# On boundary detection

Catherine Aaron, Alejandro Cholaquidis

# On boundary detection

Catherine Aaron and Alejandro Cholaquidis
LMBP : Université Clermont-Auvergne
Universitad de la República

July 4, 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, the power is one for $n$ large enough and that there exists consistent decision rule. Heuristics for choosing a convenient value for the $k_n$ parameter and identify observations close to the boundary are also given. Finally we provide a simulation study of the test.

**Keyword: Geometric Inference, Boundary, Test, Nearest-Neighbors.**

**MSclass: 62G10,62H15.**

## 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, $\partial M$, 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 the Lebesgue measure, i.e. when the support is full dimensional. We refer to Chevalier (1976) and Devroye and Wise (1980) for precursor works on support estimation, Cuevas and Fraiman (2010) for a review on support estimation, Cuevas and Rodriguez-Casal (2004) for boundary estimation, Cuevas et al. (2007) for boundary measure estimation, Berrendero et al. (2014) for integrated mean curvature estimation or Aaron and Bodart (2016) for 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 relevance due to its connection with non-linear dimensionality reduction techniques (also known as *manifold learning*), as well as *persistent homology*. See for instance Fefferman, et al (2016), Niyogi et al. (2008), Niyogi et al. (2011). Considering support estimation it would be natural to think that some of the proposed estimators (in the full dimensional framework) are still suitable. For instance in Niyogi et al. (2008), assuming that $M$ is smooth enough, it is proved that, for $\varepsilon$ 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 Niyogi et al. (2008)). 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) this idea is hopeless (See Figure 1).

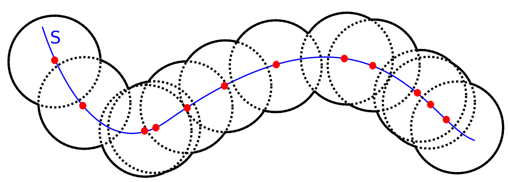


Figure 1: A one dimensional set $M$ with boundary (the two extremities of the line), sample drawn on $M$ and the associated Devroye-Wise $\hat{M}_r$ estimator of $M$, note that $\partial\hat{M}_r$ is far from $\partial M$

To our knowledge only one $d'$-dimensional support estimator exists and have only been studied recently, in the case of support with no boundary (see Aamari and Levrard (2016)). Thus the classical plug-in idea of estimating the boundary of the support using the boundary of an estimator can not be used.

Before trying to estimate the boundary of the support, in the lower dimensional case, one has to be able to decide whether it has a boundary or not. The answer provides topological information on the manifold that may be useful. For instance, if there is no boundary, the support estimator proposed in Aamari and Levrard (2016) can be used. Moreover, a compact , simply connected manifolds without boundary is homomorphic to a sphere, as it follows from the well known (and now proved)

Poincaré's conjecture. When the test decides the presence of 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 there orientability, there Euler characteristic and the number of the components of the boundary).

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

This work 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 and 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 Notations 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. The $k$-dimensional closed ball of radius $\varepsilon$ centered at $x$ will be denoted by $\mathcal{B}_k(x,\varepsilon) \subset \mathbb{R}^d$ (when $k = d$ the index will be removed) and its Lebesgue measure will be denoted as $\sigma_k = |\mathcal{B}_k(x,1)|$. When $A = (a_{ij})$, $(i = 1,\dots,m\ ,\ j = 1,\dots,n)$ is a matrix, we will write $\|A\|_\infty = \max_{i,j} |a_{ij}|$. The transpose of $A$ will be denoted $A'$. For the case $n = m$, we will denote by $\det(A)$ and $\mathrm{tr}(A)$ 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 $d'$-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 $\partial M$. For $x \in M$, $T_xM$ denotes the tangent space at $x$ and $\varphi_x$ the orthogonal projection on the affine tangent space $x+T_xM$. When $M$ is orientable it has a unique associated volume form $\omega$ such that $\omega(e_1,\dots,e_{d'}) = 1$ for all oriented orthonormal basis $e_1,\dots,e_{d'}$ of $T_xM$. Then if $g : M \to \mathbb{R}$ is a density function, we can define a new measure $\mu(B) = \int_B g\omega$, 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 work $X_1, \dots, X_n$ is an i.i.d. sample of a random variable $X$, whose probability distribution, $\mathbb{P}_X$, fulfills the condition P and that the sequence $(k_n)$ fulfills the condition K:

P. A probability distribution $\mathbb{P}_X$ fulfills condition P if there exists $M$ a compact $d'$-dimensional manifold of class $\mathcal{C}^2$ and $f$ a function such that:

1. $\partial M$ is either empty or of class $\mathcal{C}^2$,
2. for all $x \in M$, $f(x) \geq f_0 > 0$, $f$ is $K_f-$Lipschitz continuous and, for all $A \subset M$, $\mathbb{P}_X(A) = \int_A f\omega$. In the following $f_1 = \max_{x\in M} f(x)$.

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

**Definition 1.** *Given an i.i.d. sample $X_1, \dots, X_n$ of a random variable $X$ with support $M \subset \mathbb{R}^d$, where $M$ is $d'$-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 $i = 1, \dots, n$,*

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

*Consider now $Q_{i,k_n}$ the $d'$-dimensional plane spanned by the $d'$ eigenvectors of $\hat{S}_{i,k_n}$ associated to the $d'$ largest eigenvalues of $\hat{S}_{i,k_n}$. Let $X^*_{k(i)}$ be the normal projection of $X_{k(i)} - X_i$ on $Q_{i,k_n}$ and $\overline{X}_{k_n,i} = \frac{1}{k_n}\sum_{j=1}^{k_n} X^*_{j(i)}$.*

*Let us 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:*

$$\Delta_{n,k_n} = \max_i \delta_{i,k_n}.$$

Let us explain the heuristic behind the test we will propose. It will be proved that, under conditions P. and K. we have $r_n \xrightarrow{a.s.} 0$. Let us consider an observation $X_{i_0}$ such that $d(X_{i_0}, \partial M) \geq r_{i_0,k_n}$. Regularity of the manifold and continuity of the density given by condition P will entail that the sample $\{r_{i_0,k_n}^{-1} X_{1(i_0)}^*, \ldots, r_{i_0,k_n}^{-1} X_{k_n(i_0)}^*\}$ "converges" toward a uniform sample on $\mathcal{B}_{d'}(0,1)$ and then $\|\overline{X}_{k_n,i_0}\| r_{i_0,k_n}^{-1} \xrightarrow{a.s.} 0$. It will also be proved that $\delta_{i_0,k_n} \longrightarrow \chi_2(d')$ in distribution. If $\partial M = \emptyset$ all the observations satisfies $d(X_i, \partial M) \geq r_{i,k_n}$. Even though the $\{\delta_{i,k_n}\}_i$ are not independent we will obtain an asymptotic result on the $\Delta_{n,k_n}$ that involves the $\chi_2(d')$ distribution. If $\partial M \neq \emptyset$ and we consider a point $X_{i_0}$ such that $d(X_{i_0}, \partial M) \ll r_{i_0,k_n}$ (conditions P. and K. will ensure the a.s. existence of such a point) the sample $\{r_{i_0,k_n}^{-1} X_{1(i_0)}^*, \ldots, r_{i_0,k_n}^{-1} X_{k_n(i_0)}^*\}$ "converges" to a uniform sample on $\mathcal{B}_{d'}(0,1) \cap \{x : \langle u, x \rangle \geq 0\}$ and $\|\overline{X}_{k_n,i_0}\| r_{i_0,k_n}^{-1} \xrightarrow{a.s.} a_{d'} > 0$. Asymptotic behavior of the test statistic is given in the following four theorems. The first theorem provides a bound for the $p$-value when testing $H_0 : \partial M = \emptyset$ versus $H_1 : \partial M \neq \emptyset$ using the test statistic $\Delta_{n,k_n}$ and rejection region $\{\Delta_{n,k_n} \geq t_n\}$ for some suitable $t_n$. The second theorem states that, under $H_0$, the empirical distribution of $\delta_{i,k_n}$ converges in mean square towards a $\chi^2(d')$ distribution. We will use this result to choose the parameter $k_n$ (see Section 3.2). The third theorem states that, with probability one, the power of the test is one for $n$ large enough. The last one provides a consistent decision rule.

