Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2020, Vol. 56, No. 3, 2028–2050
https://doi.org/10.1214/19-AIHP1027

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES
www.imstat.org/aihp

On boundary detection

Catherine Aaron[a] and Alejandro Cholaquidis[b]

[a] Université Clermont Auvergne – LMBP-UMR 6620-CNRS, Clermont-Ferrand, France. E-mail: catherine.aaron@uca.fr
[b] Facultad de Ciencias, Universitad de la Republica, Montevideo, Uruguay. E-mail: acholaquidis@hotmail.com

Received 5 July 2017; revised 24 July 2019; accepted 30 September 2019

Abstract. Given a sample of a random variable supported by a smooth compact manifold $M \subset \mathbb{R}^d$, we propose a test to decide whether the boundary of M is empty or not with no preliminary support estimation. The test statistic is based on the maximal distance between a sample point and the average of its k_n-nearest neighbors. We prove that the level of the test can be estimated, that, with probability one, its power is one for n large enough, and that there exists a consistent decision rule. Heuristics for choosing a convenient value for the k_n parameter and identifying observations close to the boundary are also given. We provide a simulation study of the test.

Résumé. Soit un n-échantillon issus d'une loi supportée par M, une variété compacte suffisament régulière. On propose un test de l'hypothèse nulle $\partial M = \varnothing$ contre l'hypothèse alternative $\partial M \neq \varnothing$ qui ne nécessite pas d'estimation de M préliminaire. La statistique de test est la distance maximale (adéquatement renormalisée) entre une observation et la moyenne de ses k_n-plus proches voisins. On montre que le niveau du test peut être estimé, que sa puissance est 1 lorsque n est suffisament grand et, enfin, qu'il existe une règle de décision consistente. De manière pratique, on propose aussi une heuristique pour le choix de k_n et pour l'indentification des observations proches du bord. Ces résultats sont illustrés par des simulations.

MSC: 62G10; 62H15

Keywords: Geometric inference; Boundary; Test; Nearest-neighbors

1. Introduction

Given an i.i.d. sample $X_1, \dots, X_n$ of X drawn according to an unknown distribution $\mathbb{P}_X$ on $\mathbb{R}^d$, geometric inference deals with the problem of estimating the support, M, of $\mathbb{P}_X$, its boundary, ∂M, or any possible functional of the support, such as the measure of its boundary, for instance. These problems have been widely studied when $\mathbb{P}_X$ is uniformly continuous with respect to Lebesgue measure, i.e. when the support is full dimensional. We refer to [14] and [19] for prior work on support estimation, [15] for a review of support estimation, [17] for estimation of the boundary, [16] for estimation of the measure of the boundary, [7] for estimation of the integrated mean curvature and [3] for the recognition of topological properties having a support estimator homeomorphic to the support. The lower dimensional case (that is, when the support of the distribution is a d'-dimensional manifold with $d' < d$) has recently gained importance due to its connection with non-linear dimensionality reduction techniques (also known as *manifold learning*), as well as *persistent homology*. See [11] For links between data analysis and topological analysis and [13] for one of the later work on persistent diagrams. [30] illustrates the link between topology and unsupervised learning. In [21] a test deciding whether the support lies near a lower dimensional manifold or not is proposed. In [22] or [23] minimax rates for manifold estimation are given under different hypotheses. In [2] non-asymptotic bounds for manifold estimation and related quantities such as tangent spaces and curvature are derived. In these papers the manifolds are supposed without boundary.

Regarding support estimation, it would be natural to think that some of the proposed estimators (in the full dimensional framework) would still be suitable. For instance, in [29], assuming that M is smooth enough, it is proved that for ε small enough, the Devroye–Wise estimator $\hat{M}_\varepsilon = \bigcup_{i=1}^n \mathcal{B}(X_i, \varepsilon)$ deformation retracts to M and therefore the homology of $\hat{M}_\varepsilon$ equals the homology of M (see Proposition 3.1 in [29]). Considering boundary estimation, it is not possible to directly adapt the "full dimensional" methods since in this case the boundary is estimated by the boundary of the estimator. Unfortunately, when the support estimator is full dimensional (which is typically the case, as for example in the Devroye–Wise estimator but also for more recent manifold estimators) this idea is hopeless (see Figure 1).

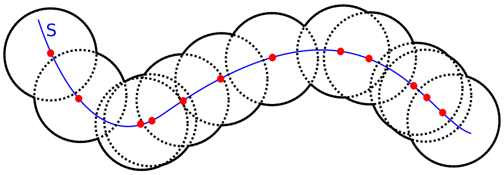

Fig. 1. A one dimensional set M with boundary (the two extremities of the line), sample drawn on M and the associated Devroye–Wise $\hat{M}_r$ estimator of M. Note that $\partial \hat{M}_r$ is far from ∂M.

As far as our knowledge extends, there are only a few d'-dimensional support estimators, see [1] or [28]; they all require support without boundary thus the classical plug-in idea of estimating the boundary of the support using the boundary of an estimator can not be used.

In the lower dimensional case, before trying to estimate the boundary of the support, one has to be able to decide whether it has a boundary or not. The answer provides topological information about the manifold that may be useful. For instance, if there is no boundary, the support estimator proposed in [1] can be used. Moreover, a compact, simply connected manifold without boundary is homomorphic to a sphere, as follows from the well known (and now proved) Poincaré conjecture. When the test decides there is a boundary, one can naturally want to estimate it, or at least estimate the number of its connected components, which is an important topological invariant (for instance the surfaces, i.e. the 2-dimensional manifolds, are topologically determined by their orientability, their Euler characteristic, and the number of the components of the boundary). Testing for the presence of boundary can also be useful as a preliminary step when considering the problem of density estimation on a manifold. Roughly speaking, when the support is smooth enough and has no boundary, a kernel density estimator will work. However, when the support has a boundary, a bias appears near to it. In [8] a correction taking into account the distance to the boundary, also based on a barycenter moving statistics (calculated with a kernel instead of nearest neighbors) is proposed. It allows decreasing the bias but may increase the variance and so should only be performed when necessary, that is, when the support has a boundary.

The aim of the present paper is to provide a statistical test to decide whether the boundary of the support is empty or not and, when there is a boundary, to provide an heuristic method to identify observations close to the boundary and estimate the number of connected components of the boundary.

This paper is organized as follows. In Section 2 we introduce the notation used throughout the paper. In Section 3 we present the test statistic, the associated theoretical results, a way to select suitable values for the parameter k_n and perform a small simulation study. In Section 4 we present an heuristic algorithm that identifies points located close to the boundary and estimates the number of connected components of the boundary. Finally, Section 5 is devoted to the proofs.

2. Notation and geometric framework

If $B \subset \mathbb{R}^d$ is a Borel set, we will denote by $|B|$ its Lebesgue measure and by $\overline{B}$ its closure. Given a set A on a topological space, the interior of A with respect to the underlying topology is denoted by $\mathring{A}$. The k-dimensional closed ball of radius ε centred at x will be denoted by $\mathcal{B}_k(x, \varepsilon) \subset \mathbb{R}^d$ (when $k = d$ the index will be omitted) and its Lebesgue measure will be denoted by $\sigma_k = |\mathcal{B}_k(x, 1)|$. When $A = (a_{ij})$ $(i = 1, \dots, m,\ j = 1, \dots, n)$ is a matrix, we will write, $\|A\|$ the euclidean norm of A, $\|A\|_\infty = \max_{i,j} |a_{ij}|$ and $\|A\|_{\text{op}}$ the operator norm of A. The transpose of A will be denoted A'. For the case $n = m$, we will write $\det(A)$ and $\operatorname{tr}(A)$ for the determinant and trace of A, respectively.

Given a $\mathcal{C}^2$ function f, $\vec{\nabla} f$ denotes its gradient and H_f its Hessian matrix. We will denote by $\Psi_{d'}(t)$ the cumulative distribution function of a $\chi^2(d')$ distribution and $F_{d'}(t) = 1 - \Psi_{d'}(t)$.

In what follows $M \subset \mathbb{R}^d$ is a d'-dimensional compact manifold of class $\mathcal{C}^2$ (also called a d'-regular surface of class $\mathcal{C}^2$). We will consider the Riemannian metric on M inherited from $\mathbb{R}^d$. When M has a boundary, as a manifold, it will be denoted by ∂M. For $x \in M$, $T_x M$ denotes the tangent space at x and φ_x the orthogonal projection on the affine tangent space $x + T_x M$. When M is orientable it has a unique associated volume form ω such that $\omega(e_1, \dots, e_{d'}) = 1$ for all oriented orthonormal bases $e_1, \dots, e_{d'}$ of $T_x M$. Then if $g : M \to \mathbb{R}$ is a density function, we can define a new measure $\mu(B) = \int_B g\, d\omega$, where $B \subset M$ is a Borel set. Since we will only be interested in measures, which can be defined even if the manifold is not orientable, although in a slightly less intuitive way, the orientability hypothesis will be dropped in the following.

3. The test

3.1. *Hypotheses, test statistics and main results*

Throughout this paper, $X_1,\dots,X_n$ is an i.i.d. sample of a random variable X whose probability distribution, $\mathbb{P}_X$, fulfills condition P, and the sequence (k_n) fulfills condition K:

P. A probability distribution $\mathbb{P}_X$ fulfills condition P if there exists a compact, path connected d'-dimensional manifold of class $\mathcal{C}^2$ M and a density function f such that:
1. ∂M is either empty or of class $\mathcal{C}^2$,
2. for all $x \in M$, $f(x) \geq f_0 > 0$, f is Lipschitz continuous with constant K_f, and, for all measurable $A \subset M$, $\mathbb{P}_X(A) = \int_A f\omega$. In the following $f_1 = \max_{x\in M} f(x)$.

K. A sequence $\{k_n\}_n \subset \mathbb{R}$ fulfills condition K if $k_n/n^{1/(d'+1)} \to 0$ and if $k_n/(\ln(n))^4 \to \infty$ when $d' > 1$ and if $k_n/\sqrt{n \ln n} \to +\infty$ when $d' = 1$.

Definition 1. Given an i.i.d. sample $X_1,\dots,X_n$ of a random row vector X with support $M \subset \mathbb{R}^d$, where M is a d'-dimensional manifold with $d' \leq d$, we will denote by $X_{j(i)}$ the j-nearest neighbor of X_i. For a given sequence of positive integers k_n, let us define, for $i = 1,\dots,n$,

$$r_{i,k_n} = \|X_i - X_{k_n(i)}\|; \qquad r_n = \max_{1\leq i\leq n} r_{i,k_n}; \qquad \mathcal{X}_{i,k_n} = \begin{pmatrix} X_{1(i)} - X_i \\ \vdots \\ X_{k_n(i)} - X_i \end{pmatrix}; \qquad \hat{S}_{i,k_n} = \frac{1}{k_n}(\mathcal{X}_{i,k_n})(\mathcal{X}_{i,k_n})',$$

where $X_{j(i)} - X_i$ is a row vector, for all $j = 1,\dots,k_n$. Consider Q_{i,k_n} the d'-dimensional space spanned by the d' eigenvectors of $\hat{S}_{i,k_n}$ associated to its d' largest eigenvalues. Let $X^*_{k(i)}$ be the normal projection of $X_{k(i)} - X_i$ on Q_{i,k_n} and $\overline{X}_{k_n,i} = \frac{1}{k_n}\sum_{k=1}^{k_n} X^*_{k(i)}$.

Define $\delta_{i,k_n} = \frac{(d'+2)k_n}{r^2_{i,k_n}}\|\overline{X}_{k_n,i}\|^2$, for $i = 1,\dots,n$. Then the proposed test statistic is

$$\Delta_{n,k_n} = \max_{1\leq i\leq n} \delta_{i,k_n}.$$

We will now explain the heuristic behind the test we will propose. It will be proved that, under conditions P and K we have $r_n \xrightarrow{\text{a.s.}} 0$ (using that the density is bounded from below and the classic condition $k_n/n \to 0$ as in [27] where the concept of nearest neighbors was introduced). Consider an observation X_{i_0} such that $d(X_{i_0}, \partial M) \geq r_{i_0,k_n}$. The regularity of the manifold and the continuity of the density given by condition P will imply that the sample $\{r^{-1}_{i_0,k_n} X^*_{1(i_0)}, \dots, r^{-1}_{i_0,k_n} X^*_{k_n(i_0)}\}$ "converges" to an uniform sample on $\mathcal{B}_{d'}(0,1)$, and then $\|\overline{X}_{k_n,i_0}\| r^{-1}_{i_0,k_n} \xrightarrow{\text{a.s.}} 0$. It will also be proved that $\delta_{i_0,k_n} \longrightarrow \chi^2(d')$ in distribution. If $\partial M = \varnothing$, all the observations satisfy $d(X_i, \partial M) \geq r_{i,k_n}$. Even though the $\{\delta_{i,k_n}\}_i$ are not independent, we will obtain an asymptotic result for Δ_{n,k_n} that involves the $\chi^2(d')$ distribution. If $\partial M \neq \varnothing$, condition P (the regularity of the boundary and the fact that the density is bounded from below) allows us to (lower) bound the probability that X belongs to a neighborhood of the boundary. With this bound we can ensure a.s. the existence of an observation X_{i_0} with $d(X_{i_0}, \partial M) = O(\ln n/n)$, and then condition K ($k_n/(\ln n)^4 \to +\infty$) ensures that $d(X_{i_0}, \partial M) \ll r_{i_0,k_n}$. Note that this condition is stronger than the usual $k_n \to +\infty$ as in [27]. The sample $\{r^{-1}_{i_0,k_n} X^*_{1(i_0)}, \dots, r^{-1}_{i_0,k_n} X^*_{k_n(i_0)}\}$ thus "looks like" an uniform sample on a half ball and $\|\overline{X}_{k_n,i_0}\| r^{-1}_{i_0,k_n} \xrightarrow{\text{a.s.}} \alpha_{d'} > 0$. The asymptotic behavior of the test statistic is given in the following four theorems. The first theorem provides a bound for the level when testing H_0: $\partial M = \varnothing$ versus H_1: $\partial M \neq \varnothing$ using the test statistic Δ_{n,k_n} and rejection region $\{\Delta_{n,k_n} \geq t_n\}$ for some suitable t_n. The second theorem states that, with probability one, the power of the test is one for n large enough. The third theorem provides a consistent decision rule.

Theorem 1. *Let k_n be a sequence fulfilling condition K. Assume that $X_1,\dots,X_n$ is an i.i.d. sample drawn according to an unknown distribution $\mathbb{P}_X$ which fulfills condition P. The test*

$$\begin{cases} H_0\colon & \partial M = \varnothing, \\ H_1\colon & \partial M \neq \varnothing \end{cases} \tag{1}$$

with the rejection zone

$$W_n = \bigl\{\Delta_{n,k_n} \geq F^{-1}_{d'}\bigl(9\alpha/\bigl(2e^3 n\bigr)\bigr)\bigr\}, \tag{2}$$

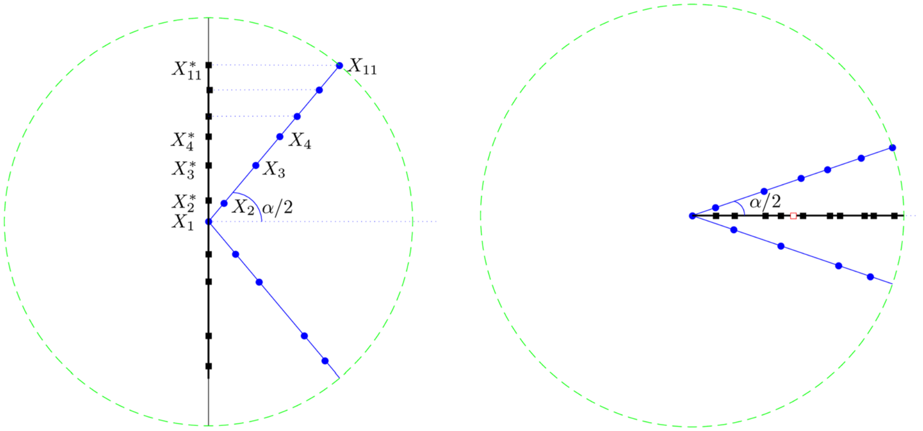

Fig. 2. Behaviour when there is an angle at X_1. Blue: manifold and observations, black: estimated tangent space and projections. Red: mean of the projections, dashed green: the sphere of radius $\|X_1 - X_{11}\|$, centred at X_1. Left when $\alpha > \pi/2$, the tangent space is "correct" but not the normalization radius. Right, when $\alpha < \pi/2$, the tangent space is not at all the expected one.

satisfies $\mathbb{P}_{H_0}(W_n) \leq \alpha + o(1)$.

