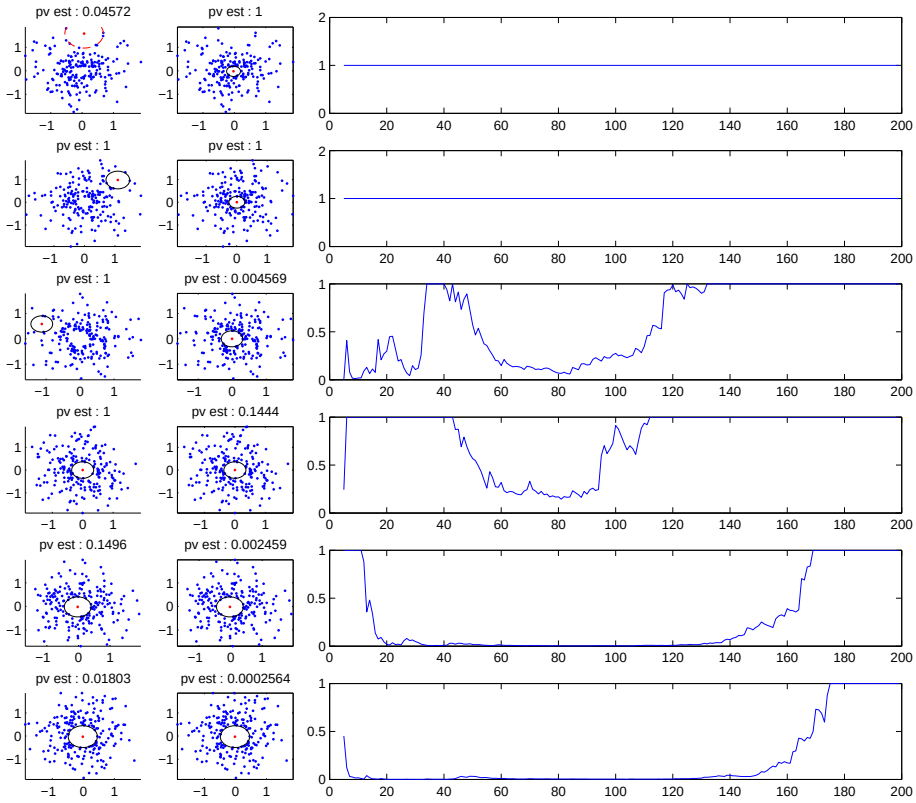


Figure 5: Estimation of the $p-$value for $N = 100$ samples of size $n = 200$ with $d = 2$ uniformly randomized on a $d-$dimensional ball for $k$ from 1 to 20 (for higher $k$ the estimated $p-$value is always 1).

- Can a confidence interval on the $p-$value be found instead of a mupper bound ?
- The strong hypothesis on the density $\min_S f > 0$ cannot be removed (as mentioned in [10]), but when $\min_S f = 0$ can our test be used to test the convexity of level subsest of the density ($E_\lambda = \{x$ such that:$f(x) \geq \lambda\}$)?

Various applications can easily be imagined for a test for convexity. Let us mention here two kinds of application for which the work has been started and the results encouraging:

- As mentioned in the introduction, it can help compute the geodesic distance, and so is useful for any statistical method that is based on geodesic distance. Such methods originally began with isomap[**?**]. Working with the geodesic distance instead of the euclidian one is known to solve many problems and so has become quite popular (see for instance [6] or [1] for new developments on the use of geodesic distance). In [10] it is proposed to use convexity test to select a result with an a posteriori criteria, but we think that it can be used to compute the geodesic distance a priori. The idea is very intuitive: build the graph as follows : connect $x$ to all the observations that are in $\mathcal{B}(x, r_x)$, with $r_x$ the highest value such that $S \cap \mathcal{B}(x, r_x)$ is convex.
- As mentioned in 2.3, a small adaptation from our test can be used to compute a density support estimation based on a restricted Delaunay complex with a better rule than this provided in [2].

# References


[1] S. H. Joshi A. Srivastava, C. Samir and M. Daoudi. Elastic shape models for face analysis using curvilinear coordinates. *Journal of Mathematical Imaging and Vision*, 33:253–265, 2009.

[2] C. Aaron. Using the k-nearest neighbor restricted delaunay complex to estimate the density support and its topological properties. submitted to Electronic Journal Of Statistics.

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

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

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

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

[7] C. Penrod L Devroye. The strong uniform convergence of multivariate variable kernel estimates. *The Canadian Journal of Statistics*, 14:211–219, 1986.

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

[9] S. Vempala L. Rademacher. Testing geometric convexity. presentation at FSTTCS 2004, IMSc, Chennai.