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

$$\begin{cases} H_0 : & \partial M = \emptyset \\ H_1 : & \partial M \neq \emptyset \end{cases} \tag{1}$$

*with the rejection zone*

$$W_n = \left\{ \Delta_{n,k_n} \geq F_{d'}^{-1}(9\alpha/(2e^3 n)) \right\}, \tag{2}$$

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

**Theorem 2.** *Let $k_n$ be a sequence fulfilling condition* K. *Let us assume that $X_1, \ldots, X_n$ is an i.i.d. sample drawn according to an unknown distribution $\mathbb{P}_X$ which fulfills condition* P *with $\partial M = \emptyset$. If we define*

$$\hat{\Psi}_{n,k_n}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{\{\delta_{i,k_n} \leq x\}},$$

*then, for all $x \in M$,*

$$\mathbb{E}\big(\hat{\Psi}_{n,k_n}(x) - \Psi_{d'}(x)\big)^2 \to 0 \qquad \text{as } n \to \infty.$$

**Theorem 3.** *Let $k_n$ be a sequence fulfilling condition* K. *Let us 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 4.** *Let $k_n$ be a sequence fulfilling condition* K. *Let us 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. *Then, with probability one, the decision rule: $\partial M = \emptyset$ if and only if $\Delta_{n,k_n} \leq \beta_n$ with $\lambda \ln n \leq \beta_n \leq \mu k_n$ with $\lambda > 4$ and $\mu \leq (d'+2)\left(\frac{\Gamma\left(\frac{d'+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d'+3}{2}\right)}\right)^2$ is consistent.*

## 3.2 Automatic choice for $k_n$

Theorem 2 ensures that when $\partial M = \emptyset$, the empirical distribution of $\delta_{i,k_n}$ converges to a $\chi^2(d')$ distribution. One can easily conjecture that when $\partial M \neq \emptyset$ the distribution of $\delta_{i,k_n}$ conditioned to the points $X_i$ "far enough" from the boundary also converges to a $\chi^2(d')$ distribution. We define $d_{\chi^2}(k)$ as follows:

i. If the estimated $p$-value (using $k$-nearest neighbors) is greater than $\alpha$ ($H_0$ is decided) compute:

$$d_{\chi^2}(k) = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{\Psi}_{n,k}(\delta_{i,k}) - \Psi_{d'}(\delta_{i,k})\right|.$$

ii. If the estimated $p$-value is less than $\alpha$, first identify the points "far from the boundary" as the observations $i \in I_k = \{F_{d'}(\delta_{i,k}) \geq \alpha\}$. Then, if we define

$$\hat{\psi}_{\alpha,n,k}(x) = \frac{1}{\#I_k}\sum_{i\in I_k}\mathbb{I}_{\{\delta_{i,k}\leq x\}},$$

compute

$$d_{\chi^2}(k) = \frac{1}{\#I_k}\sum_{i\in I_k}\left|\hat{\Psi}_{\alpha,n,k}(\delta_{i,k}) - \Psi_{\alpha,d'}(\delta_{i,k})\right|,$$

where $\Psi_{\alpha,d'}(x) = (1-\alpha)^{-1}\Psi_{d'}(x)\mathbb{I}_{\{\Psi_{d'}(x)\leq 1-\alpha\}}$.

Finally choose $k = \text{argmin}_k d_{\chi^2}(k)$. In practice we choose $\alpha = 0.05$.

### 3.3 Discussion on the hypotheses

We assume that the dimension, $d'$, is known. In practice it can be estimated using a dimension estimation method. Estimation of the intrinsic dimension has been widely studied, (see Camastra and Staiano (2016) for a review).