Theorem 2. *Let k_n be a sequence fulfilling condition K. Assume that $X_1, \dots, X_n$ is an i.i.d. sample drawn according to an unknown distribution $\mathbb{P}_X$ which fulfills condition P. The test* (1) *with rejection zone* (2) *has power* 1 *for n large enough.*

Theorem 3. *Let k_n be a sequence fulfilling condition K. Assume that $X_1, \dots, X_n$ is an i.i.d. sample drawn according to an unknown distribution $\mathbb{P}_X$ which fulfills condition P. For all $\lambda > 6$, the decision rule $\partial M = \varnothing$ if, and only if, $\Delta_{n,k_n} \leq \lambda \ln n$ is consistent for n large enough.*

3.2. *Discussion of the hypotheses*

The two main hypotheses in this paper consist in the smoothness of the support and the continuity of the density. These two hypotheses can not be weakened and we now exhibit examples of manifolds without boundary for which our test fails, the first one being not smooth enough and the second one with a discontinuous density.

Suppose that $d = 2$, $d' = 1$, X is uniformly drawn on M that has no boundary, but there exists a corner at the origin with an angle α (see Figure 2). Introduce $S = \frac{1}{r}\mathbb{E}YY'$ where $Y = X|\{\|X\| \leq r\}$. Then a short calculation gives

$$S = \frac{\cos^2(\alpha/2)}{3}\begin{pmatrix} 1 & 0 \\ 0 & \tan(\alpha/2)^2 \end{pmatrix}.$$

- If $\alpha > \pi/2$, the projection direction is "the vertical one", that can be considered as a "correct tangent space". The only problem is that we should rescale by $\|X_i^* - X_{k_n(i)}^*\|$ instead of $r_{i,k_n} = \|X_i^* - X_{k_n(i)}^*\|$.
- If $\alpha < \pi/2$, the projection direction is "the horizontal one", this fails in recognizing the tangent space, and induces a barycentre moving as in the boundary case and the test will decide falsely that there is a boundary.

The continuity of the density is also necessary: if this is not the case, we may reject H_0 for any support, with or without boundary. In order to see this, consider the circular support $M = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$ with a "density" $1/(4\pi)$ when $x \leq 0$ and $3/(4\pi)$ when $x > 0$. In this case it can be proved that $\Delta_{n,k_n}/k_n \to 1/2$ (considering points located near the discontinuity points), which also corresponds to a "boundary-type" behavior.

The other hypotheses can be weakened by pre-processing the data. For instance, the intrinsic dimension can be estimated by several existing methods (see [9] for a consitent method or [10] for a review). Observe that this is costless in terms of sample size dependency. Even more, there are minimax bounds for dimension estimation (see [26]).

With our approach the assumption that there is no noise, i.e. that the dimension of the support is lower than the dimension of the ambient space, can not be replaced by a noisy model in which the support is "around" a lower dimensional manifold. However, in such a case, performing a preliminary manifold estimation before running our test (see for instance [22] or [4]) can be used to overcome this problem. Even if the manifold estimator is not a d'-dimensional manifold, we may expect that by imposing stronger conditions on the sequence k_n, our approach can work.

Even if, due to [24,32] and [25] we can avoid assuming the compactness of the support for some geometrical inference problem we are not sure that it is possible for the boundary detection case.

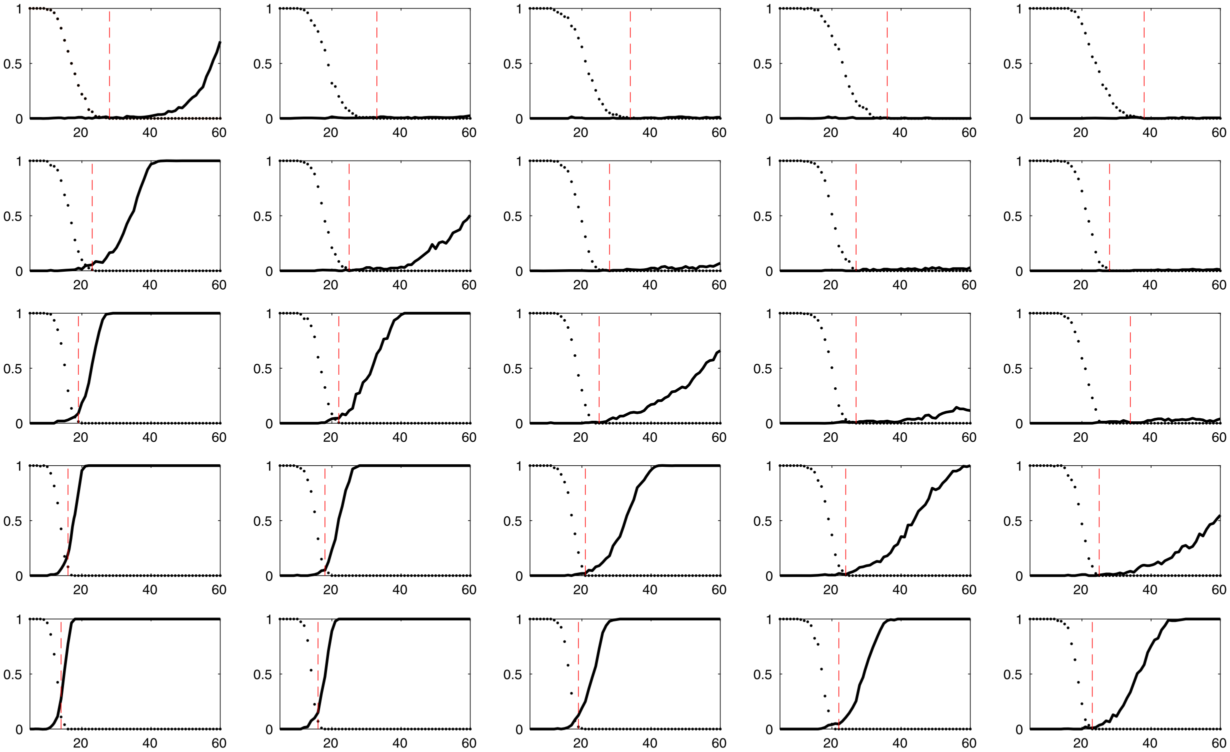

Fig. 3. e_0 (dashed) and e_1 (plain) for different values of n and d' (from left to right, increasing values of n in $\{100; 200; 500; 1000; 2000\}$ and from top to bottom increasing values of d' in $\{1; 2; 3; 4; 5\}$), the chosen value for k_n is indicated by the vertical dashed line.

Lastly, the $\mathcal{C}^2$ smoothness of the whole boundary is not necessary, the existence of a compact $\mathcal{C}^2$ subset of ∂M is enough. When the manifold has a boundary, the hypothesis $f(x) > 0$ on M can also be weakened to the usual condition $f(x) \geq ad(x, \partial S)^b$ (for some positive constants a and b), which change only the convergence rates.

3.3. *Numerical simulations and k_n calibration*

In this section we are going to explain intuitively the underlying idea regarding the parameter k_n. We think that, at least asymptotically, the “optimal” choice of k_n should only depend on d'. Other parameters, such as density variations, or the curvature of the manifold, should slow down the convergence rate. That is, we believe that the quality of p-value estimation asymptotically behaves like $C_{f,M,d}g(n, d', k_n')$. Intuitively, we have that

1. Under H_0:
   (a) if we let $U_1, \dots, U_k$ be an uniform random sample on the d'-dimensional unit ball, $\overline{U}_{k_n} = (1/k_n)\sum_{i=1}^{k_n} U_i$ and $\delta_k^U = (d'+2)k_n\|\overline{U}_{k_n}\|^2$. Then k_n should be large enough to ensure that $\delta_{k_n}^U$ is “close enough”, in law, to a $\chi^2(d')$ distribution.
   (b) On the other hand, k_n should be small enough so that, locally, the nearest neighbors to every sample point behave like an uniform sample on a d'-dimensional ball.

   As can be seen in Figure 3 and Table 1, $k_n \geq 10$ is sufficient to guarantee 1(a). Regarding 1(b), the greater the curvature of M, or the more variations in the density, the smaller the k_n should be (see Figure 3). When n is large enough, this still provides a large interval of acceptable values for k_n.
2. Under H_1:
   (a) k_n should be large enough to ensure the existence of an observation X_{i_0} such that its k_n nearest neighbors “look” like an uniform sample on a half ball. More precisely, k_n should be large enough to guarantee that $r_{i_0,k_n} \gg d(X_{i_0}, \partial M)$.
   (b) On the contrary, k_n should be small enough so that, locally, the nearest neighbors “look” like an uniform sample on a subsets of the d'-dimensional ball.

Part 2(b) is analogous to part 1(b) and does not add more constraints on k_n. Considering 2(a), the (only) important parameter is the $(d'-1)$ measure of the boundary. The smaller this measure is, the larger k_n should be. Conversely, if the

Table 1
Proposed values for k_n

	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
$d' = 1$	30	30	40	40	40
$d' = 2$	24	26	28	28	28
$d' = 2$	20	24	26	26	26
$d' = 4$	18	22	22	24	26
$d' = 5$	18	18	20	22	24

Table 2
For different samples, the % of times where H_0 is rejected when there is no boundary

	$n = 100$	$n = 200$	$n = 500$	$n = 10^3$	$n = 2000$
S_1	0.96%	0.53%	0.37%	0.41%	0.33%
S_2	4.01%	1.39%	0.71%	0.38%	0.29%
S_3	12.09%	4.81%	1.63%	0.9%	0.95%
S_4	20.93%	7.8%	3.08%	2.06%	1.06%
Trefoil	100%	99.93%	12.87%	2.05%	0%
Torus	100%	99.61%	27.46%	4.69%	0%

Table 3
For different samples, the % of times where H_0 is accepted when there is a boundary

	$n = 100$	$n = 200$	$n = 500$	$n = 10^3$	$n = 2000$
S_1^+	0%	0%	0%	0%	0%
S_2^+	0%	0%	0%	0%	0%
S_3^+	0%	0%	0%	0%	0%
S_4^+	0%	0%	0%	0%	0%
Spire	0.5%	3.5%	1.5%	2%	5%
Moebius	0%	0%	0%	0%	0%

measure of the boundary is large, we will have more observations close to it, so the condition $r_{i,k_n} \gg d(X_i, \partial M)$ will be fulfilled. Due to the well known curse of dimensionality, for small values of n and for high dimensions, we have more observations located close to the boundary, which has the following unexpected effect: k_n decreases with the dimension.

All this is illustrated in two simulation studies, first for $S_{d'} = \{x \in \mathbb{R}^{d'+1}, \|x\| = 1\}$ the d'-dimensional sphere and $S_{d'}^+ = \{x = (x_1, \dots, x_{d'+1}), \|x\| = 1, x_1 \geq 0\}$ the d'-dimensional half sphere. Consider the test with a level $\alpha = 5\%$. For a given $d' \in \{1, 2, 3, 4, 5\}$ and a given $n \in \{100, 200, 500, 1000, 2000\}$ we estimate $e_0(k) = \mathbb{P}_{H_0}(\Delta_{n,k} \geq F_{d'}^{-1}(9\alpha/(2e^3 n)))$ as the percentage of wrong decisions for samples of size n, uniformly drawn on $S_{d'}$ and $e_1(k) = \mathbb{P}_{H_1}(\Delta_{n,k} \leq F_{d'}^{-1}(9\alpha/(2e^3 n)))$ as the percentage of wrong decisions for samples of size n, uniformly drawn on $S_{d'}^+$. Each time the percentages are estimated with 200 repetitions of the experiment. The results are presented in Figure 3. For $d' \in \{1, 2, 3\}$ we observe that e_0 can be neglected (for $k \in [10, 60]$) when $n \geq N_{d'}$ (with $N_1 = 200$, $N_2 = 500$ and $N_3 = 1000$). We propose the following criteria to choose k_n.

1. If $\{k \text{ such that } e_0(k) + e_1(k) \leq 0.01\} \neq \varnothing$ then $k_n = \min\{k \text{ such that } e_0(k) + e_1(k) \leq 0.01\}$.
2. If $\{k \text{ such that } e_0(k) + e_1(k) \leq 0.01\} = \varnothing$ then choose $k_n = \operatorname{argmin}_k(e_0(k) + e_1(k))$

The values of k_n are given in Table 1. They are also presented in Figure 3.

We also considered the trefoil knot, a torus, a spire and a Moebius ring. The percentage of times (over 50,000 replicates for each manifold and sample size) where H_0 is rejected is shown in Table 2 when there is no boundary. In Table 3 it is shown the percentage of times (over 50,000 replicates) where H_0 is accepted when there is a boundary. As can be seen, the test almost never fails under H_1, which is not surprising considering the way we chose the sequence k_n. Under H_0 the convergence to an error rate inferior to 5% depends on the dimension d' and the curvature of the manifold.

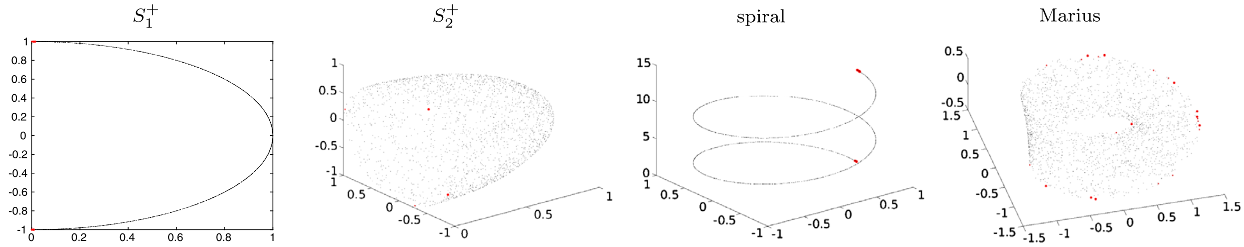

Fig. 4. Some examples for support with boundary, the associated sample ($n = 2000$) is in black, and points that are identified as "close to the boundary" are in red, the size of the points depending of the associated α, the boundary identification starts with $\alpha = 20\%$ (small red points), and finish with $\alpha = 5\%$ (larger red points).

4. Empirical detection of points close to the boundary and estimation of the number of its connected components

A natural second step after deciding that the support has a boundary is to estimate it, or at least identify observations "close" to it. To get an insight into the topological properties of the boundary, a third step could be to estimate the number of its connected components. In this section we will tackle empirically both problems.

4.1. *Detection of "boundary observations"*

Theorem 1 suggests selecting $\{X_i : \delta_{i,k_n} \geq F_{d'}^{-1}(9\alpha/(2ne^3))\}$ as "boundary observations". However, when applying this method with the previously proposed values for k_n, it identifies "too few" boundary observations for $d' = 2$. We think that this is due to the $2e^3/9$ factor, which deals with the problem of the maximum of dependant variables but, for a given observation, underestimates probability to be close to the boundary. Allowing "large" values for α is not sufficient to overcome this problem, as it can be observed in Figure 4 where $\alpha = 20\%$ is considered. For this reason we will adapt, using tangent spaces, the method given in [4] to detect "boundary balls".