[10] A. Hernandez P. Delicado and G. Lugosi. Testing the convexity of the support of a distribution.

[11] 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.

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

[13] C. Schütt. On the affine surface area. *Proceedings of the American Mathematical Society*, 118:1213–1218, 1993.

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

[15] C. Schütt E. Werner. Polytopes with vertices chosen randomly from the boundary of a convex body. In *Israel Seminar 20012002 (V. D. Milman, G. Schechtman, eds.), pp. 241422, Lecture Notes in Math., vol. 1807, Springer, New York*, 2003.

[16] M Rosenblatt Y.P Mack. Multivariate k-nearest neighbor density estimates. *Journal of Multivariate Analysis*, 9:1–15, 1979.

# Appendices

## A Proof of Theorem 1

### A.1 Theoretical upper bound on the probability (Point $(iii-1)$)

Let us start with some elementary lemmas:

**Lemma 1.** *Let us denote*

$$\alpha_S[r] = \inf_{x\in S, \rho\leq r} \frac{V(\mathcal{B}(x,\rho)\cap S)}{V(\mathcal{B}(x,\rho))}$$

*If $\partial S$ is of differentiability class $\mathcal{C}^2$ then:*

- *there exists $r_0(S)$ such that, when $r \leq r_0(S)$, $\alpha_S[r]$ is a decreasing function (i.e. for all $r < r' \leq r_0$, $\alpha_S[r] < \alpha_S[r']$).*
- $\lim_{r\to 0} \alpha_S[r] = 0.5$*, and more precisely there exists $r'_0$ and $a_S > 0$ such that, when $r \leq r'_0$, $|\alpha_S[r] - 0.5| < a_S r$.*

This is a direct corollary of Lemma 2.3 of Penrose [11].

**Lemma 2.** *Let us pick $x$ deterministically in $S$. Then*
$P(\mathcal{B}(x,r)$ *does not contains any observation* $) = (1 - \frac{V(\mathcal{B}(x,r)\cap S)}{V(S)})^n$
$P(\mathcal{B}(x,r)$ *does not contains any observation* $) \leq (1 - \frac{\alpha[r]\theta_d r^d)}{V(S)})^n$

**Corollary 1.** *Let us pick points $(x_1, ..., x_{\nu(S,\varepsilon)})$ deterministically in $S$ such that $S \subset \cup\mathcal{B}(x_i, \varepsilon)$. Then*
$P(\exists i$ *such that:* $\mathcal{B}(x_i, r) \cap \mathcal{X}_n = \emptyset) \leq \nu(S,\varepsilon)(1 - \frac{\alpha[r]\theta_d r^d)}{V(S)})^n$.

*Proof.* Direct consequence of Lemma 2. □

**Corollary 2.** *When $r \leq r_0(S)$, for all $\varepsilon \in ]0, r]$, $P(\delta(\mathcal{X}_n > r) \leq \nu(S,\varepsilon)(1 - \frac{\alpha[r]\theta_d (r-\varepsilon)^d)}{V(S)})^n$.*

*Proof.* For a fixed $\varepsilon$ let us cover $S$ with small balls of radius $\varepsilon$ as in Corollary 1. Let us suppose that $\delta(\mathcal{X}_n) > r$. The compacitness of $S$ implies the existence of $x \in S$ such that $\mathcal{B}(x,r)\cap\mathcal{X}_n = \emptyset$. There exists $i$ such that $x \in \mathcal{B}(x_i,\varepsilon)$ and $\mathcal{B}(x_i, r-\varepsilon) \subset \mathcal{B}(x,r)$ does not contains any observation. Lemma 1 allows us to replace $\alpha[r-\varepsilon]$ by $\alpha[r]$ and Corollary 1 gives the conclusion. □

**Corollary 3.** *For all $\delta_0 > 0$ let us denote $\delta_{0,n} = (\frac{log(n)\delta_0}{n})^{1/d}$ and $\varepsilon_{0,n} = \frac{2V(\mathcal{H}(\mathcal{X}_n))}{\theta_d \delta_0 \log(n)}$.*
*There exists $n_0$ such that, for all $n \geq n_0$:*

$$P\left(\delta(\mathcal{X}_n) > \delta_{0,n}\right) \leq p_n(S,\delta_0) = \nu(S, \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d \alpha_S[\delta_{0,n}]\delta_{0,n}^d (1-\varepsilon_{0,n})^d}{V(S)}\right)^n.$$

*Proof.* This is a direct consequence of the previous corollary with $r = \delta_{0,n}$, $\varepsilon = \delta_{0,n}\varepsilon_{0,n}$ and $n_0 = \min\{n$ such that: $\delta_{0,n} \leq r_0$ and $\varepsilon_{0,n} < 1\}$. □