The noiseless assumption, i.e., the support is a lower dimensional manifold, can not be changed by a noisy model, that is the support is "around" a lower dimensional manifold, with our approach. To see this, let us consider that the support is $M \oplus \varepsilon\mathcal{B} = \{x, d(x, M) \leq \varepsilon\}$ (with $M$ a lower dimensional manifold) our test will asymptotically decide that $M \oplus \varepsilon\mathcal{B}$ is a manifold with boundary. However this case is not hopeless. Indeed, if were able to find a functional sequence $\varphi_n$ such that $\varphi_n(\mathcal{B}(M, \varepsilon)) \subset \mathcal{B}(M, \varepsilon_n)$ with $\varepsilon_n \to 0$ "quickly enough" (i.e. such that $\varepsilon_n/(\min_i r_{i,k_n}) \xrightarrow{a.s.} 0$) and such that the distribution of $\varphi_n(X)$ converges toward a distribution that satisfies the condition P, one could probably apply our test on the sample $\{Y_1, \ldots, Y_n\}$ where $Y_i = \varphi_n(X_i)$. Note that such a "de-noising" process is a current research topic, see for instance Aaron et al. (2017) where a de-noising process is proposed (unfortunately with no guarantee on the existence and regularity of the limit distribution).

Smoothness of the support is necessary for the proposed test. One can imagine that, when the support has no boundary but is not smooth enough, the proposed test will reject the null hypothesis. Indeed, let us consider the case $d = 2$ and a uniform sample on the boundary of the unit square $[0, 1] \times [0, 1]$, see Figure 2 left. For observations near a corner, the normalization parameter should be $r_{i,k_n}/\sqrt{2}$ instead of $r_{i,k_n}$. In a polyhedron, when a corner becomes acute, the local $PCA$ fails to estimate a "tangent" plane at the corner, see Figure 2 right.

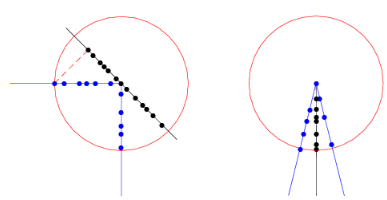

Figure 2: Behavior for polyhedron. When the angle does not allow to estimate the "tangent" plane the normalization is not suitable. When the angle is too acute the projection is not accurate. The manifold, and sample points are in blue, the estimated tangent plane and projected observations are in black.

The continuity of the density is also necessary: if this is not the case, we may reject $H_0$ for any supports with or without boundary. In order to see this, let us 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 correspond to a "boundary-type" behavior. Although we will assume in general that $f$ is bounded away from zero, this can weakened by asking that $f(x) \geq d(x, \partial M)^\alpha$ for some $\alpha > 0$, for the sake of simplicity in the notation, and length of the calculus we kept the hypothesis $f > 0$. By contrast, the $\mathcal{C}^2$ smoothness of the boundary (if it exists), can be weakened. The proofs of Theorems 3 and 4 are similar (just a bit more complicated to write) when only a part of the boundary is $\mathcal{C}^2$ (namely if there exists $x \in \partial M$ and $r > 0$ such that $\partial M \cap \mathcal{B}(x,r)$ is a $\mathcal{C}^2$ manifold).

### 3.4 Numerical simulations

We now present some results for different manifolds. First, we study the behavior of our test for a sample with uniform distribution on $S_{d'}$, the $d'-$dimensional sphere in $\mathbb{R}^{d'+1}$ and on $S_{d'}^+$ the $d'-$dimensional half-sphere in $\mathbb{R}^{d'+1}$. We also present some results for manifolds with non constant curvature, such as the trefoil knot ($d' = 1$ and $d = 3$), a spiral, a Möebius ring, and a torus (for these two last examples the samples are not uniform).

First we observe that the proposed rule to find a suitable value for $k$ is practically efficient. Here we choose the sample size $n = 3000$. In Figure 3 we present results for supports without boundary. Two curves are plotted, the estimated $p-$value (red) and $d_{\chi^2}$ (blue). In order to have comparable curves $d_{\chi^2}$ has been artificially normalized to be in $[0,1]$. Notice that each time, at the selected value for $k$, i.e. $k = \text{argmin}(d_{\chi^2})$, the estimated $p-$value is large enough to accept $H_0$ (the support has no boundary). In Figure 4 we present the result of the same experiment but for support with boundary. On the first line of the figure the curves of the estimated $p-$value and $d_{\chi^2}$ are presented. Here also the choice of $k = \text{argmin}(d_{\chi^2})$ allows us to decide well (i.e. here to reject $H_0$). On the second line of the figure we draw the sample point and underline the points $X_i$ such that $\frac{2e^3}{9} F_{d'}(\delta_{i,k}) \leq 0.05$ that is the one that are expected be located "near to" the boundary.