In [4], M is d-dimensional and boundary observations are identified as those with large Voronoi cells (recall that $\text{Vor}(X_i) = \{x : \|x - X_i\| \leq \|x - X_j\| \ \forall j\}$). More precisely, define $\rho_i = \sup\{\|x - X_i\| : x \in \text{Vor}(X_i)\}$. Then boundary observations are those X_i such that $\rho_i \geq \varepsilon_n$, where ε_n is a smoothing parameter. Two different ideas inspired this characterization. The first one was to consider the Devroye–Wise estimator of the support $\hat{S}_{\varepsilon_n} = \bigcup_i \mathcal{B}(X_i, \varepsilon_n)$ (see [14] or [19]), in which case it is quite intuitive that sample points X_i fulfilling $\mathcal{B}(X_i, \varepsilon_n) \cap \partial\hat{S} \neq \varnothing$ are close to the boundary. The second one was to look for observations in $\partial C_{\varepsilon_n}$, the ε_n-convex hull of the sample (see [12]). These two approaches are in fact the same, the boundary observations can be easily identified considering the size of the Voronoi cells (see Figure 5 left side). This can be explained as follows. Choose $\varepsilon_n > d_H(\{X_1, \dots, X_n\}, M)$, where d_H denotes the Hausdorff distance, suppose that there exists $x \in \text{Vor}(X_i)$ with $\|x - X_i\| > \varepsilon_n$, then $x \notin M$. Using the fact that $X_i \in M$, it follows that there exists $t \in [X_i, x] \cap \partial M$ (because M is d-dimensional) and then $d(X_i, \partial M) \leq \varepsilon_n$ (when ∂M is smooth enough we have an even better inequality).

When M has dimension $d' < d$, every observation has a large Voronoi cell (this can be observed considering directions normal to M, see Figure 5 right side). Then the previously suggested method requires a small adjustment, naturally done using projections on the tangent space, which can be estimated via local PCA. The idea being to locally lie in the full dimensional case. More precisely, recalling that Q_{i,k_n} denotes estimation via local PCA of the tangent space at X_i, the tangential boundary observations are defined as follows.

Definition 2. X_i is a (k_n, ε_n)-tangential boundary observation if

$$\rho_i \equiv \sup\Big\{\|x\| : x \in Q_{i,k_n} \text{ and } \|x\| \leq \big\|x - X^*_{j(i)}\big\|, \forall 1 \leq j \leq k_n\Big\} \geq \varepsilon_n.$$

As in [4], we suggest choosing $\varepsilon_n = 2\max_i \min_j \|X_i - X_j\|$.

4.2. *Building a "boundary graph"*

Once we have identified $\mathcal{Y}_m = \{Y_1, \dots, Y_m\}$ as the set of the centers of the (k_n, ε_n)-tangential boundary balls, a natural second step is how to estimate ∂M. In this respect, we think that the tangential weighted Delaunay complex (see [1]) should work. To prove this is far beyond the scope of this paper. Here, we propose, as an initial step, an estimator based on a graph with vertices $\mathcal{Y}_m$, building edges between the vertices in such a way that the resulting graph captures the "shape" of the boundary. To do this, we are going to "connect" each Y_i to those Y_j such that $\|Y_i - Y_j\| \leq R_i$. As usual,

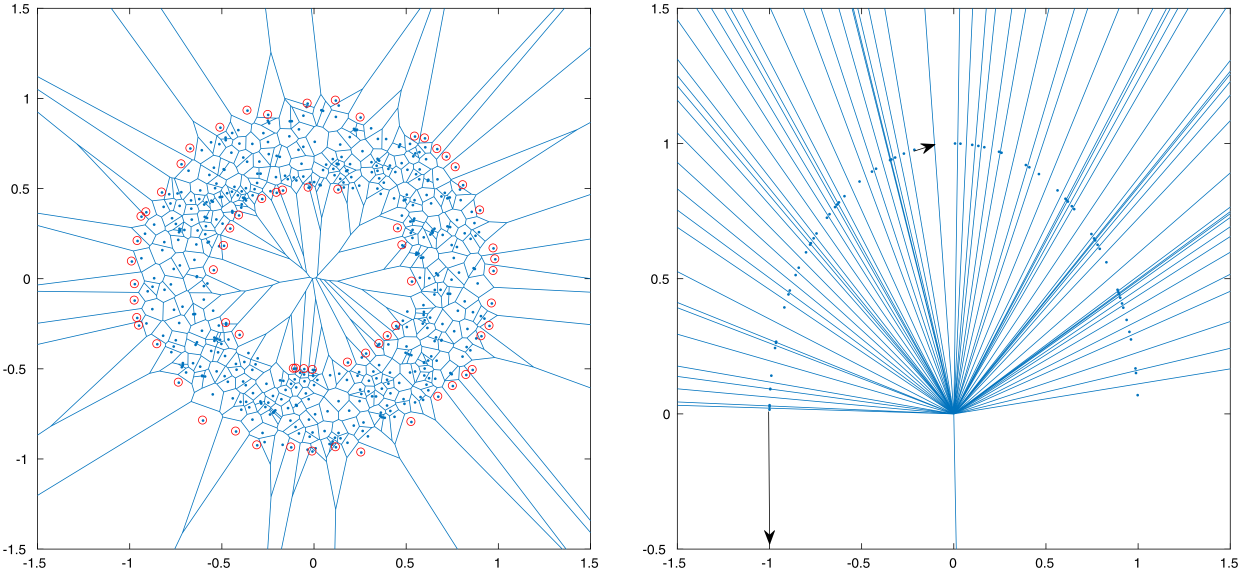

Fig. 5. Left side, $d = d' = 2$, 500 points drawn on $M = \mathcal{B}(0, 1) \setminus \mathcal{B}(0, 0.5)$, observations and Voronoi cells are presented. Observations with an associated radius larger than 0.3 are highlighted. Right side, $d = 2$, $d' = 1$, 70 points uniformly drawn on a half circle, all the Voronoi cells are large, but considering the tangential direction (highlighted by arrows at two points) helps to identify boundary observations.

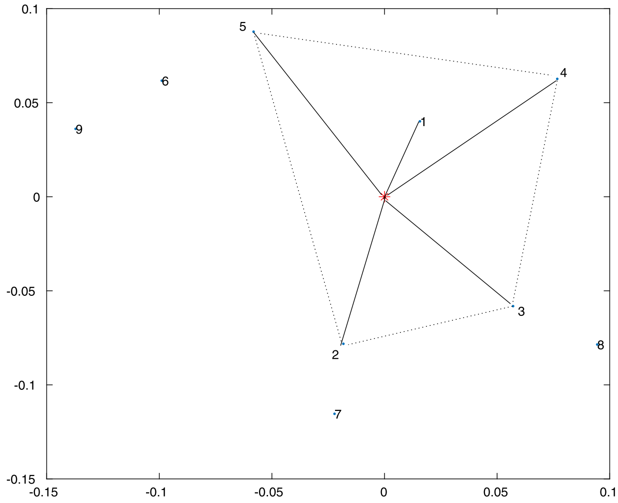

Fig. 6. Consider the point $(0, 0)$ (the red $*$) in $\mathcal{Y}$ and its 9 nearest neighbors. We will connect $(0, 0)$ to its 5 nearest neighbors.

the choice of R_i depends on striking a balance. On the one hand, R_i should be small enough to connect a point only with its neighbors. On the other hand, R_i should be large enough to allow capturing the global structure of ∂M. The idea for selecting R_i is based on the following. As ∂M is a $(d' - 1)$-dimensional manifold without boundary, then for all $x \in \partial M$, for r small enough, the projection onto the space tangent to ∂M at the point x, $\pi_x(\mathcal{B}(x, r) \cap \partial M)$, should be close to $\mathcal{B}(x, r) \cap T_x \partial M$. As a plug-in version we introduce

1. $\mathcal{Z}_{i,r} = \{Y_j : \|Y_j - Y_i\| \le r\}$, the empirical neighborhood of Y_i,
2. $\hat{\pi}_i(\mathcal{Z}_{i,r})$ the orthogonal projection onto the $(d' - 1)$ first axis of a PCA based on $\mathcal{Z}_{i,r}$.

Naturally $\hat{\pi}_i(\mathcal{Z}_{i,r})$ estimates $\pi_x(\mathcal{B}(x, r) \cap \partial M)$ and so should be close to a $(d' - 1)$-dimensional ball centred at Y_i. We quantify this closeness as follows. We say that r is large enough for i if Y_i is in $\mathring{H}_i$ where H_i is the convex hull of $\hat{\pi}_i(\mathcal{Z}_{R_i})$.

Lastly, for all $i = 1, \ldots, n$, choose R_i as the smallest value r that is large enough for i. This is illustrated in Figure 6.

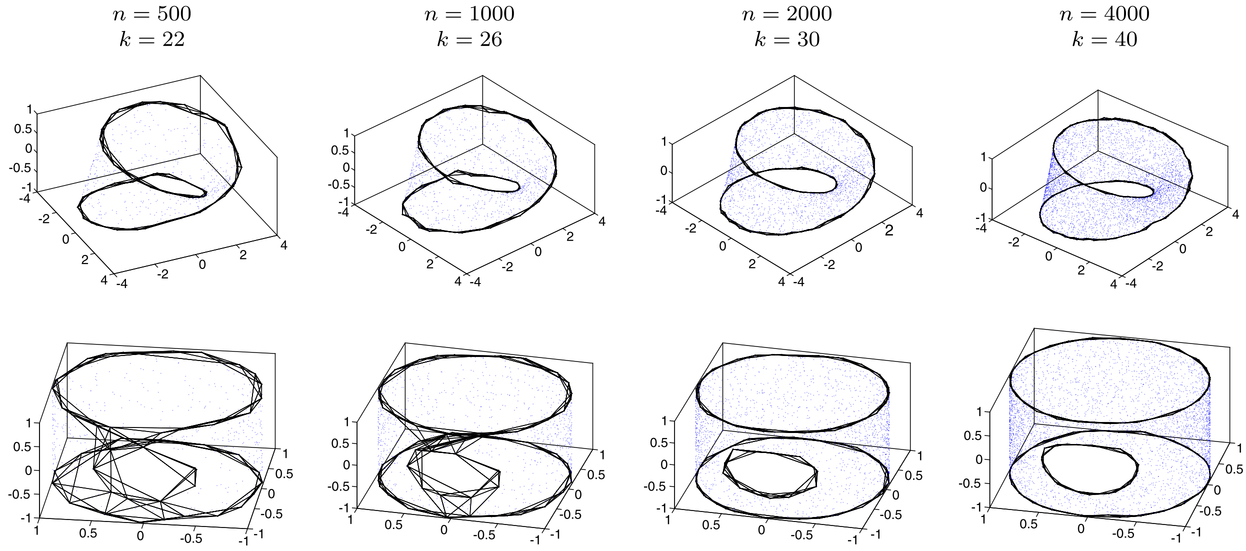

Fig. 7. Boundary ball detection and associated graph for different sample sizes. In the first row the Moebius ring and in the second the truncated cylinder with a hole in a cap. Observations are represented as blue dots while boundary centres are large black dots. The graph is represented by black lines.

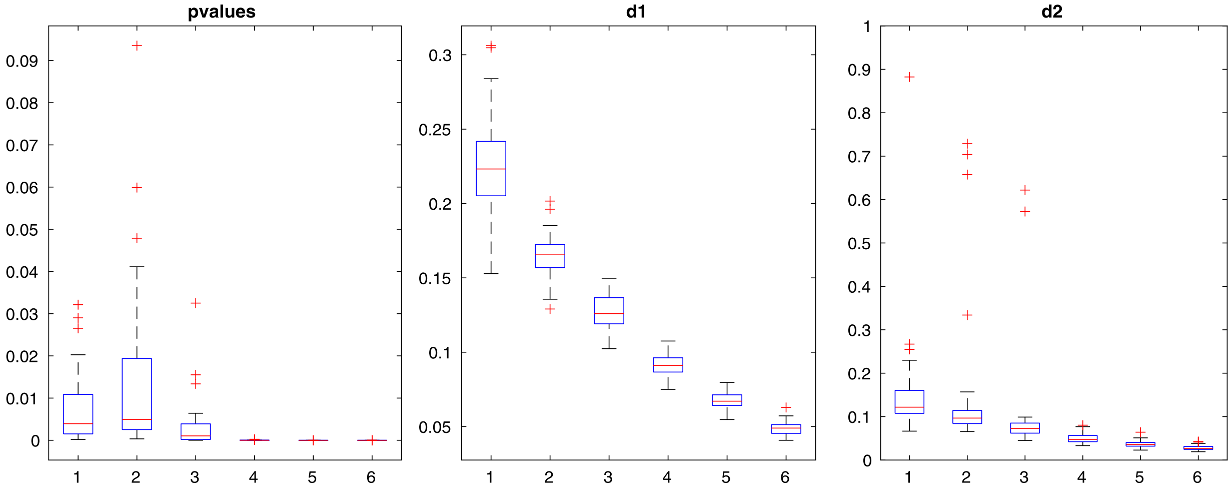

Fig. 8. $d = 3$, on abscissa $1 : (n = 500, k = 25)$, $2 : (n = 1000, k = 25)$, $3 : (n = 2000, k = 30)$, $4 : (n = 4000, k = 40)$, $5 : (n = 8000, k = 50)$, $6 : (n = 16{,}000, k = 50)$.

4.3. *Some experiments*

To illustrate the procedure introduced we consider the Moebius ring and the truncated cylinder with a hole in a cap, (see Figure 7). Both are 2-dimensional sub-manifolds of $\mathbb{R}^3$. The boundary of the first one has one connected component while the boundary of the second one has three.

As expected, in the cylinder the sample size required to have a "coherent" graph is higher.

Second, we consider uniform draws of sizes $n \in \{500, 1000, 2000, 4000, 8000, 16{,}000\}$ on the $(d-1)$-dimensional half sphere $\{x_1^2 + \cdots + x_d^2 = 1, x_d \geq 0\} \subset \mathbb{R}^d$ for $d = \{3, 4, 5\}$. Define $d_1 = \max_{x \in \partial M} \min_i \|x - Y_i\|$ and $d_2 = \max_i \min_{x \in \partial M} \|x - Y_i\|$. They are estimated via a Monte Carlo method, drawing 50,000 points on ∂M. For each value of n and d, the box plot over 50 repetitions of the p-values of the test and the estimations of d_1 and d_2 are shown in Figures 8, 9 and 10.

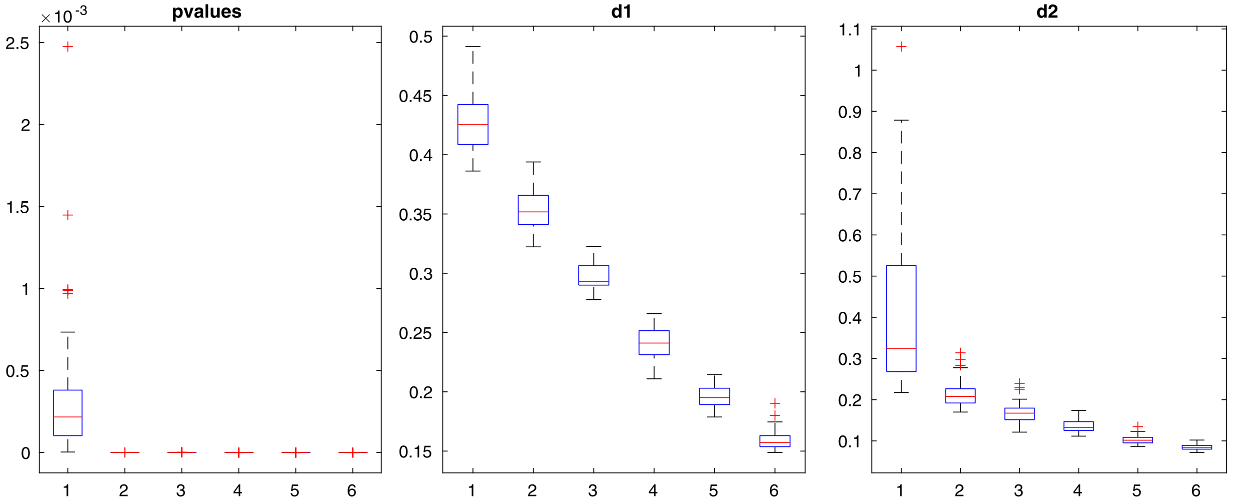

Fig. 9. $d = 4$, on abscissa $1 : (n = 500, k = 30)$, $2 : (n = 1000, k = 50)$, $3 : (n = 2000, k = 50)$, $4 : (n = 4000, k = 60)$, $5 : (n = 8000, k = 70)$, $6 : (n = 16{,}000, k = 70)$.