**Remark:** The choice for $\varepsilon_{0,n}$ has been empirically made as follows. When $n \to +\infty$, $c(S, \delta_{0,n}\varepsilon_{0,n}) \to c(S)$ and so $p_n \sim c(S)\varepsilon^{-d}(1-\frac{0.5\theta_d\theta_d(\delta_{0,n}-\varepsilon)^d)}{V(S)})^n$ and we want to minimize this function. If $\varepsilon^{-d}(1-\frac{0.5\theta_d\theta_d(\delta_{0,n}-\varepsilon)^d)}{V(S)})^n$ is minimized for $x_n = o(\delta_{0,n})$ (which is empirically observed) then $x_n \sim \frac{2V(S)}{\theta_d\delta_0\log(n)}\delta_{0,n} \sim \frac{2V(\mathcal{H}(\mathcal{X}_n))}{\theta_d\delta_0\log(n)}\delta_{0,n}$.

The first part of Theorem 1 is thus proved.

## A.2 Proof of convergence (point $(iii-2)$)

It has now to be proved that

$$\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0) = \nu(\mathcal{H}(\mathcal{X}_n), \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d\delta_{0,n}^d(1-\varepsilon_{0,n})^d}{2V(\mathcal{H}(\mathcal{X}_n))}\right)^n$$

gives a good approximation of

$$\nu(S, \delta_{0,n}\varepsilon_{0,n}) \left(1 - \frac{\theta_d\alpha_S[\delta_{0,n}]\delta_{0,n}^d(1-\varepsilon_{0,n})^d}{V(S)}\right)^n$$

i.e. that:

- $\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0)/p_n(S, \delta_0) \xrightarrow{L^2} 1$, and
- $\frac{n^{1/(d)}}{\log(n)}(\hat{p}_n(\mathcal{H}(\mathcal{X}_n), \delta_0)/p_n(S, \delta_0) - 1) = O_p(1)$.

### A.2.1 Convergence for the first term of the product

We now study the convergence of $\nu(\mathcal{H}(\mathcal{X}_n), \varepsilon)/\nu(S, \varepsilon)$ toward 1. Let us first remark that if $S$ is convex then $\nu(\mathcal{H}(\mathcal{X}_n), \varepsilon)/\nu(S, \varepsilon) \leq 1$. The main problem is to find a lower bound for this quantity. For that, let us remark that:

**Lemma 3.** *Let us denote $S_a^- = \{x \in S, \mathcal{B}(x, a) \subset S\}$. For all $r \leq a$,*

$$P(\exists x \in S_a^- \text{ such that } \mathcal{B}(x, r) \cap \mathcal{X}_n = \emptyset) \leq \nu(S, \varepsilon)(1 - \theta_d(r-\varepsilon)^d/V(S))^n.$$

The proof is the same as that of Corollary 2.

**Corollary 4.** *Let us denote*

$$\delta_a(\mathcal{X} = n) = max_{x \in S_a^-}\{r \text{ such that } \mathcal{B}(x, r) \cap \mathcal{X}_n = \emptyset\}.$$

*Then*

$$\delta_a(\mathcal{X}_n) \leq \left(\frac{2V(S)\log(n)}{\theta_d n}\right)^{1/d} \text{ almost surely.}$$

*Proof.* Let $\rho_n = (\frac{2V(S)\log(n)}{\theta_d n})^{1/d}$. Apply Lemma 3 replacing:

- $r$ by the series $\rho_n(\varepsilon') = \rho_n(1+\varepsilon')$,
- $\varepsilon$ by the series $e(\varepsilon') = \rho_n\varepsilon'/2$.

Then $P(\delta(\mathcal{X}_n) \geq \rho_n(\varepsilon') \leq C(S)(\rho_n\varepsilon'/2)^{-d}\frac{n}{log(n)}O(n^{-2(1+\varepsilon'/2)^d})$
so $\sum P(\delta(\mathcal{X}_n) \geq \rho_n(\varepsilon') < +\infty$. The Borrel-Cantelli lemma gives the conclusion.

□

In the following the notation $\rho_n = (\frac{2V(S)\log(n)}{\theta_d n})^{1/d}$ is kept.

**Corollary 5.** *If $S$ is convex $S^-_{2\rho_n} \subset \mathcal{H}(\mathcal{X}_n)$ almost surely*

*Proof.* Let us suppose the converse. Then there exists a point $x \in S^-_{2\rho_n}$ that is not in $\mathcal{H}(\mathcal{X}_n)$. As $\mathcal{H}(\mathcal{X}_n)$ is obvioulsy convex, there exists $\overrightarrow{u}$ with $||\overrightarrow{u}|| = 1$ and, for all $i$, $\overrightarrow{xX_i}.\overrightarrow{u} \leq 0$. Let us define $y$ as follows: $\overrightarrow{xy} = \rho_n\overrightarrow{u}$. It is easy to see that:

- $y \in S^-_{\rho_n}$,
- $\mathcal{B}(y, \rho_n)$ does not contains any observation.