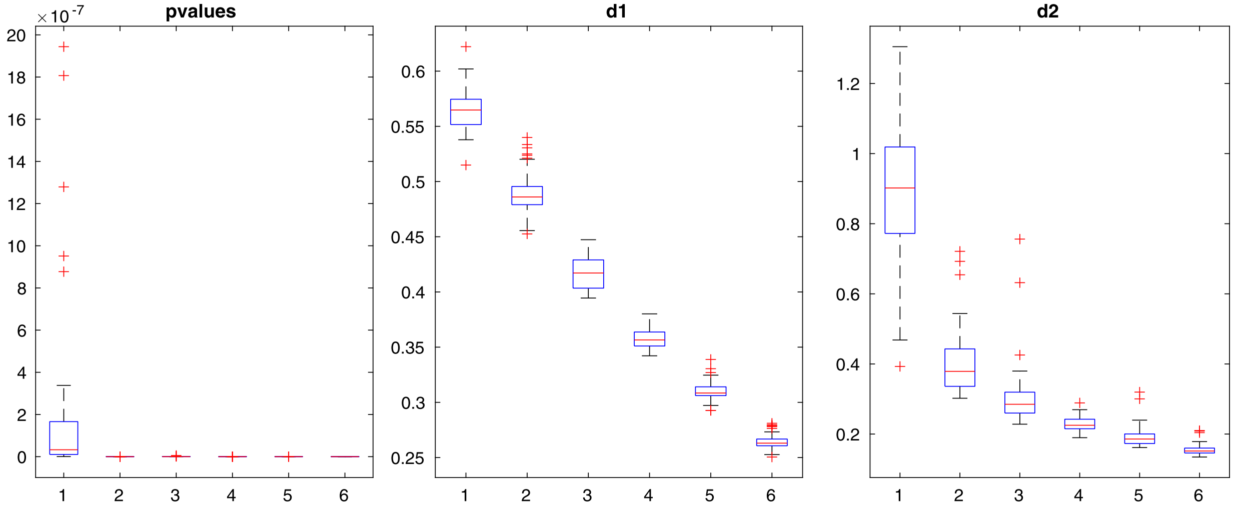

Fig. 10. $d = 5$, on abscissa $1 : (n = 500, k = 50)$, $2 : (n = 1000, k = 70)$, $3 : (n = 2000, k = 80)$, $4 : (n = 4000, k = 90)$, $5 : (n = 8000, k = 100)$, $6 : (n = 16{,}000, k = 100)$.

5. Proofs

5.1. *Proofs under H_0 $(\partial M = \varnothing)$*

In this section we give the details of the proofs when $\partial M \neq \varnothing$. First we prove that the empirical distribution of the δ_i converges to a χ^2 distribution, then we prove that the proposed test has, asymptotically, level α (which proves Theorem 1).

For ease of writing, in what follows, a denotes a general constant that may have different values and should be understood as "there exists an uniform constant such that…".

First we introduce $\xi_n^* \equiv (\ln(n)/n)^{1/2d'}$, $\xi_n^\blacktriangledown \equiv (k_n/n)^{1/d'}$, $\xi_n^\circ \equiv \sqrt{\ln(n)/k_n}$, $\rho_n = \max(\xi_n^*, \xi_n^\blacktriangledown)$ and $\xi_n \equiv \max\{\xi_n^*, \xi_n^\blacktriangledown, \xi_n^\circ\}$. Observe that by condition K, $(\ln n)^2 \xi_n \to 0$, then

1. the maximum distance from an observation to its k_nth nearest neighbor converges (almost surely) to 0, i.e. $r_n \to 0$ (this is a consequence of Lemma 1);
2. the local PCA step converges to the projection onto the tangent space (the rate, ξ_n°, is given in Lemma 3).

For a given $i \in \{1, \dots, n\}$, denote by $x_0 \equiv X_i$, and by $x_1, \dots, x_{k_n}$ the k_n-nearest neighbors of X_i. Recall that $r_{i,k_n} = \max_{1 \le j \le k_n} \|x_0 - x_j\|$ (see Definition 1). For all $j \in \{1, \dots, k_n\}$, write x_j^* for the local PCA projection of $x_j - x_0$, and y_j for the (orthogonal) projection onto the tangent space $T_{x_0}M$ (at the point x_0) of $x_j - x_0$.

Write $\delta_i = (d'+2)k_n r_{i,k_n}^{-2}\|(1/k_n)\sum_j x_j^*\|^2$ and $\delta_i^Y = (d'+2)k_n r_{i,k_n}^{-2}\|(1/k_n)\sum_j y_j\|^2$.

By Lemma 3, for all $i \in \{1,\dots,n\}$ we have, with probability greater than $1-n^{-6}$,

$$\delta_i = \frac{(d'+2)k_n}{r_{i,k_n}^2}\left\|\frac{1}{k_n}\sum_j y_j + E_{i,n}\left(\frac{1}{k_n}\sum_j y_j\right) + \frac{1}{k_n}\sum_j e_j\right\|^2$$

with $\|E_{i,n}\|_{\mathrm{op}} \le a\xi_n$ and $\|e_j\| \le a\xi_n\|y_j\|^2$.

From where it follows that,

$$\begin{aligned}\delta_i = {}& \delta_i^Y + \frac{(d'+2)k_n}{r_{i,k_n}^2}\left\|E_{i,n}\left(\frac{1}{k_n}\sum_j y_j\right)\right\|^2 + \frac{(d'+2)k_n}{r_{i,k_n}^2}\left\|\frac{1}{k_n}\sum_j e_j\right\|^2\\ &+ 2\frac{(d'+2)k_n}{r_{i,k_n}^2}\left\langle\frac{1}{k_n}\sum_j y_j, E_{i,n}\left(\frac{1}{k_n}\sum_j y_j\right)\right\rangle + 2\frac{(d'+2)k_n}{r_{i,k_n}^2}\left\langle\frac{1}{k_n}\sum_j y_j, \frac{1}{k_n}\sum_j e_j\right\rangle\\ &+ 2\frac{(d'+2)k_n}{r_{i,k_n}^2}\left\langle\frac{1}{k_n}\sum_j e_j, E_{i,n}\left(\frac{1}{k_n}\sum_j y_j\right)\right\rangle.\end{aligned}$$

So, with probability greater than $1-n^{-6}$ for all i, we have $\delta_i = \delta_i^Y + \varepsilon_{i,1}$ with:

$$|\varepsilon_{i,1}| \le a^2\xi_n^2\delta_i^Y + a^2\xi_n^2(d'+2)k_n r_{i,k_n}^2 + 2a\xi_n\delta_i^Y + 2a\xi_n\sqrt{(d'+2)k_n\delta_i^Y}\,r_{i,k_n} + 2a^2\xi_n^2\sqrt{(d'+2)k_n\delta_i^Y}\,r_{i,k_n}.$$

By Lemma 1 we have $\mathbb{P}(r_n \ge a\rho_n) \le n^{-7}$, where $r_n = \max_i(r_{i,k_n})$. Because $\rho_n \le \xi_n$ we have, with probability greater than $1-2n^{-6}$, for all i

$$|\varepsilon_{i,1}| \le a\xi_n\delta_i^Y + a\xi_n^2\sqrt{\delta_i^Y} + a\xi_n^4. \tag{3}$$

First we will prove that $\delta_i \to \chi^2(d')$ in distribution. Consider the distribution of the random variable y_j for $j = 1,\dots,k_n$. By Proposition 4 it is the same as the following mixture law: with probability $1-p_n$: $z_i \equiv y_j/r_{i,k_n}$ is drawn according to an uniform law on $\mathcal{B}_{d'}(O, 1-cr_{i,k_n})$ and with probability p_n: $z_j \equiv y_j/r_{i,k_n}$ is drawn according to a residual law (supported by $\mathcal{B}_{d'}(O,1)$) with $p_n \le a\rho_n$. Denote by K_i the number of y_j belonging to the uniform part of the mixture (K_i has distribution Binom$((1-p_n), k_n)$), and introduce $\kappa_n = \max_i|(k_n - K_i)/\sqrt{k_n}|$. By application of Lemma 2 (with $k_n' = k_n$ and $q_n = a\rho_n$, because $k_n \ll n^{1/(d'+1)}$ we have $\rho_n\sqrt{k_n}\ln(n) \to 0$) we have, for n large enough:

$$\mathbb{P}(\ln(n)\kappa_n \ge a) \le n^{-6}. \tag{4}$$

For ease of writing let us suppose that $z_1,\dots,z_{K_i}$ are the observations belonging to the uniform part of the mixture. Consider $z_{K_i+1}^*,\dots,z_n^*$ i.i.d., uniformly distributed on $\mathcal{B}_{d'}(O,1)$. We will write $u_j \equiv z_j$ if $j \le K_i$, and $u_j \equiv z_j^*$ if $j > K_i$. If we define now $e_j \equiv z_j - z_j^*$ if $j > K_i$, then

$$\begin{aligned}\delta_i^Y|\{r_n \le a\xi_n\} &= (d'+2)k_n\left\|\frac{1}{k_n}\sum_{j=1}^{K_i}u_j + \frac{1}{k_n}\sum_{j=K_i+1}^{k_n}z_j^* + \frac{1}{k_n}\sum_{j=K_i+1}^{k_n}z_j - \frac{1}{k_n}\sum_{j=K_i+1}^{k_n}z_j\right\|^2\\ &= (d'+2)k_n\left\|\frac{1}{k_n}\sum_{j=1}^{k_n}u_j - \frac{1}{k_n}\sum_{j=K_i+1}^{k_n}e_j\right\|^2\\ &= (d'+2)k_n\left[\left\|\frac{1}{k_n}\sum_{j=1}^{k_n}u_j\right\|^2 + \left\|\frac{1}{k_n}\sum_{j=K_i+1}^{k_n}e_j\right\|^2 - 2\left\langle\frac{1}{k_n}\sum_{j=1}^{k_n}u_j, \frac{1}{k_n}\sum_{j=K_i+1}^{k_n}e_j\right\rangle\right].\end{aligned}$$

Consider $u_j/(1-cr_{i,k_n})$ for $i = 1,\dots,n$, which is an uniform sample on a d'-dimensional unit ball, and $\delta_i^U = (d'+2)k_n\|\sum_j u_j/(1-cr_{i,k_n})\|^2$. Then,

$$\delta_i^Y|\{r_n \le a\xi_n\} = (1-cr_{i,k_n})^2\delta_i^U + \varepsilon_{2,i} \quad \text{with } |\varepsilon_{2,i}| \le a\sqrt{\delta_i^U}\kappa_n + a\kappa_n^2. \tag{5}$$

By Proposition 1, $\delta_i^U \xrightarrow{\mathcal{L}} \chi^2(d')$ when $k_n \to +\infty$. This and (4) implies that $\varepsilon_{2,i} \xrightarrow{a.s} 0$. From $\mathbb{P}(\{r_n \le a\xi_n\}) \to 0$ we obtain $\delta_i^Y \xrightarrow{\mathcal{L}} \chi^2(d')$. That in turns, by (3) implies that $\varepsilon_{i,1} \xrightarrow{\mathcal{L}} 0$.

Lastly,

$$\delta_i \xrightarrow{\mathcal{L}} \chi^2(d'). \tag{6}$$

Regarding Theorem 1, we need an upper bound for $\mathbb{P}(\max_i \delta_i > t)$. If we use the classical rough bound $\mathbb{P}(\max_i \delta_i > t) \le n\mathbb{P}(\delta_i > t)$, we get $\mathbb{P}(\max_i \delta_i > t) \le n\Psi_{d'}(t) + no(1)$, which is useless because we have no control on the $no(1)$ term. To solve this problem we aim to get a better upper bound for $\mathbb{P}(\max_i \delta_i > t)$. This is done using Theorem 2.4 in [31], which states that for all $i = 1, \dots, n$

$$\mathbb{P}(\delta_i^U > t) \le \frac{2e^3}{9} F_{d'}(t). \tag{7}$$

Now the aim is to prove that, conditionally to $r_n \le a\xi_n$, $(\ln n)^{1/3} \max_i |\varepsilon_{i,2}| \xrightarrow{\text{a.s.}} 0$. First we have

$$\mathbb{P}\left(|\varepsilon_{i,2}| > \frac{\lambda}{(\ln n)^{1/3}}\right) \le \mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U}\kappa_n \ge \frac{\lambda}{(\ln n)^{1/3}}\right).$$

As

$$\mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U}\kappa_n \le \frac{\lambda}{(\ln n)^{1/3}}\right) \ge \mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U} \le \frac{\lambda(\ln n)^{2/3}}{a} \text{ and } \kappa_n \le \frac{a}{\ln n}\right)$$

we have

$$\mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U}\kappa_n \ge \frac{\lambda}{(\ln n)^{1/3}}\right) \le \mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U} \ge \frac{\lambda(\ln n)^{2/3}}{a} \text{ or } \kappa_n \ge \frac{a}{\ln n}\right)$$

and, finally, by (4) and (7)

$$\mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U}\kappa_n \ge \lambda\right) \le n\frac{2e^3}{9} F_{d'}\left(\frac{\lambda^2(\ln n)^{4/3}}{a^2}\right) + n^{-6}.$$

From

$$n\frac{2e^3}{9} F_{d'}\left(\frac{\lambda^2(\ln n)^{4/3}}{a^2}\right) \sim \frac{2e^3 n}{9} \frac{\exp(-\lambda^2 \ln n^{4/3}/(2a^2))(\lambda \ln n/2a)^{d'-2}}{\Gamma(d'/2)},$$

we obtain that

$$\sum \mathbb{P}\left(\max_{1\le i\le n} \sqrt{\delta_i^U}\kappa_n \ge \lambda\right) < +\infty$$

so, by Borel–Cantelli's Lemma, $(\ln n)^{1/3} \max_i |\varepsilon_{i,2}| \xrightarrow{\text{a.s.}} 0$.