That is (almost surely) not possible. □

**Corollary 6.** *There exists a constant $c_*(S)$ such that*

$$1 - c_*(S)\left(\frac{\log(n)}{n}\right)^d \frac{\nu(\mathcal{H}(X_n), \delta_{0,n}\varepsilon_{0,n})}{\nu(S,)\delta_{0,n}\varepsilon_{0,n}} \leq 1.$$

*Proof.* For a fixed $n$ let us cover $\partial S$ with deterministic points $x^*_{1,n}, ..., x^*_{i,n}, ...x^*_{\nu(\partial S, 2\rho_n),n}$. $\nu(\partial S, 2\rho_n), n < C(\partial S)\rho_n^{-d+1}$. It is easy to see that $S \setminus S^-_{2\rho_n} \subset \cup_i \mathcal{B}(x^*_{i,n}, 4\rho_n)$
so

$$\nu(S \setminus S^-_{2\rho_n}, \delta_{0,n}\varepsilon_{0,n}) \leq \sum_i \nu(\mathcal{B}(x^*_{i,n}, 4\rho_n), \delta_{0,n}\varepsilon_{0,n}),$$

$$\nu(S \setminus S^-_{2\rho_n}, \delta_{0,n}\varepsilon_{0,n}) \leq C(\partial S)(\rho_n/2)^{-d+1}\nu(\mathcal{B}(0,1), \delta_{0,n}\varepsilon_{0,n}/(4\rho_n)),$$

$$\nu(S \setminus S^-_{2\rho_n}, \delta_{0,n}\varepsilon_{0,n}) \leq 2^{d+1}C(\partial S)C(\mathcal{B}(0,1))\rho_n(\delta_{0,n}\varepsilon_{0,n})^{-d}.$$

On the other hand,

$$\nu(S, \delta_{0,n}\varepsilon_{0,n}) \geq C_-(S)(\delta_{0,n}\varepsilon_{0,n})^{-d},$$

so according to the previous corollary,

$$\frac{\nu(S \setminus \mathcal{H}(\mathcal{X}_n), \delta_{0,n}\varepsilon_{0,n})}{\nu(S, \delta_{0,n}\varepsilon_{0,n})} \leq \rho_n \frac{2^{d+1}C(\partial S)C(\mathcal{B}(0,1))}{C_-(S)}.$$

□

### A.2.2 Convergence for the second term of the product

**Corollary 7.** *Let $S$ be a $d-$dimensional convex set that has a boundary, $\partial S$, of differentiability class $\mathcal{C}^2$:*

*let $\mathcal{X}_n = \{X_1, ..., X_n\}$ be a set of $n$ independant and uniformly distributed points on $S$.*

*Let $\delta_{0,n}$ be the series $\delta_{0,n} = (\frac{log(n)\delta_0}{n})^{1/d}$. Then*

$$n\delta_{0,n}^d \theta_d \left( \frac{\alpha_S[\delta_{0,n}]}{V(S)} - \frac{1}{2V(\mathcal{H}(\mathcal{X}_n))} \right) \xrightarrow{a.s.} 0.$$

*More precisely there exist constants $a_1$ and $a_2$ such that*

$$a_1 \leq \frac{n^{1/d}}{\log(n)^{1+1/d}} \left[ n\delta_{0,n}^d \theta_d \left( \frac{\alpha_S[\delta_{0,n}]}{V(S)} - \frac{1}{2V(\mathcal{H}(\mathcal{X}_n))} \right) \right] \leq a_2 \ a.s..$$

*Proof.* This is a direct consequence of Theorem 5 and Lemma 1. Let us write $\frac{\alpha_S[\delta_{0,n}]}{V(S)} - \frac{1}{2V(\mathcal{H}(\mathcal{X}_n))} = \frac{\alpha_S[\delta_{0,n}]-0.5}{V(S)} + \frac{1}{2}\left( \frac{1}{V(S)} - \frac{1}{V(\mathcal{H}(\mathcal{X}_n))} \right)$. Lemma 1 and Theorem 5 imply that the two parts of the sum are both $O((\log(n)/n)^{1/d})$. □

**Corollary 8.** *Let $S$ be a $d-$dimensional convex set that has a boundary, $\partial S$, of differentiability class $\mathcal{C}^2$:*

*Let $\mathcal{X}_n = \{X_1, ..., X_n\}$ be a set of $n$ independant and uniformly distributed points on $S$.*

*Let $\delta_{0,n}$ be the series $\delta_{0,n} = (\frac{log(n)\delta_0}{n})^{1/d}$ and $\varepsilon_{0,n} \in [0,1], \varepsilon_{n,0} \to 0$. Then*

$$\left( \frac{1 - \frac{\theta_d \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{2V(\mathcal{H}(\mathcal{X}_n))}}{1 - \frac{\theta_d \alpha_S[\delta_{0,n}] \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{V(S)}} \right)^n \xrightarrow{a.s.} 1,$$

*and there exists $b_1$ and $b_2$ such that*

$$b_1 \leq \frac{n^{1/d}}{\log(n)^{1+1/d}} \left[ \left( \frac{1 - \frac{\theta_d \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{2V(\mathcal{H}(\mathcal{X}_n))}}{1 - \frac{\theta_d \alpha_S[\delta_{0,n}] \delta_{0,n}^d (1-\varepsilon_{0,n})^d}{V(S)}} \right)^n - 1 \right] \leq b_2 \ a.s.$$

The convergence part of Theorem 1 is a direct consequence of Corrolaries 6 and 8.

## A.3 Note on the convergence speed

For Theorem 1 our easy work on $\mathcal{X}_n$ that lead to Corrolaries 6, 7 and 8 is sufficient since, in the first case the convergence speed can not be higher than $n^{1/d}/(\log n^{1+1/d})$ because of the $\alpha_S$ term. When proving Theorem 2 every step is the same but getting rid of the $\alpha$ term.

the existence of the following theorem due to [3],[14],[12] may indicates that the true convergence speed is higher than ours.

**Theorem 5.** *Let $S$ be a $d-$dimensional convex set and $\mathcal{X}_n = \{X_1, ..., X_n\}$ be a random sample of $n$ independant and uniformly distributed points in $S$. Then*

$n^{2/(d+1)}(V(S) - V(\mathcal{H}(\mathcal{X}_n))) \xrightarrow{L^1} \gamma_d V(S)^{2/(d+1)}\Omega(S)$.

*If now the boundary of $S$, $\partial S$ is assumed to have a differentiability class $\mathcal{C}^2$ then*

$n^{2/(d+1)}(V(S) - V(\mathcal{H}(\mathcal{X}_n))) \xrightarrow{L^2} \gamma_d V(S)^{2/(d+1)}\Omega(S)$

*with $\gamma_d$ a constant that only depends on $d$ and $\Omega(S)$ the affine surface area.*

However working with high refinement on the convex hull is a very difficult task (see for instance the 20 years between results on the mean and results on the variance or the length of [15]), and it is reasonable to satisfy ourselves with a quick result that allows almost surely convergence for the volume and that allows us to prove the convergence of the covering number.

## A.4 Proof of the two first points

Recall that the first point of Theorem 1 is :

$(i)$ If $H_0$ is true and $\lambda > 4$ is a constant then:

$$\frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \leq \frac{4V(S)}{\theta_d} \text{ almost surely.}$$

$$\frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \geq \frac{\lambda V(\mathcal{H}(\mathcal{X}_n))}{\theta_d} \text{ for finitely many } n.$$

The first inequality can be proved exactly as in Corollary 4 (using Corollary 1 instead of Lemma 3). To prove the second part one has to combine this with Corollary 4 and Corollary 5.

Let us now focus on the second point, which does not have a proof similar to those above:

**Lemma 4.** *Given a fixed support that satisfies the hypthesis $H_1$, there exists a constant $d_S$ such that*

$$\delta(\mathcal{X}_n) \leq d_S \text{ for finitely many } n \text{ and so } \frac{n\delta(\mathcal{X}_n)^d}{\log(n)} \to +\infty.$$

*Proof.* Let us assume that $S$ is not convex, then there exists $x_0 \in \mathcal{H}(S)$ with $x_0 \notin S$. As $S$ is a closed set this can be improved to : $x_0 \in \mathring{\mathcal{H}}(S)$ with $x_0 \notin S$. We are now going to prove that:

$$x_0 \notin \mathcal{H}(\mathcal{X}_n) \text{ for finitely many } n.$$

Since $x_0 \in \mathring{\mathcal{H}}(S)$, there exists $y_1, .., y_d \in S$ such that $x_0 \in \mathring{\mathcal{H}}(\{y_1, .., y_d\})$. So there exists $\varepsilon > 0$ such that for any $(y'_1, .., y'_d) \in \Pi\mathcal{B}(y_i, \varepsilon)$, $x_0 \in \mathring{\mathcal{H}}(\{y'_1, .., y'_d\})$. Thus :

$$P(x_0 \notin \mathcal{H}(\mathcal{X}_n)) \leq P(\exists i \text{ such that: } \mathcal{B}(y_i, \varepsilon) \cap \mathcal{X}_n = \emptyset),$$

$$P(x_0 \notin \mathcal{H}(\mathcal{X}_n)) \leq d(1 - \alpha_S[\varepsilon]\theta_d \varepsilon^d / V)^n.$$