Applying exactly same calculus it can be obtained from $(\ln n)^2\xi_n \to 0$ and (3) that, conditionally to $r_n \le a\xi_n$ $\max_i \delta_i \le \max_i \delta_i^U + \varepsilon_{3,n}$ with $(\ln n)^{1/3}\varepsilon_{3,n} \xrightarrow{\text{a.s.}} 0$. As a result,

$$\mathbb{P}\left(\max_{1\le i\le n} \delta_i \ge t \,\middle|\, \left\{\{r_n \le a\xi_n\} \cap \left\{|\varepsilon_{3,n}| \le a(\ln n)^{-1/3}\right\}\right\}\right) \le \frac{2e^3 n}{9} F_{d'}\left(t - a(\ln n)^{-1/3}\right).$$

Introduce $t_n = t_{n,\alpha} \equiv F_{d'}^{-1}(9\alpha/(2e^3 n))$. Notice that $t_n \to +\infty$ so that we can use the usual equivalent of $F_{d'}(t_n)$ and get

$$\frac{2e^3 n}{9} \frac{e^{-t_n/2}(t_n/2)^{d'/2-1}}{\Gamma(d'/2)} \to \alpha \quad \text{when } n \to +\infty.$$

Now note that $2e^3 n/9F_{d'}(x_n) \to \alpha \Leftrightarrow x_n = 2\ln n + (d'-2)\ln(\ln n) + 2\ln(2e^3/(9\alpha\Gamma(d'/2))) + o(1)$. Thus:

$$\mathbb{P}\left(\max_{1\le i\le n} \delta_i \ge t_n \middle| \left\{\{r_n \le a\xi_n\} \cap \left\{|\varepsilon_{3,n}| \le a(\ln n)^{-1/3}\right\}\right\}\right) \le \alpha + o(1).$$

Lastly, because e.a.s. $r_n \le a\xi_n$ (which follows from Lemma 1) and because $|\varepsilon_{3,n}|(\ln n)^{1/3} \xrightarrow{\text{a.s.}} 0$ we have $\mathbb{P}(\{r_n \le a\xi_n\} \cap \{|\varepsilon_{3,n}| \le a(\ln n)^{-1/3}\}) \to 1$, and so

$$\mathbb{P}\Big(\max_{1\le i\le n} \delta_i \ge t_n\Big) \le \alpha + o(1),$$

which proves Theorem 1. For $\lambda > 6$ we have

$$\mathbb{P}\Big(\max_{1\le i\le n} \delta_i \ge \lambda \ln n\Big) \le an^{1-\lambda/2}(\ln n)^{d'/2-1}$$

so that, once again, by the Borel–Cantelli's lemma, we obtain that if $\lambda > 6$,

$$\text{Under } H_0\text{:} \quad \Delta_{n,k_n} \ge \lambda \ln n \quad \text{e.a.s.} \tag{8}$$

5.2. *Proofs under* H_1 ($\partial M \neq \varnothing$)

The idea of the proof is the following. When $\partial M \neq \varnothing$, there exists an observation X_{i_0} close enough to the boundary (that is, such that $d(X_{i_0}, \partial M) \ll r_{i_0,k_n}$). Then $\mathcal{B}(X_{i_0}, r_{i_0,k_n}) \cap M$ looks like a "half ball", so that $\Delta_{n,k_n} \ge \delta_{i_0,k_n} \ge (d' + 2)k_n(\alpha_{d'} + o(1)) \to \infty$, $\alpha_{d'}$ being a positive constant (obtained from Proposition 2).

More precisely, set $\varepsilon_n \equiv a\ln(n)/n$. We will first prove that for a suitably chosen constant a, with probability one, for n large enough there exists an $X_{i_0} \in \partial M \oplus \varepsilon_n\mathcal{B} \equiv \{x : d(x, \partial M) \le \varepsilon_n\}$. Indeed, as ∂M is a compact $(d' - 1)$-manifold of class $\mathcal{C}^2$, by Proposition 14 in [33] it has positive reach. Then by Theorem 5.5 in [20], for n large enough $|\partial M \oplus \varepsilon_n\mathcal{B}| = C_{\partial M}\varepsilon_n(1 + o(1))$ where $C_{\partial M} > 0$ is a constant depending only on ∂M.

Thus,

$$\mathbb{P}\big((\partial M \oplus \varepsilon_n\mathcal{B}) \cap \mathcal{X}_n = \varnothing\big) \le (1 - f_0C_{\partial M}\varepsilon_n\big(1 - o(1)\big)^n \le n^{-f_0C_{\partial M}a+o(1)}.$$

If we choose $a > (f_0C_{\partial M})^{-1}$, then as a direct application of the Borel–Cantelli's lemma, with probability one, for n large enough, $\exists i_0, d(X_{i_0}, \partial M) \le \varepsilon_n$. Now we are going to prove that

$$\text{for all } X_{i_0} \in \partial M \oplus \mathcal{B}(0, \varepsilon_n), \quad \text{we have } r_{i_0,k_n} \ge \sqrt{\varepsilon_n} \quad \text{e.a.s.} \tag{9}$$

This will allow us to apply Proposition 3 part 5, which implies that $\mathcal{B}(X_{i_0}, r_{i_0,k_n})$ is "close" to a half ball.

First we assume n large enough to ensure that $\varepsilon_n < 1$. Cover ∂M with $\nu_n \le B\varepsilon_n^{(1-d')/2}$ balls, centred at $\{x_1, \dots x_{\nu_n}\} \subset \partial M$ with radius $\sqrt{\varepsilon_n}$. Observe that

$$\mathbb{P}(\exists X_{i_0} : r_{i_0,k_n} \le \sqrt{\varepsilon_n}) = \mathbb{P}\big(\exists X_{i_0} : \#\big\{\mathcal{B}(X_{i_0} : \sqrt{\varepsilon_n}) \cap \mathcal{X}_n\big\} \ge k_n\big).$$

Now, if $X_{i_0} \in \partial M \oplus \varepsilon_n\mathcal{B}$, then there exists a $y_i \in \partial M$ such that $\|X_{i_0} - y_i\| \le \varepsilon_n$ and y_i belongs to some ball $\mathcal{B}(x_r, \sqrt{\varepsilon_n})$ for $r = 1, \dots, \nu_n$. Then

$$\mathbb{P}(\exists X_{i_0} \in \partial M \oplus \varepsilon_n\mathcal{B} : r_{i_0,k_n} \le \sqrt{\varepsilon_n}) \le \sum_{i=1}^{\nu_n} \mathbb{P}\big(\#\big\{\mathcal{B}(x_i, 3\sqrt{\varepsilon_n}) \cap \mathcal{X}_n\big\} \ge k_n\big). \tag{10}$$

Applying Corollary 1 part 1 together with $f \le f_1$, we get that there exists a constant b such that

$$\mathbb{P}\big(\#\big\{\mathcal{B}(x_i, 3\sqrt{\varepsilon_n}) \cap \mathcal{X}_n\big\} \ge k_n\big) \le \sum_{j=k_n}^{n} \binom{n}{j} \big(b\varepsilon_n^{d'/2}\big)^j.$$

Now from the bounds $n!/(n - j)! \le n^j$ and $\sum_{j=k}^{n} x^j/j! \le x^k e^x/k!$, we obtain

$$\mathbb{P}\big(\#\big\{\mathcal{B}(x_i, 3\sqrt{\varepsilon_n}) \cap \mathcal{X}_n\big\} \ge k_n\big) \le \sum_{j=k_n}^{n} \frac{1}{j!}\big(bn\varepsilon_n^{d'/2}\big)^j \le \frac{(bn\varepsilon_n^{d'/2})^{k_n}}{k_n!} \exp\big(bn\varepsilon_n^{d'/2}\big). \tag{11}$$

Finally, (10), (11) and the upper bound on ν_n imply

$$\mathbb{P}(\exists X_{i_0} \in \partial M \oplus \varepsilon_n\mathcal{B}, r_{i_0,k_n} \le \sqrt{\varepsilon_n}) \le B\varepsilon_n^{(1-d')/2} \frac{(bn\varepsilon_n^{d'/2})^{k_n}}{k_n!} \exp\big(bn\varepsilon_n^{d'/2}\big).$$

If we apply Stirling's formula, for n large enough

$$\mathbb{P}(\exists X_{i_0} \in \partial M \oplus \varepsilon_n \mathcal{B}, r_{i_0,k_n} \leq \sqrt{\varepsilon_n}) \leq \exp\left\{-k_n \ln(k_n) + k_n + \frac{1-d'}{2}\ln(\varepsilon_n) + k_n \ln\bigl(bn\varepsilon_n^{d'/2}\bigr) + bn\varepsilon_n^{d'/2}\right\}.$$

From $k_n \gg \sqrt{n\ln(n)}$ when $d'=1$ and $k_n \gg \ln(n)$ for any dimension $d'>1$, it follows that

$$\mathbb{P}(\exists X_{i_0} \in \partial M \oplus \varepsilon_n \mathcal{B}, r_{i_0,k_n} \leq \sqrt{\varepsilon_n}) \leq \exp\bigl(-k_n \ln(k_n)\bigl(c_{d'} + o(1)\bigr)\bigr)$$

with $c_2 = 2$ and $c_{d'} = 1$ when $d' \neq 2$.

Then, $k_n \gg (\ln(n))$ ensures that

$$\sum_n \mathbb{P}(\exists X_{i_0} \in \partial M \oplus \varepsilon_n \mathcal{B}, r_{i_0,k_n} \leq \sqrt{\varepsilon_n}) < \infty.$$

The proof of (9) follows by a direct application of the Borel–Cantelli's lemma.

For an observation X_{i_0} such that $d(X_{i_0}, \partial M) \leq c_{\partial M}\ln(n)/n$, denote by x_0 its projection onto ∂M. Recall that u_{x_0} denotes the unit vector tangent to M and normal to ∂M pointing outward. Now introduce $Y = \varphi_{X_{i_0}}(X)|\{\|X - X_{i_0}\| \leq r_{i_0,k_n}\}$.

On the one hand, a direct consequence of Proposition 5 is that

$$\mathbb{E}\left(\left\langle \frac{Y - X_{i_0}}{r_{i_0,k_n}}, -u_{x_0}\right\rangle\right) \geq \alpha_{d'} - ar_{i_0,k_n} \geq \alpha_{d'} - ar_n.$$

On the other hand, by Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{k_n}\sum_{k=1}^{k_n}\left\langle \frac{Y_{k(i_0)} - X_{i_0}}{r_{i_0,k_n}}, -u_{x_0}\right\rangle - \mathbb{E}\left(\left\langle \frac{Y - X_{i_0}}{r_{i_0,k_n}}, -u_{x_0}\right\rangle\right) \leq -t\right) \leq \exp\bigl(-2t^2 k_n\bigr).$$

Thus

$$\mathbb{P}\left(\frac{1}{k_n}\sum_{k=1}^{k_n}\left\langle \frac{Y_{k(i_0)} - X_{i_0}}{r_{i_0,k_n}}, -u_{x_0}\right\rangle \leq \alpha_{d'} - ar_n - (\ln n)^{-1}\right) \leq 2\exp\bigl(-2k_n/(\ln n)^2\bigr).$$

Let us denote

$$Z = \frac{1}{k_n}\sum_{k=1}^{k_n}\frac{Y_{k(i_0)} - X_{i_0}}{r_{i_0,k_n}} \quad \text{and} \quad Z^* = \frac{1}{k_n}\sum_{k=1}^{k_n}\frac{X^*_{k(i_0)} - X_{i_0}}{r_{i_0,k_n}},$$

by Lemma 4 we have that there exists a sequence ϵ'_n such that, with probability greater than $1-n^{-6}$, $Z^* = Z + E_{i_0,n}Z + \epsilon'_n$ with $\|E_{i_0,n}\|_{\mathrm{op}} \leq a\xi_n$ and $\|\epsilon'_n\| \leq a\xi_n r_{i_0,n}$ with

$$\xi_n = \max\bigl((\ln n/n)^{1/(2d')}, (k_n/n)^{1/d'}, \sqrt{\ln n/k_n}\bigr)$$

as in previous section, and so with probability greater than $1 - n^{-6}$, $\langle Z^*, -u_{x_0}\rangle \geq (1 - a\xi_n)\langle Z, -u_{x_0}\rangle - a\xi_n r_{i_0,n}$ thus, we have that

$$\mathbb{P}\left(\frac{1}{\sqrt{(d'+2)k_n}}\sqrt{\delta_{i_0,k_n}} \leq (1 - a\xi_n)\bigl(\alpha_{d'} - ar_n - (\ln n)^{-1}\bigr) - a\xi_n r_{i_0,n}\right) \leq 2\exp\bigl(-2k_n/(\ln n)^2\bigr) + n^{-6}.$$

From $k_n \gg (\ln n)^4$, we get $\sum_n n(\exp(-2k_n/(\ln n)^2) + n^{-6}) < +\infty$, so that, by Borel–Cantelli's lemma for all i_0 such that $d(X_{i_0}, \partial M) \leq c_{\partial M}\ln(n)/n$, we have

$$\delta_{i_0,k_n} \geq \bigl(d'+2\bigr)k_n\bigl((1 - a\xi_n)\bigl(\alpha_{d'} - ar_n - (\ln n)^{-1}\bigr) - a\xi_n r_{i_0,n}\bigr)^2,$$

with probability one for n large enough. As by Lemma 1 $r_n \overset{\text{a.s.}}{\to} 0$, and because $\Delta_{n,k_n} \geq \delta_{i_0,k_n}$ we have for all $\lambda < 1$,

$$\mathbb{P}_{H_1}\bigl(\Delta_{n,k_n} \geq \bigl(d'+2\bigr)\alpha_{d'}^2\lambda k_n\bigr) = 1 \quad \text{for } n \text{ large enough.} \tag{12}$$

Now, observe that $k_n \gg (\ln(n))^4$ ensures the existence of an n_1 such that for all $n \geq n_1$, $k_n(d'+2)\alpha_{d'}^2/2 \geq t_{n,\alpha} \sim 2\ln n$, which together with (12) prove Theorem 2.

Similarly, for all $\lambda > 6$, $\mathbb{P}_{H_1}(\Delta_{n,k_n} \geq \lambda \ln n) = 1$ for n large enough and by (8) we also have $\mathbb{P}_{H_0}(\Delta_{n,k_n} \leq \lambda \ln n) = 1$ for n large enough, which concludes the proof of Theorem 3.

5.3. *Useful lemmas*

We will now give the details of the proofs of the lemmas and propositions used in the proofs of the main theorems. First we focus on the asymptotic behavior of the "centroid movement" when considering uniform samples on a ball or on a half ball.

Proposition 1. *Let $X_1, \dots, X_n$ be an i.i.d. sample uniformly drawn on $\mathcal{B}(x,r) \subset \mathbb{R}^d$ and write $\overline{X}_n \equiv \frac{1}{n}\sum_{i=1}^n X_i$. We have*

$$\frac{(d+2)n\|\overline{X}_n - x\|^2}{r^2} \xrightarrow{\mathcal{L}} \chi^2(d). \tag{13}$$

Proof. Taking $(X-x)/r$ we can assume that X obeys the uniform distribution on $\mathcal{B}(0,1)$.