The series is summable, so the Borrel-Cantelli Lemma implies that $x_0 \notin \mathcal{H}(\mathcal{X}_n)$ for finitely many $n$.

This implies that $\delta(\mathcal{X}_n) < d(x_0, S)$ for finitely many $n$. The fact that $d(x_0, S) > 0$ because of the closeness of $S$ completes the proof.

□

# B Proof of Theorem 3

The proof of theorem 3 will not be as detailed as that of theorem 1 when arguments are very similar (as for the two first points for instance).

## B.1 Theoretical bound on the probability (points *iii* − 1)

**Lemma 5.** *Let us again write* $r_{0,n} = r_0 \left(\frac{n}{\log(n)}\right)^{1/d}$.

*Let us denote by* $\mathcal{A}_n(r_0)$ *the event* $\{\exists x, \mathcal{B}(x, r_{0,n} f(x)^{-1/d}) \cap \mathcal{X}_n = \emptyset\}$.

*Let us denote by* $B$ *the constant* $B = 1 + K_{-1/d}$

*For all* $\varepsilon \leq r_{0,n}$,

$$P(\mathcal{A}_n(r_0)) \leq \nu\left(S, \frac{\varepsilon}{f_M^{1/d}}\right)\left(1-\theta_d\alpha_S[r_{0,n}](r_{0,n}-\varepsilon)^d(1-K_1 r_{0,n})(1-B\varepsilon)^d\right)^n$$

*Proof.* For a deterministic $x$:

$$P(\mathcal{B}(x, r/f(x)^{1/d}) \cap \mathcal{X}_n = \emptyset) \leq \left(1-\theta_d\alpha[r/f_m^{1/d}]r^d(1-rK_1)\right)^n.$$

Let us now cover $S$ with determinitic balls $\mathcal{B}(x_i, \varepsilon/f(x_i)^{1/d})$ with $\varepsilon \leq r$. Let us suppose that $\mathcal{B}(x, \rho/f(x)^{1/d}$ does not contain any observation. There exists $i$ such that $x \in \mathcal{B}(x_i, \varepsilon/f(x_i)^{1/d})$ and so:

- $f(x)^{-1/d} \geq f(x_i)^{-1/d}(1-K_{-1/d}\varepsilon)$, so $\mathcal{B}(x, \rho(1-K_{-1/d}\varepsilon)/f(x_i)^{1/d})$ does not contain any observations.
- Then $\mathcal{B}(x_i, \rho(1-(K_{-1/d}+1)\varepsilon)/f(x_i)^{1/d})$ does not contain any observation.

To finish the proof, replace $\rho$ by $r_{0,n}$ and remark that $\varepsilon/f(x_i)^{1/d} \geq \varepsilon/f_M^{1/d}$. □

**Corollary 9.** *Let* $m_n(=m_n(\mathcal{X}_n))$ *denote* $\min_x f(x)/\hat{f}_{k_n}(x)$
*For all* $\varepsilon_n = o(r_{0,n})$ *there exists* $n_1$ *such that, for all* $n \geq n_1$*:*

$$P(\Delta(\mathcal{X}_n) \geq r_{0,n}) \leq p_n^*(S, r_0, \varepsilon_n, \mathcal{X}_n)$$

*with*

$$p_n^*(S, r_0, \varepsilon_n, \mathcal{X}_n) = \nu\left(S, \frac{\varepsilon_n}{f_M^{1/d}}\right)\left(1-\theta_d\alpha_S[r_{0,n}m_n](r_{0,n}m_n-\varepsilon_n)^d(1-K_1 r_{0,n}m_n)(1-B\varepsilon_n)^d\right)^n.$$

This is a direct consequence of the previous lemma with $n_1$ such that $r_{0,n_1}m_n < r_0(S)$, $K_1 r_{0,n}m_n \leq 1$ and $B\varepsilon_n \leq 1$ . In fact for a completely rigorous proof one need the previous Lemma 5, which proves that there exists $n_2$ such that for all $n \geq n_2$, $m_n \in [0.5, 2]$.

## B.2 Convergence (Points $iii-2$)

**Corollary 10.** *Let us define*

$$\hat{p}_n^+(S, r_0, r_{0,n}\varepsilon_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1-2^{-1}\theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n.$$

*and*

$$\hat{p}_n^-(S, r_0, r_{0,n}\varepsilon_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - \theta_d r_{0,n}^d(1-\varepsilon_n)^d)^n.$$

*If* $\log(n)(m_n^d - 1) \geq 0$ *almost surely and if* $\log(n)(m_n^d - 2) \leq 0$ *almost surely and if* $S$ *is convex with a* $\mathcal{C}^2$ *boundary* $\partial S$*, then* $\hat{q}_n^-(S, r_0, r_{0,n}\varepsilon_n) \leq q_n(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) \leq \hat{q}_n^+(S, r_0, r_{0,n}\varepsilon_n)$ *almost surely.*

The proof requires two parts: the convergence of the covering numbers (for which the proof is very similar to the uniform case) and the proof of the convergence for the second component of the product (wich is obvious).

We are now going to prove that $\log(n)(m_n^d - 1) \geq 0$ almost surely and if $\log(n)(m_n^d - 2) \leq 0$ almost surely .

**Lemma 6.** *Let us denote* $\tilde{m}_n = \min\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d}$.

*If* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $k_n/\log(n)^3 \to +\infty$

*then* $\log(n)(\tilde{m}_n - 1) \geq 0$ *a.s.*

*Remark:* Results on uniform convergence (as in [7]) can not be applied here because they requires uniform continuity on $\mathbb{R}$ for $f$. Obviously our hypotheses on $f$ are not compatible with the uniform continuity.

*Proof.* Let us assume that $\frac{k_n \log(n)^d}{n} \to 0$ and $k_n/\log(n)^3 \to +\infty$. Start to pick a deterministic $x$ in $S$. Let us denote $\rho_n(\varepsilon, x) = (1-\varepsilon/(3\log(n))(k_n/(n\theta_d f(x)))^{1/d}$. Then

$$\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d} \leq (1-\varepsilon/(3\log(n))) \Leftrightarrow \mathcal{B}(x, \rho_n(\varepsilon, x)) \text{ contains most than } k_n \text{ points}$$

The probability for an observation to be in $\mathcal{B}(x, \rho_n(\varepsilon, x))$ is bounded above by:

$$\lambda_n = (1 + K_1\rho_n(\varepsilon, x))\frac{k_n}{n}(1 - \varepsilon/(3\log(n)))^d.$$

Thus

$$P\left(\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d} \leq (1 - \varepsilon/(3\log(n)))\right) = O\left(\Phi\left(\frac{n\lambda_n - k_n}{\sqrt{n\lambda_n}}\right)\right)$$

and since $\frac{k_n \log(n)^d}{n} \to 0$, $n\lambda_n - k_n \sim -d\frac{\varepsilon}{3\log(n)}$
Hence

$$P\left(\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d} \leq (1-\varepsilon/(3\log(n)))\right) = O\left(\frac{\log(n)}{k_n}\exp\left(-\frac{9d^2\varepsilon^2 k_n}{\log(n)^2}\right)\right)$$

Let us now cover $S$ with $\nu(S, (\frac{k_n}{n\theta_d f_M})^{1/d}\frac{\varepsilon}{3\log(n)})$ deterministic balls centered on $x_i$ with radius $a_n = (\frac{k_n}{n\theta_d f_M})^{1/d}\frac{\varepsilon}{3\log(n)}$.

The probability $\lambda'_n$ that there exists a $x_i$ with $\left(\frac{f(x_i)}{\tilde{f}_{k_n}(x_i)}\right)^{1/d} \leq (1-\varepsilon/(3\log(n)))$ satisfies:

$$\lambda'_n = O\left(\frac{n\log(n)^{d+1}}{k_n^{3/2}}\exp\left(-\frac{9d^2\varepsilon^2 k_n}{\log(n)^2}\right)\right).$$

One easily shows that if $k_n/\log(n)^3 \to +\infty$ then $\sum \lambda'_n < +\infty$.
Let us now suppose that there exists $x \in S$ such that:

$$\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d} \leq (1-\varepsilon/(\log(n))).$$