If we write $X = (X_{\cdot,1}, \dots, X_{\cdot,d})$, then the density of $X_{\cdot,i}$ is

$$f(x) = \frac{1}{\sigma_d}\sigma_{d-1}\left(1-x^2\right)^{(d-1)/2}\mathbb{I}_{[-1,1]}(x),$$

and so

$$\operatorname{Var}(X_{\cdot,i}) = \int_{-1}^{1} x^2 \frac{1}{\sigma_d}\sigma_{d-1}\left(1-x^2\right)^{(d-1)/2}\,dx = \frac{\sigma_{d-1}}{\sigma_d}B\bigl(3/2, (d+1)/2\bigr),$$

where $B(x,y)$ is the Beta function. If we use the fact that $\sigma_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ and that $B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, we get

$$\frac{\sigma_{d-1}}{\sigma_d}B\bigl(3/2, (d+1)/2\bigr) = \frac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\Gamma(\frac{d+1}{2})} \times \frac{\Gamma(\frac{3}{2})\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d+4}{2})} = \frac{\Gamma(\frac{d+2}{2})\Gamma(\frac{3}{2})}{\sqrt{\pi}\Gamma(\frac{d+4}{2})}.$$

Since $\Gamma(z+1) = z\Gamma(z)$ and $\Gamma(1/2) = \sqrt{\pi}$, we obtain that

$$\frac{\sigma_{d-1}}{\sigma_d}B\bigl(3/2, (d+1)/2\bigr) = \frac{\sqrt{\pi}\frac{1}{2}}{\sqrt{\pi}\frac{d+2}{2}} = \frac{1}{d+2}.$$

Now, to prove (13), observe that

$$(d+2)n\|\overline{X}_n\|^2 = \left\|\sqrt{(d+2)n}\overline{X}_n\right\|^2 \xrightarrow{\mathcal{L}} N(0, I_d).$$

Then, $\|\sqrt{(d+2)n}\overline{X}_n\|^2 \xrightarrow{\mathcal{L}} \|N(0,I_d)\|^2$. Lastly, it is well known that $\|N(0,I_d)\|^2 \overset{\mathcal{L}}{=} \chi^2(d)$. □

Proposition 2. *Let X be uniformly drawn on $\mathcal{B}_u(x,r) = \mathcal{B}(x,r) \cap \{z \in \mathbb{R}^d : \langle z-x, u\rangle \geq 0\}$ where u is a unit vector. Then,*

$$\mathbb{E}\left(\frac{\langle X-x, u\rangle}{r}\right) = \alpha_d, \quad \textit{where } \alpha_d = \left(\frac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\Gamma(\frac{d+3}{2})}\right). \tag{14}$$

Proof. First assume that $r=1$, $x=0$ and $u = e_1 = (1,0,\dots,0)$. The marginal density of X_1 is

$$f_{X_1}(t) = \frac{2}{\sigma_d}\sigma_{d-1}\left(1-t^2\right)^{(d-1)/2}\mathbb{I}_{[0,1]}(x),$$

so

$$\mathbb{E}(X_1) = \int_0^1 2\frac{\sigma_{d-1}}{\sigma_d}x\left(1-x^2\right)^{d-1}\,dx = \frac{\sigma_{d-1}}{\sigma_d}\frac{\Gamma(1)\Gamma(\frac{d+1}{2})}{\Gamma(\frac{d+3}{2})} = \frac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\Gamma(\frac{d+3}{2})} = \alpha_d.$$

For a general value of r, x and u, define $Y = A_u(X - x)/r$ where A_u is a rotation matrix that sends u to $(1, 0, \dots, 0)$ (with $r > 0$). Then Y is uniformly distributed on $\mathcal{B}_{e_1}(0, 1)$ and so (14) holds. □

Now we aim to make explicit how close to an uniform sample on a ball or a half ball are the nearest neighbors statistics as $n \to +\infty$. First we detail some consequences of the regularity of M and ∂M. For $x \in M$ we denote by $N_x M$ the normal space at x. For $x \in \partial M$ we denote by u_x the unit normal outer vector to ∂M, that is, $\|u_x\| = 1$, $u_x \in T_x M \cap N_x \partial M$ and for all $\varepsilon > 0$ there exists an r_ε such that $\|y - x\| \leq r_\varepsilon \Rightarrow \langle \frac{y-x}{\|y-x\|}, u_x \rangle \leq \varepsilon$. Write $\varphi_x : M \to x + T_x M$ for the orthogonal projection onto the affine tangent space. Let $J_x(y)$ be the Jacobian matrix of φ_x^{-1} and $G_x(y) = \sqrt{\det(J_x'(y) J_x(y))}$.

Proposition 3. *Let $M \subset \mathbb{R}^d$ be a compact $\mathcal{C}^2$ d'-dimensional manifold with either $\partial M = \varnothing$ or ∂M is a $\mathcal{C}^2$ $(d' - 1)$-dimensional manifold. Then, there exists an $r_M > 0$ and $c_M > 0$ such that for all $r \leq r_M$:*

1. *For all $x \in M$, φ_x is a $\mathcal{C}^2$ bijection from $M \cap \mathcal{B}(x, r)$ to $\varphi_x(M \cap \mathcal{B}(x, r))$ for all $r \leq r_M$.*
2. *For all $x \in M$ and $y \in x + T_x M$ such that $\|x - y\| \leq r_M$, $|G_x(y) - 1| \leq c_M \|x - y\|$.*
3. *For all $x, y \in M$ such that $\|x - y\| \leq r_M$, $\|\varphi_x(y) - y\| \leq c_M \|x - \varphi_x(y)\|^2 \leq c_M \|x - y\|^2$.*
4. *For all $x \in M$, if $d(x, \partial M) \geq r$, then*
$$\mathcal{B}\big(x, r - c_M r^2\big) \cap (x + T_x M) \subset \varphi_x\big(\mathcal{B}(x, r) \cap M\big) \subset \mathcal{B}(x, r) \cap (x + T_x M).$$
5. *For all $x \in M$ with $d(x, \partial M) \leq r^2$, write x^* for its projection onto ∂M and define $H_x^- = \{y : \langle y - x, u_{x^*}\rangle \leq -c_M r^2\}$ and $H_x^+ = \{y : \langle y - x, u_{x^*}\rangle \leq c_M r^2\}$. Then,*
$$H_x^- \cap \mathcal{B}\big(x, r - c_M r^2\big) \cap (x + T_x M) \subset \varphi_x\big(\mathcal{B}(x, r) \cap M\big) \subset H_x^+ \cap \mathcal{B}(x, r) \cap (x + T_x M).$$

Proof. 1. When the manifold has no boundary, this result is classic (see, for instance Lemma 16 in [28]), but, as far as our knowledge extends, it has not been proved when M has a boundary.

It only has to be proved that there exists a radius $\rho_{M,0} > 0$ such that all the φ_x restricted to $M \cap \mathcal{B}(x, \rho_{M,0})$ are one to one. Proceeding by contradiction, let $r_n \to 0$, x_n, y_n and z_n be such that $\{y_n, z_n\} \subset \mathcal{B}(x_n, r_n)$ and $\varphi_{x_n}(y_n) = \varphi_{x_n}(z_n)$. Since M is compact, we can assume that (by taking a subsequence if necessary) $x_n \to x \in M$. Put $w_n \equiv (y_n - z_n)/\|y_n - z_n\| \to w$. Since $\varphi_{x_n}(y_n) = \varphi_{x_n}(z_n)$, we have $w_n \in (T_{x_n} M)^\perp$. Since M is of class $\mathcal{C}^2$, we have $w \in (T_x M)^\perp$. Let γ_n be a geodesic curve on M that joins y_n to z_n (there exists at least one since M is compact and path connected). As M is compact and $\mathcal{C}^2$, it has an injectivity radius $r_{\text{inj}} > 0$. Therefore (see Proposition 88 in [6]), if we take n so large that $r_n \leq r_{\text{inj}}/2$, we may take γ_n to be the (unique) geodesic which is the image, by the exponential map, of a vector $v_n \in T_{y_n} M$. The Taylor expansion of the exponential map shows that $w_n = v_n/\|y_n - z_n\| + o(1)$. Then, taking the limit as $n \to \infty$, we get $w \in T_x M$, which contradicts the fact that $w \in (T_x M)^\perp$.

As a conclusion, there exists an r_0 such that for all $x \in M$, φ_x is one to one from $M \cap \mathcal{B}(x, r)$ to $\varphi_x(M \cap \mathcal{B}(x, r))$ (then the existence of an r_1 such that for all $x \in M$ and $r \leq r_1$, φ_x is one to one and $\mathcal{C}^2$ is easily obtained).

2 and 3. For all $x \in M$ there exist k functions $\Phi_{x,k} : \varphi_x(M \cap \mathcal{B}(x, r_1)) - x \to \mathbb{R}$ such that

$$\varphi_x^{-1} : \varphi_x\big(M \cap \mathcal{B}(x, r_1)\big) \to M \cap \mathcal{B}(x, r_1)$$
$$x + \begin{pmatrix} y_1 \\ \vdots \\ y_{d'} \\ 0_{d-d'} \end{pmatrix} \mapsto x + \begin{pmatrix} y \\ \Phi_{x,d'+1}(y) \\ \vdots \\ \Phi_{x,d}(y) \end{pmatrix}.$$

The compactness of M together with its $\mathcal{C}^2$ regularity allows us to find a (uniform) radius r_2 such that all the $\Phi_{x,k}$ are $\mathcal{C}^2$ on $\varphi_x(M \cap \mathcal{B}(x, r_2))$. Note that as φ_x is the orthogonal projection, we have, for all x and k, that $\nabla_0 \Phi_{x,k} = 0$. Once again the smoothness and compactness assumptions guarantee that the eigenvalues of the Hessian matrices $H(\Phi_{x,k})(0)$ are uniformly bounded from above by some $\lambda_M > 0$.

Thus, first

$$\big\|\varphi_x^{-1}(y) - y\big\|^2 = \sum_{k=1}^{d-d'} \big(\Phi_{x,d'+k}(y - x)\big)^2 \leq \big(d - d'\big)\lambda_M \|x - y\|^4 + o\big(\|x - y\|^4\big), \tag{15}$$

and then there exist a c_3 and r_3 such that for all $(x, y) \in M \times \varphi_x(M \cap B(x, r_2))$ such that $\|x - y\| \leq r_3$,

$$\big\|\varphi_x^{-1}(y) - y\big\| \leq c_3 \|x - y\|^2. \tag{16}$$

Second:

$$J_x(y) = \begin{pmatrix} I_{d'} \\ \nabla_y \Phi_{x,d'+1} \\ \vdots \\ \nabla_y \Phi_{x,d} \end{pmatrix} = \begin{pmatrix} I_{d'} \\ O(\|y\|) \\ \vdots \\ O(\|y\|) \end{pmatrix} \quad \text{and} \quad J_x(y)' J_x(y) = I_{d'} + O\big(\|y\|\big).$$

This, together with the differentiability of the determinant, implies that there exist a $c_4 > 0$ and $r_4 > 0$ such that for all $x, y \in M$ fulfilling $\|x - y\| \leq r_4$,

$$\big|G_x(y) - 1\big| \leq c_4 \|x - y\|.$$

4. Only the first inclusion has to be proved: the second one is obvious. Introduce $\tilde{r} = \min\{r_1, r_2, r_3, 1/c_3\}$. Proceeding by contradiction, suppose that there are r, x and y such that $0 < r \leq \tilde{r}$, $x \in M$, $d(x, \partial M) > r$, $y \in \mathcal{B}(x, r(1 - c_3 r)) \cap T_x M$ and $y \notin \varphi_x(\mathcal{B}(x, r) \cap M)$. As $x \in \varphi_x(\mathcal{B}(x, r) \cap M)$, the segment $[x, y]$ intersects $\partial(\varphi_x(\mathcal{B}(x, r) \cap M))$. Let $z \in [x, y] \cap \partial\varphi_x(\mathcal{B}(x, r) \cap M)$. On the one hand, we have $\|x - z\| < \|x - y\| \leq r(1 - c_3 r)$. On the other hand, since φ_x^{-1} is a continuous function, $\partial\varphi_x(\mathcal{B}(x, r) \cap M) = \varphi_x(\partial(\mathcal{B}(x, r) \cap M))$, and, because $d(x, \partial M) > r$, one has that $\partial\varphi_x(\mathcal{B}(x, r) \cap M) = \varphi_x(M \cap \partial\mathcal{B}(x, r)))$. Then, there exist a z_0, $\|x - z_0\| = r$, and $\varphi_x(z_0) = z$. Now by (16),

$$r^2 = \|x - z\|^2 + \|z - z_0\|^2 < r^2(1 - c_3 r)^2 + c_3^2 r^4 = r^2 - 2c_3 r^3(1 - c_3 r) \leq r^2,$$

which is a contradiction. Then there exist a c_5 and r_5 such that for all $r \leq r_5$ and for all $x \in M$ with $d(x, \partial M) > r$,

$$\mathcal{B}\big(x, r - c_5 r^2\big) \cap (x + T_x M) \subset \varphi_x\big(\mathcal{B}(x, r) \cap M\big) \subset \mathcal{B}(x, r) \cap (x + T_x M). \tag{17}$$

5. Sketch of proof. Suppose that $\partial M \neq \varnothing$. For each $x^* \in \partial M$ write $\varphi_{x^*}^*$ for the affine projection on $x^* + T_{x^*}\partial M$. First note that for all y we have $\varphi_{x^*}^*(y) = \varphi_{x^*}(y) - \langle y - x^*, u_{x^*}\rangle u_{x^*}$. Thus, by the triangle inequality, $|\langle y - x^*, u_{x^*}\rangle| \leq \|\varphi_{x^*}^*(y) - y\| + \|\varphi_{x^*}(y) - y\|$.

Recall that ∂M is of class $\mathcal{C}^2$ and take $y \in \partial M$. Then by applying (17) (to M and ∂M) we have that there are r_6 and c_6 such that for all $x^* \in \partial M$ and for all $y \in \partial M$ with $\|x^* - y\| \leq r_6$, $|\langle y - x^*, u_{x^*}\rangle| \leq c_6 \|x^* - y\|^2$. Thus, for all $r \leq r_6/2$ and for all x with $d(x, \partial M) \leq r_6/2$, and denoting by x^* the projection of x onto ∂M, we have

$$\partial M \cap \mathcal{B}(x, r) \subset \mathcal{B}(x, r) \cap \big\{y : \big|\langle y - x^*, u_{x^*}\rangle\big| \leq c_6 \big\|x^* - y\big\|^2\big\}.$$

Taking now an x with $d(x, \partial M) \leq r^2$ gives

$$\begin{aligned} \varphi_x\big(\partial M \cap \mathcal{B}(x, r)\big) &\subset \varphi_x\big(\mathcal{B}(x, r) \cap \big\{y : \big|\langle y - x, u_{x^*}\rangle\big| \leq c_7 r^2\big\}\big) \\ &\subset \varphi_x\big(\mathcal{B}(x, r)\big) \cap \varphi_x\big(\big\{y : \big|\langle y - x, u_{x^*}\rangle\big| \leq c_7 r^2\big\}\big). \end{aligned}$$

Clearly $\varphi_x(\partial M \cap \mathcal{B}(x, r)) \subset \mathcal{B}(x, r) \cap (x + T_x M)$.

Recall that, as ∂M is a compact $\mathcal{C}^2$ manifold it has a positive reach (see Proposition 14 in [33]). Let us denote by c the reach of ∂M, so for all $(x^*, y) \in (\partial M)^2$ we have from Theorem 4.8 part 7 in [20].

$$\big|\langle y - x^*, u_{x^*}\rangle\big| < \frac{\|y - x^*\|^2}{2c}. \tag{18}$$

Notice now that for all $y \in \partial M \cap \mathcal{B}(x, r)$ we have $y \in \partial M \cap \mathcal{B}(x^*, r + r^2)$, and

$$\big|\langle \varphi_x(y) - x, u_{x^*}\rangle\big| \leq \big|\langle \varphi_x(y) - y, u_{x^*}\rangle\big| + \big|\langle y - x^*, u_{x^*}\rangle\big| + \big|\langle x^* - x, u_{x^*}\rangle\big|$$

thus

$$\big|\langle \varphi_x(y) - x, u_{x^*}\rangle\big| \leq \big\|\varphi_x(y) - y\big\| + \big|\langle y - x^*, u_{x^*}\rangle\big| + \big|\langle x^* - x, u_{x^*}\rangle\big|.$$

Equations (16) and (18) entails,

$$\big|\langle \varphi_x(y) - x, u_{x^*}\rangle\big| \leq c_3 \|x - y\|^2 + \frac{\|y - x^*\|^2}{2c} + \big\|x^* - x\big\|.$$

Recall that $\|x - y\| \leq r$ and $\|x - x^*\| \leq r^2$, then $|\langle \varphi_x(y) - x, u_{x^*}\rangle| \leq r^2(c_3 + (1 + r)^2/(2c) + 1)$.

Lastly, we proved that there exists c_7 such that,

$$\varphi_x\big(\partial M \cap \mathcal{B}(x,r)\big) \subset \mathcal{B}(x,r) \cap (x + T_x M) \cap \big\{y : \big|\langle y - x, u_{x^*}\rangle\big| \leq c_7 r^2\big\}.$$

Now, when $r \leq r_1$, we have $\partial\varphi_x(M \cap \mathcal{B}(x,r)) = \varphi_x(\partial(M \cap \mathcal{B}(x,r))) = \varphi_x(\partial M \cap \mathcal{B}(x,r)) \cup \varphi_x(M \cap \partial\mathcal{B}(x,r))$ As in the proof of previous part, we easily obtain

$$\partial\varphi_x\big(M \cap \mathcal{B}(x,r)\big) \subset (x + T_x M) \cap \big\{y : \big|\langle y - x, u_{x^*}\rangle\big| \leq c_7 r^2\big\} \cup \big(\mathcal{B}(x,r) \setminus \big(\mathcal{B}\big(x, r - c_3 r^2\big)\big).$$

Thus, arguing on the basis of connectedness arguments, we have:

$$\begin{aligned}(x + T_x M) \cap \big\{y : \langle y - x, u_{x^*}\rangle \leq -c_7 r^2\big\} \cap \mathcal{B}\big(x, r - c_3 r^2\big) \subset \varphi_x\big(M \cap \mathcal{B}(x,r)\big) \\ \subset (x + T_x M) \cap \big\{y : \langle y - x, u_{x^*}\rangle \leq -c_7 r^2\big\} \cap \mathcal{B}(x,r)\end{aligned} \tag{19}$$

or

$$\begin{aligned}&(x + T_x M) \cap \big\{y : \langle y - x, u_{x^*}\rangle \geq c_7 r^2\big\} \cap \mathcal{B}\big(x, r - c_3 r^2\big) \\ &\quad \subset \varphi_x\big(M \cap \mathcal{B}(x,r)\big) \subset (x + T_x M) \cap \big\{y : \langle y - x, u_{x^*}\rangle \geq c_7 r^2\big\} \cap \mathcal{B}(x,r).\end{aligned} \tag{20}$$

Because u_x is the normal outer vector to ∂M we have (19) and not (20). The choice of (19) comes from the orientation of u_{x^*}. □

Recall the change of variables formula

$$V \subset \mathcal{B}(x, r_{0,M}) \quad \Rightarrow \quad \mathbb{P}_X(V) = \int_{V \cap M} f \, d\omega = \int_{\varphi_x(V)} f\big(\varphi_x^{-1}(y)\big) G_x(y) \, dy. \tag{21}$$

Corollary 1. *Let $X_1, \dots, X_n$ be an i.i.d. sample of X, a random variable whose distribution $\mathbb{P}_X$ fulfills condition P. Then, there exist positive constants r_M, A, B and C such that if $r \leq r_M$, then*:

1. *For all $x \in M$, $Ar^{d'} \leq \mathbb{P}_X(\mathcal{B}(x,r)) \leq Br^{d'}$.*
2. *For all $x \in M$ such that $d(x, \partial M) \geq r$, $|\mathbb{P}_X(\mathcal{B}(x,r)) - f(x)\sigma_{d'} r^{d'}| \leq Cr^{d'+1}$.*

Proof. For any $r \leq r_M$ and any $x \in M$,

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \leq f_1 \int_{\varphi_x(\mathcal{B}(x,r) \cap M)} G_x(y) \, dy.$$

Thus by Proposition 3, part 2 we have

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \leq f_1 \sigma_{d'} r^{d'} (1 + c_M r). \tag{22}$$

For any $r > 0$ let us consider first $x \in M$ such that $d(x, \partial M) \geq r/2$. Then

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \geq \mathbb{P}_X\big(\mathcal{B}(x, r/2)\big) \geq f_0 \int_{\varphi_x(\mathcal{B}(x,r/2) \cap M)} G_x(y) \, dy.$$

Since $r \leq 2r_M$, applying Proposition 3 parts 2 and 4 we obtain

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \geq f_0 \sigma_{d'} \big(r - c_M r^2\big)^{d'} (1 - c_M r). \tag{23}$$

Let $x \in M$ such that $d(x, \partial M) \leq r/2$, let x^* be the projection of x onto ∂M, then we have

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \geq \mathbb{P}_X\big(\mathcal{B}\big(x^*, r/2\big)\big) \geq f_0 \int_{\varphi_{x^*}(\mathcal{B}(x^*, r/2) \cap M)} G_{x^*}(y) \, dy.$$

Since $r \leq 2r_M$, applying Proposition parts 2 and 5, we obtain

$$\mathbb{P}_X\big(\mathcal{B}(x,r)\big) \geq f_0 \bigg(\frac{\sigma_{d'}}{2} (r)^{d'} - c_M \sigma_{d'-1} r^{d'+1}\bigg)(1 - c_M r). \tag{24}$$

Lastly, part 1 is a direct consequence of (22), (23) and (24).

To prove part 2, assume $r \leq r_M$. From the Lipschitz hypothesis on f, we get

$$\left|\mathbb{P}_X\big(\mathcal{B}(x,r)\big) - f(x)\int_{\mathcal{B}(x,r)\cap M} d\omega\right| \leq rK_f \int_{\mathcal{B}(x,r)\cap M} d\omega.$$

By (21), $\int_{\mathcal{B}(x,r)\cap M} d\omega = \int_{\varphi_x(\mathcal{B}(x,r)\cap M)} G_x(y)\,dy$. Applying Proposition 3 part 2 there follows

$$\left|\int_{\mathcal{B}(x,r)\cap M} d\omega - \int_{\varphi_x(\mathcal{B}(x,r)\cap M)} dy\right| \leq c_{M,1} r \int_{\varphi_x(\mathcal{B}(x,r)\cap M)} dy.$$

By Proposition 3 part 4,

$$\left|\int_{\mathcal{B}(x,r)\cap M} d\omega - \int_{\mathcal{B}(x,r)\cap T_x M} 1\,dy\right| \leq \int_{(\mathcal{B}(x,r)\setminus\mathcal{B}(x,r-c_{M,2}r^2))\cap T_x M} dy + c_{M,1} r \int_{\mathcal{B}(x,r)\cap T_x M} dy.$$

This implies

$$\begin{aligned}\left|\mathbb{P}_X\big(\mathcal{B}(x,r)\big) - f(x)\sigma_{d'} r^{d'}\right| &\leq rK_f\big(\sigma_{d'} r^{d'}\big(1-(1-c_{M,2}r)^{d'}\big)\big)\\ &\quad + f(x)\big(\sigma_{d'} r^{d'}\big(1-(1-c_{M,2}r)^{d'}\big) + c_{M,1}\sigma_{d'} r^{d'+1}\big).\end{aligned}$$

Thus, the choice of any constant $C_1 > \sigma_{d'}(K_f + f_1 d c_{M,2} + c_{M,1})$ allows us to find a suitable R_1. □

This in turns implies the following lemma.

Lemma 1. *Let $X_1, \ldots, X_n$ be an i.i.d. sample of X, a random variable whose distribution $\mathbb{P}_X$ fulfills condition P. Introduce $\rho_n = (2A^{-1}((\ln(n)/n)^{1/2} + k_n/n))^{1/d'}$ where A is the constant introduced in Corollary 1. Then $\mathbb{P}(r_n \geq \rho_n) \leq n^{-7}$, where r_n was introduced in Definition 1.*

Proof. Let us introduce the random variables $Z_i \equiv \#\{\{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\} \cap \mathcal{B}(X_i, \rho_n)\}$. Z_i follows a binomial distribution. We can bound $\mathbb{P}(r_n \geq \rho_n) \leq \sum_i \mathbb{P}(Z_i \leq k_n)$. Put $p_i = \mathbb{P}_X(\mathcal{B}(X_i, \rho_n))$. By Corollary 1 part 1, we have $k_n/n \leq p_i$. Then, by Hoeffding's inequality, $\mathbb{P}(r_{i,k_n} \geq \rho_n) = \mathbb{P}(Z_i - np_i < k_n - np_i) \leq \exp(-2n(k_n/n - p_i)^2)$, from which it follows that $\mathbb{P}(r_n \geq \rho_n) \leq \sum_i \exp(-2n(k_n/n - p_i)^2)$. Using again Corollary 1 and the definition of ρ_n, we obtain

$$\mathbb{P}(r_n \geq \rho_n) \leq n\exp\big(-2n\big(k_n/n + \big(\ln(n)/n\big)^{1/2}\big)^2\big) \leq n^{-7},$$

which concludes the proof. □

Now that we have guaranteed that $r_n \to 0$, the following proposition will make explicit how close the projection of the sample onto the tangent space of k_n-nearest neighbors is to an uniform random sample on a d'-dimensional sphere when the manifold M has no boundary.

Proposition 4. *Let X be a random variable whose distribution $\mathbb{P}_X$ fulfills condition P with $\partial M = \varnothing$. For each $x_0 \in M$, put $Y_1 = \varphi_{x_0}(X)$ the projection onto the tangent space and $Y = Y_1|\{\|X - x_0\| \leq r\}$. Then there exists a constant $a > 0$ such that if r is small enough, $Y \overset{\mathcal{L}}{=} Z$, where Z has a mixture law with density $g_{x_0} = (1-p)g_u + pg_v$ where g_u is the density of a random variable uniformly distributed on $\mathcal{B}_{d'}(O, r - cr^2)$, g_v is a density supported by $\mathcal{B}_{d'}(O, r)$, and $p \leq ar$.*

Proof. Observe that $X|\{\|X - x_0\| \leq r\}$ has density $f_{x_0}(x) = \frac{f(x)}{\mathbb{P}_X(\mathcal{B}(x_0,r))}\mathbb{I}_{M\cap\mathcal{B}(x_0,r)}$. By Corollary 1 part 2, for r small enough,

$$\frac{f(x)}{f(x)\sigma_{d'} r^{d'}(1 + \frac{Cr}{f_0\sigma_{d'}})}\mathbb{I}_{M\cap\mathcal{B}(x_0,r)} \leq f_{x_0}(x) \leq \frac{f(x)}{f(x)\sigma_{d'} r^{d'}(1 - \frac{Cr}{f_0\sigma_{d'}})}\mathbb{I}_{M\cap\mathcal{B}(x_0,r)}.$$

The random variable Y has density $g_{x_0}(x) = f_{x_0}(\varphi_{x_0}^{-1}(x))G_{x_0}(x)\mathbb{I}_{B_{x_0}}$, where $B_{x_0} = \varphi_{x_0}(M \cap \mathcal{B}(x_0, r))$. By Proposition 3, $|G_{x_0}(x) - 1| \leq c_M r$, and so

$$\frac{1 - c_M r}{\sigma_{d'} r^{d'}(1 + \frac{Cr}{f_0 \sigma_{d'}})}\mathbb{I}_{B_{x_0}} \leq g_{x_0}(x) \leq \frac{1 + c_M r}{\sigma_{d'} r^{d'}(1 - \frac{Cr}{f_0 \sigma_{d'}})}\mathbb{I}_{B_{x_0}}.$$

Note that by Proposition 3 we have

$$\mathcal{B}\big(x_0, r(1 - c_M r)\big) \cap (x_0 + T_{x_0}M) \subset B_{x_0} \subset \mathcal{B}(x_0, r) \cap (x_0 + T_{x_0}M).$$

Put $B^-(x_0, r) \equiv \mathcal{B}(x_0, r(1 - c_M r)) \cap (x_0 + T_{x_0}M)$, and define

$$p \equiv (1 - c_M r)^{d'+1}\left(\frac{C}{f_0 \sigma_{d'}} r + 1\right)^{-1}.$$

Observe that g_{x_0} is a density and has the property that $g_{x_0}(x) \geq p g_u(x)$, $g_{x_0}(x) = 0$ if $\|x - x_0\| > r$, and $p = O(r)$. This concludes the proof. □

Proposition 5. *Let X be a random variable whose distribution $\mathbb{P}_X$ fulfills condition P with $\partial M \neq \varnothing$. For each $x_0 \in M$ with $d(x_0, \partial M) \leq r^2$, put $Y_1 = \varphi_{x_0}(X)$ the projection onto the tangent space and $Y = Y_1 | \{\|X - x_0\| \leq r\}$. Then there exists a constant $a > 0$ such that if r is small enough, $Y \overset{\mathcal{L}}{=} Z$, where Z has a mixture law with density $g_{x_0} = (1 - p)g_u + p g_v$ where g_u is the density of a random variable uniformly distributed on $\mathcal{B}_{d'}(O, r - cr^2) \cap \{x, \langle x, -u_{x_0^*}\rangle \geq cr^2\}$, g_v is a density supported by $\mathcal{B}_{d'}(O, r)$ and $p \leq ar$.*

The proof is similar to the previous one and is left to the reader.

In the proofs of Theorems 1 and 2 we also needed to control the number of points in the mixture that are drawn with the non-uniform random variable. This is done with the following lemma.

Lemma 2. *Suppose $T_n \rightsquigarrow \mathrm{Binom}(k'_n, q_n)$ with $q_n \sqrt{k'_n} \ln(n) \to 0$ and $k'_n / (\ln(n))^4 \to +\infty$.*

Then, for all $\lambda > 0$, for all $b > 0$, and for n large enough, $n\mathbb{P}(\ln(n) T_n / \sqrt{k'_n} > \lambda) < n^{-b}$.

Proof. By Bernstein Inequality we have

$$\mathbb{P}\left(\frac{T_n}{k'_n} \geq q_n + \sqrt{\frac{2 q_n u}{k'_n}} + \frac{u}{k'_n}\right) \leq e^{-u}$$

then

$$\mathbb{P}\left(\frac{T_n \ln n}{\sqrt{k'_n}} \geq \sqrt{k'_n} q_n \ln(n) + \sqrt{2 q_n u (\ln n)^2} + \frac{u \ln n}{\sqrt{k'_n}}\right) \leq e^{-u}.$$

Thus, taking $u = \lambda \sqrt{k'_n}/(2 \ln n)$ and considering n large enough to ensure

$$\sqrt{k'_n} q_n \ln(n) + \sqrt{2 q_n \lambda \sqrt{k'_n} (\ln n)} \leq \lambda/2,$$

which is possible according to the condition $\sqrt{k'_n} q_n \ln(n) \to 0$, we have:

$$\mathbb{P}\left(\frac{T_n \ln n}{\sqrt{k'_n}} \geq \lambda\right) \leq \exp\left(-\lambda \frac{\sqrt{k'_n}}{2 \ln n}\right) \leq \exp\left(-(\ln n)\left(\lambda \sqrt{\frac{k'_n}{(\ln n)^4}}\right)\right).$$

Lastly, the results follows from $k'_n / (\ln(n))^4 \to +\infty$, taking n large enough to ensure $\lambda \sqrt{k'_n}/(\ln n)^2 \geq b + 1$. □

We have proved that the projection of the k_n nearest neighbors onto the tangent space is close to an uniform draw. The following proposition quantifies how this (unknown) projection is close to the estimation via a local PCA.

Proposition 6. *Let $X_1, \dots, X_n$ be an i.i.d. sample in $\mathbb{R}^d$ of a law whose support is included in the unit ball. Let $\hat{S}_n = \frac{1}{n}\sum_i X'_i X_i$ and $S = \mathbb{E}(X'X)$. Then*

(i) $\mathbb{P}(\|\hat{S}_n - S\|_\infty > s) \le 2d^2 \exp(-s^2 n/2)$;
(ii) *If, moreover, X_i is uniformly drawn in the unit ball, then*

$$\mathbb{P}\left(\left\|\hat{S}_n - \frac{1}{d+2} I_d\right\|_\infty > s\right) \le 2d^2 \exp\bigl(-s^2 n/2\bigr)$$

and there exist a and s_0 such that for all $s < s_0$, $\mathbb{P}(\|\hat{S}_n^{-1} - (d+2)I_d\|_\infty > as) \le 2d^2 \exp(-s^2 n/2)$ for n large enough.