Then $\mathcal{B}(x, \rho_n(3\varepsilon, x))$ contains most than $k_n$ points. There exists $x_i$ such that $x \in \mathcal{B}(x_i, a_n)$. Let us first remark that, as $k_n/n \to 0$, $a_n = o(1/\log(n))$, and there exists $n_0(\varepsilon)$ such that for all $n \geq n_0(\varepsilon)$, $a_n \leq \left(\frac{1-2\varepsilon/(3\log(n))}{1-\varepsilon/\log(n)} - 1\right)\frac{1}{K_{-1/d}}$.
Now $x \in \mathcal{B}(x_i, a_n)$ and so, for $n \geq n_0(\varepsilon)$, $\mathcal{B}(x, \rho_n(3\varepsilon, x)) \subset \mathcal{B}(x, \rho_n(2\varepsilon, x_i)) \subset \mathcal{B}(x_i, \rho_n(\varepsilon, x_i))$.
So $\mathcal{B}(x_i, \rho_n(\varepsilon, x_i))$ contains at least $k_n$ points.
To conclude, $P(\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d} \leq (1-\varepsilon/(\log(n)))) \leq \lambda'_n$ and the Borrel-Cantelli lemma allows us to conclude that

$$\log(n)\,(m_n - 1) \geq 0 \text{ a.s.}$$

□

**Lemma 7.** *Let us denote* $\tilde{M}_n = \max\left(\frac{f(x)}{\tilde{f}_{k_n}(x)}\right)^{1/d}$. *If $S$ has a $\mathcal{C}^2$ boundary, if* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $k_n/\log(n)^3 \to +\infty$

$\log(n)(\tilde{M}_n - 2^{1/d}) \leq 0$ *a.s.*

The proof uses the same kind of steps.

**Lemma 8.** *Let us denote by* $m_n^* = \min_{x\in S}(\tilde{f}(x)/\hat{f}(x))$ *and* $M_n^* = \max_{x\in S}(\tilde{f}(x)/\hat{f}(x))$.
*If* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $\frac{k_n}{\log(n)^{d+1}} \to +\infty$ *then:*
$\log(n)(m_n^* - 1) \xrightarrow{a.s.} 0$
*and* $\log(n)(M_n^* - 1) \xrightarrow{a.s.} 0$

Let us denote $\rho_{kn}(x)$ the distance between $x$ and its $k_n^{\text{the}}$ neighbor. Let $y$ be a point in $Vor(x)$. Then we see that:

$$\rho_{kn}(y) \leq \rho_{kn}(x) + ||x-y|| \text{ and } \rho_{kn}(x) \leq \rho_{kn}(y) + ||x-y||,$$

If for all $x$ in $S$ we let $X(x)$ denote the observation $X_i$ such that $x \in Vor(X_i)$ then

$$\left| \left( \frac{\tilde{f}_{k_n}(x)}{\hat{f}_{k_n}(x)} \right)^{1/d} - 1 \right| \leq \frac{||x - X(x)||}{\tilde{f}_{k_n}^{1/d}(x)}.$$

With the same kind of reasoning as in Corollary 4 there exists a constant $C_1$ such that:

For all $x : ||x - X(x)|| \leq C_1 \left(\frac{\log(n)}{n}\right)^{1/d}$ a.s.

With the same kind of reasoning as in Corollary 5, there exists a constant $C_2$ such that

For all $x : \tilde{f}_{k_n}^{1/d}(x) \geq C_2 \left(\frac{k_n}{n}\right)^{1/d}$ a.s.

The additional condition on $k_n$: $\frac{k_n}{\log(n)^{d+1}} \to +\infty$ ensures the conclusion

**Corollary 11.** *If* $d \geq 2$, *if* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $\frac{k_n}{\log(n)^{d+1}} \to +\infty$ *then*
$\log(n)(m_n - 1) \geq 0$ *a.s. and* $\log(n)(m_n - 2^{1/d}) \leq 0$ *a.s.*

**Corollary 12.** *Let us define*

$$\hat{q}_n^-(S, r_0, r_{0,n}\varepsilon_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - 2^{-1}\theta_d r_{0,n}^d (1-\varepsilon_n)^d)^n$$

*and*

$$\hat{q}_n^+(S, r_0, r_{0,n}\varepsilon_n) = \nu\left(\mathcal{H}(\mathcal{X}_n), \frac{r_{0,n}\varepsilon_n}{f_M^{1/d}}\right)(1 - 2^{-1+1/d}\theta_d r_{0,n}^d (1-\varepsilon_n)^d)^n$$

*If* $\frac{k_n \log(n)^d}{n} \to 0$ *and* $\frac{k_n}{\log(n)^{d+1}} \to +\infty$ *and if* $S$ *is convex with a* $\mathcal{C}^2$ *boundary* $\partial S$ *then* $\hat{q}_n^-(S, r_0, r_{0,n}\varepsilon_n) \leq q_n(S, r_0, r_{0,n}\varepsilon_n, \mathcal{X}_n) \leq \hat{q}_n^+(S, r_0, r_{0,n}\varepsilon_n)$ *almost surely.*

This completes the proof of Point (*iii*) of Theorem 3.