Proof. Part (i) is a direct consequence of the application of Hoeffding's inequality: for all i, j we have $\mathbb{P}(|\hat{S}_n - S|_{i,j} > s) \le 2\exp(-s^2 n/2)$. Part (ii) is a consequence of part (i) (for uniformly drawn $S = (d+2)^{-1} I_d$) and of the differentiability of matrix inversion (close to the identity matrix). □

The following result provides the uniform convergence rate of the local PCA to the tangent spaces. Write $\mathcal{M}_d(\mathbb{R})$ for the space of $d \times d$ matrices with coefficients in $\mathbb{R}$. Let $I_{d',d} \in \mathcal{M}_d(\mathbb{R})$ be the block matrix

$$I_{d',d} = \begin{pmatrix} I_{d'} & 0 \\ 0 & 0 \end{pmatrix}.$$

For a symmetric matrix $S \in \mathcal{M}_d(\mathbb{R})$, put $S = Q_S \Delta_S Q'_S$, with Δ_S diagonal with $(\Delta_S)_{1,1} \ge (\Delta_S)_{2,2} \ge \cdots \ge (\Delta_S)_{d,d}$ and Q_S the matrix containing (by columns) an orthonormal basis of eigenvectors. Write $P_{S,d'} = Q_S I_{d',d} Q'_S$, that is, the matrix of the orthogonal projection on the plane spanned by the d' eigenvectors associated to the d' largest eigenvalues of S. Note that $P_{I_{d',d},d'} = I_{d',d}$.

Lemma 3. *Let $X_1, \dots, X_n$ be an i.i.d. sample drawn according to a distribution $\mathbb{P}_X$ which fulfills condition P, with $\partial M = \varnothing$. Denote by $\tilde{\varphi}_{X_i}$ the linear projection onto the tangent space at X_i and by $\hat{\varphi}_{X_i}$ the linear projection onto the estimation of the tangent space via local PCA. With probability greater than $1 - n^{-6}$ for n large enough, there exist a constant a and a matrices $E_{i,n}$ with $\|E_{i,n}\|_{\mathrm{op}} \le a(\sqrt{\ln(n)/k_n} + \rho_n)$ such that, for all i and all $y \in \mathcal{B}(X_i, \rho_n)$ we have:*

$$\bigl\|\hat{\varphi}_{X_i}(y) - (I_d - E_{i,n})\tilde{\varphi}_{X_i}(y)\bigr\| \le a\bigl(\sqrt{\ln(n)/k_n} + \rho_n\bigr)\bigl\|\tilde{\varphi}_{X_i}(y)\bigr\|^2.$$

Proof. By Proposition 6, for all i, $\mathbb{P}(\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - r_{i,k_n}^{-2} \Sigma_i\|_\infty \ge t) \le 2d^2 \exp(-t^2 k_n/2)$, where $\Sigma_i = \mathbb{E}(Y'Y \| \|Y\| \le r_{i,k_n})$ with $Y = X - X_i$ and $\hat{S}_{i,k_n}$ is as in Definition 1. Then

$$\mathbb{P}\bigl(\exists i : \bigl\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - r_{i,k_n}^{-2} \Sigma_i\bigr\|_\infty \ge t\bigr) \le n2d^2 \exp\bigl(-t^2 k_n/2\bigr).$$

Now if we apply the Borel–Cantelli lemma with $t = 4\sqrt{\ln(n)/k_n}$, we get that, with probability one, for n large enough,

$$\mathbb{P}\bigl(\exists i, \bigl\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - r_{i,k_n}^{-2} \Sigma_i\bigr\|_\infty \ge 4\sqrt{\ln(n)/k_n}\bigr) \le 2d^2 n^{-7}. \tag{25}$$

Denote by P_i the matrix whose first d' columns form an orthonormal basis of $T_{X_i}M$, completed to obtain an orthonormal base of $\mathbb{R}^d$. By Lemma 1, since $k_n/n \to 0$, we have $\rho_n \to 0$ and, for n large enough, combining Proposition 3 parts 3 and 4 and (21), there exists a c such that for n large enough

$$\mathbb{P}\bigl(\text{for all } i : \bigl\|r_{i,k_n}^{-2} \Sigma_i - \bigl(d'+2\bigr)^{-1} P'_i I_{d',d} P_i\bigr\|_\infty \le c\rho_n | \{r_n \le \rho_n\}\bigr) = 1. \tag{26}$$

Now, (25), (26) and Lemma 1 give that, for n large enough,

$$\mathbb{P}\bigl(\exists i, \bigl\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - \bigl(d'+2\bigr)^{-1} P'_i I_{d',d} P_i\bigr\|_\infty \ge 4\sqrt{\ln(n)/k_n} + c\rho_n\bigr) \le \bigl(2d^2+1\bigr)n^{-7}.$$

Thus, by usual inequality on the norms,

$$\mathbb{P}\bigl(\exists i, \bigl\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - \bigl(d'+2\bigr)^{-1} P'_i I_{d',d} P_i\bigr\|_{\mathrm{op}} \ge 4d^{-1}\sqrt{\ln(n)/k_n} + cd^{-1}\rho_n\bigr) \le \bigl(2d^2+1\bigr)n^{-7}.$$

Suppose now that, for all i we have

$$\bigl\|r_{i,k_n}^{-2} \hat{S}_{i,k_n} - \bigl(d'+2\bigr)^{-1} P'_i I_{d',d} P_i\bigr\|_{\mathrm{op}} \le 4d^{-1}\sqrt{\ln(n)/k_n} + cd^{-1}\rho_n.$$

By previous equation and Lemma 19 in [5] (based on [18])) we have that, for all i

$$\|\tilde{\varphi}_{X_i} - \hat{\varphi}_{X_i}\|_{\mathrm{op}} \leq \frac{\sqrt{2}(d'+2)}{d}\big(4\sqrt{\ln(n)/k_n} + c\rho_n\big). \tag{27}$$

Now suppose that $r_n \leq \rho_n$, which according to Lemma 1 it happens with probability greater than $1 - n^{-7}$. Consider $y \in M \cap \mathcal{B}(X_i, \rho_n) - X_i$. Introduce $E_{i,n}$ the matrix of the application $\tilde{\varphi}_{X_i} - \hat{\varphi}_{X_i}$ and $\Phi_{X_i,k}$ the function introduced in the proof of points 2 and 3 in Proposition 3, we get

$$y = \begin{pmatrix} \tilde{\varphi}_{X_i}(y) \\ \Phi_{X_i,d'+1}(\tilde{\varphi}_{X_i}(y)) \\ \vdots \\ \Phi_{X_i,d}(\tilde{\varphi}_{X_i}(y)) \end{pmatrix} \text{ so } \hat{\varphi}_{X_i}(y) = \tilde{\varphi}_{X_i}(y) + E_{i,n}\tilde{\varphi}_{X_i}(y) + E_{i,n}\begin{pmatrix} 0_{d'} \\ \Phi_{X_i,d'+1}(\tilde{\varphi}_{X_i}(y)) \\ \vdots \\ \Phi_{X_i,d}(\tilde{\varphi}_{X_i}(y)) \end{pmatrix}$$

and so, for all i, there exists $E_{i,n}$ a matrix such that,

$$\|E_{i,n}\|_{\mathrm{op}} \leq \frac{\sqrt{2}(d'+2)}{d}\big(4\sqrt{\ln(n)/k_n} + c\rho_n\big).$$

Then,

$$\big\|\hat{\varphi}_{X_i}(y) - (I_d - E_{i,n})\tilde{\varphi}_{X_i}(y)\big\| \leq \big(d - d'\big)\lambda_M \frac{\sqrt{2}(d'+2)}{d}\big(4\sqrt{\ln(n)/k_n} + c\rho_n\big)\big\|\tilde{\varphi}_{X_i}(y)\big\|^2.$$

That concludes the proof. □

Lemma 4. *Let $X_1, \dots, X_n$ be an i.i.d. sample drawn according to a distribution $\mathbb{P}_X$ which fulfills condition P. For a given $\lambda > 0$, introduce $I_n(\lambda) = \{i : d(X_i, \partial M) \leq \lambda(\ln n)/n, r_{i,k_n} \geq \sqrt{d(X_i, \partial M)}\}$. Denote by $\tilde{\varphi}_{X_i}$ the linear projection onto the tangent space at X_i and by $\hat{\varphi}_{X_i}$ the linear projection onto the estimation of the tangent space via local PCA. With probability greater than $1 - n^{-6}$ for n large enough, there exist a constant a and a matrices $E_{i,n}$ with $\|E_{i,n}\|_{\mathrm{op}} \leq a(\sqrt{\ln(n)/k_n} + \rho_n)$ such that, for all $i \in I_n(\lambda)$ and all $y \in \mathcal{B}(X_i, \rho_n)$ we have*:

$$\big\|\hat{\varphi}_{X_i}(y) - (I_d - E_{i,n})\tilde{\varphi}_{X_i}(y)\big\| \leq a\big(\sqrt{\ln(n)/k_n} + \rho_n\big)\big\|\tilde{\varphi}_{X_i}(y)\big\|^2.$$

Proof. The proof is exactly the same as the previous one, the only difference being now that, up to a change of basis, $r_{i,k_n}^{-2}\Sigma_i$ is no longer close to $(d'+2)^{-1}I_{d',d}$, but rather to a diagonal matrix with an eigenvalue $(d'+2)^{-1}$ eigenvalues of order $d'-1$ and $\beta_{d'} > 0$ eigenvalue of order 1. □

Acknowledgement

This research has been partially supported by MATH-AmSud grant 16-MATH-05 SM-HCD-HDD.

References

[1] E. Aamari and C. Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.* **59** (4) (2018) 923–971. MR3802310 https://doi.org/10.1007/s00454-017-9962-z

[2] E. Aamari and C. Levrard. Non-asymptotic rates for manifold, tangent space, and curvature estimation. *Ann. Statist.* **47** (1) (2019) 177–204. MR3909931 https://doi.org/10.1214/18-AOS1685

[3] C. Aaron and O. Bodart. Local convex hull support and boundary estimation. *J. Multivariate Anal.* **147** (2016) 82–101. MR3484171 https://doi.org/10.1016/j.jmva.2016.01.003

[4] C. Aaron, A. Cholaquidis and A. Cuevas. Detection of low dimensionality and data denoising via set estimation techniques. *Electron. J. Stat.* **11** (2) (2017) 4596–4628. MR3724969 https://doi.org/10.1214/17-EJS1370

[5] E. Arias-Castro, G. Lerman and T. Zhang. Spectral clustering based on local PCA. *J. Mach. Learn. Res.* **18** (2017) 1–57. MR3634876

[6] M. Berger. *A Panoramic View of Riemannian Geometry*. Springer-Verlag, Berlin, 2003. MR2002701 https://doi.org/10.1007/978-3-642-18245-7

[7] J. R. Berrendero, A. Cholaquidis, A. Cuevas and R. Fraiman. A geometrically motivated parametric model in manifold estimation. *Statistics* **48** (5) (2014) 983–1004. MR3259871 https://doi.org/10.1080/02331888.2013.800264

[8] T. Berry and T. Sauer. Density estimation on manifolds with boundary. *Comput. Statist. Data Anal.* **107** (2017) 1–17. MR3575055 https://doi.org/10.1016/j.csda.2016.09.011

[9] P. Bickel and E. Levina. *Maximum Likelihood Estimation of Intrinsic Dimension*. *Advances in NIPS* **17**. MIT Press, Cambridge, MA, 2005.

[10] F. Camastra and A. Staiano. Intrinsic dimension estimation. *Inform. Sci.* **328** (C) (2016) 26–41.

[11] G. Carlsson. Topology and data. *Bull. Amer. Math. Soc.* **46** (2) (2009) 255–308. MR2476414 https://doi.org/10.1090/S0273-0979-09-01249-X

[12] A. R. Casal. Set estimation under convexity type assumption. *Ann. Inst. Henri Poincaré Probab. Stat.* **43** (2007) 763–774. MR3252430

[13] F. Chazal, M. Glisse, C. Labruère and B. Michel. Convergence rates for persistence diagram estimation in topological data analysis. *J. Mach. Learn. Res.* **16** (2015) 3603–3635. MR3450548

[14] J. Chevalier. Estimation du support et du contour de support d'une loi de probabilité. *Ann. Inst. Henri Poincaré B, Probab. Stat.* **12** (4) (1976) 339–364. MR0451491

[15] A. Cuevas and R. Fraiman. Set estimation. In *New Perspectives on Stochastic Geometry* 366–389. W. S. Kendall and I. Molchanov (Eds). Oxford Univ. Press, Oxford, 2009. MR2654684

[16] A. Cuevas, R. Fraiman and A. Rodríguez-Casal. A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.* **35** (2007) 1031–1051. MR2341697 https://doi.org/10.1214/009053606000001532

[17] A. Cuevas and A. Rodriguez-Casal. On boundary estimation. *Adv. in Appl. Probab.* **36** (2004) 340–354. MR2058139 https://doi.org/10.1239/aap/1086957575

[18] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. *SIAM J. Numer. Anal.* **7** (1970) 1–46. MR0264450 https://doi.org/10.1137/0707001

[19] L. Devroye and G. Wise. Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM J. Appl. Math.* **3** (1980) 480–488. MR0579432 https://doi.org/10.1137/0138038

[20] H. Federer. Curvature measures. *Trans. Amer. Math. Soc.* **93** (1959) 418–491. MR0110078 https://doi.org/10.2307/1993504

[21] C. Fefferman, S. Mitter and H. Narayanan. Testing the manifold hypothesis. *J. Amer. Math. Soc.* **29** (2016) 983–1049. MR3522608 https://doi.org/10.1090/jams/852

[22] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli and L. Wasserman. Minimax manifold estimation. *J. Mach. Learn. Res.* **13** (2) (2012) 1263–1291. MR2930639

[23] C. R. Genovese, M. Perone-Pacifico, I. Verdinelli and L. Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40** (2) (2017) 941–963. MR2985939 https://doi.org/10.1214/12-AOS994

[24] M. Hein, J. Y. Audibert and U. Von Luxburg. From graphs to manifoldsweak and strong pointwise consistency of graph Laplacians. In *International Conference on Computational Learning Theory*. Springer, Berlin, Heidelberg, 2005. MR2203281 https://doi.org/10.1007/11503415_32

[25] M. Hein, J. Y. Audibert and U. Von Luxburg. Graph Laplacians and their convergence on random neighborhood graphs. *J. Mach. Learn. Res.* **8** (2007) 1325–1368. MR2332434

[26] J. Kim, A. Rinaldo, Wasserman and L. Minimax Rates for estimating the dimension of a manifold. Preprint. Available at arXiv:1605.01011. MR3918925

[27] D. O. Loftsgaarden and C. P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Stat.* **36** (3) (1965) 1049–1051. MR0176567 https://doi.org/10.1214/aoms/1177700079

[28] M. Maggioni, S. Minsker and N. Strawn. Multiscale dictionary learning: Non-asymptotic bounds and robustness. *J. Mach. Learn. Res.* **17** (2016) 1–51. MR3482922

[29] P. Niyogi, S. Smale and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** (2008) 419–441. MR2383768 https://doi.org/10.1007/s00454-008-9053-2

[30] P. Niyogi, S. Smale and S. Weinberger. A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* **40** (3) (2011) 646–663. MR2810909 https://doi.org/10.1137/090762932

[31] I. Pinelis. Extremal probabilistic problems and Hotelling's T^2 test under symmetry condition. *Ann. Statist.* **22** (1) (1994) 357–368. MR1272088 https://doi.org/10.1214/aos/1176325373

[32] T. Schick. Manifolds with boundary and of bounded geometry. *Math. Nachr.* **223** (2001) 103–120. MR1817852 https://doi.org/10.1002/1522-2616(200103)223:1&lt;103::AID-MANA103&gt;3.3.CO;2-J

[33] C. Thäle. 50 years sets with positive reach. A survey. *Surv. Math. Appl.* **3** (2008) 123–165. MR2443192