In Table 3.4 we present estimated level and power of the proposed test. For each example and each sample size we drew 2000 samples. It can be observed that, when the support has no boundary the percentage of rejection (i.e. the level) is less than 5% if $n \geq 500$ for every example. When the support has boundary, the percent of rejection (i.e. here the power) converges quickly to 100%. To shorten the computational time we chose $k_n$ by averaging the one obtained with the $d_{\chi^2}$ criteria with 50 samples (for each example and each sample size). The selected $k_n$ are given

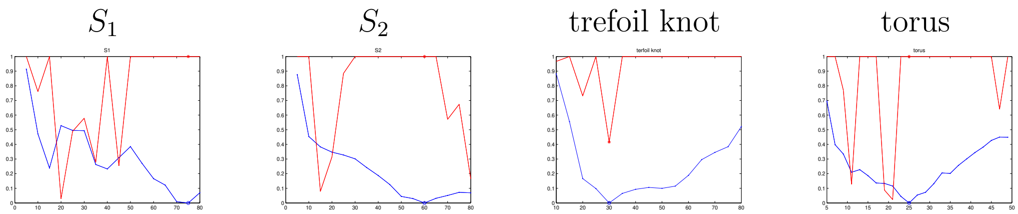


Figure 3: Some examples for support without boundary support. Abscissa: $k$, blue: $d_{\chi^2}(k)$, red: $\hat{p}_v(k)$.

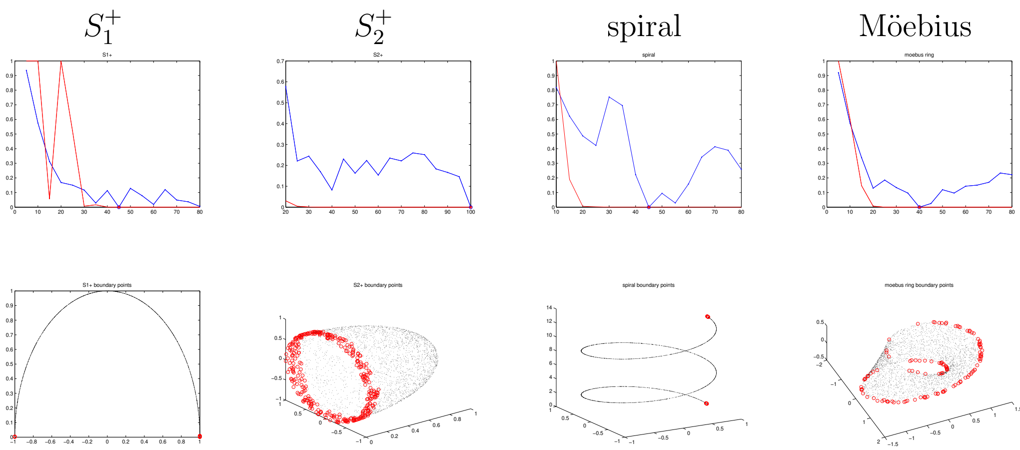


Figure 4: Some examples for support with boundary. First line: Abscissa: $k$, blue: $d_{\chi^2}(k)$, red: $\hat{p}_v(k)$. Second line: the associated sample and points that are identified as “close to the boundary”

in the Table.

# 4 Empirical detection of points close to the boundary and estimation of the number its the 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. A third step, to get an insight of the topological properties of the boundary could be to estimate the number of its connected components. In this section we will empirically tackle both problems.

| $n=$ | 100 | 500 | $10^3$ | $2.10^3$ | $n=$ | 100 | 500 | $10^3$ | $2.10^3$ |
|---|---|---|---|---|---|---|---|---|---|
| $S_1$ | $k=15$ $1,45\%$ | $k=20$ $1,05\%$ | $k=35$ $1\%$ | $k=40$ $0,9\%$ | $S_1^+$ | $k=15$ $89,25\%$ | $k=20$ $60,7\%$ | $k=35$ $97,1\%$ | $k=40$ $99,3\%$ |
| $S_2$ | $k=15$ $3\%$ | $k=20$ $1,6\%$ | $k=25$ $1,4\%$ | $k=30$ $1,35\%$ | $S_2^+$ | $k=17$ $84,8\%$ | $k=30$ $100\%$ | $k=50$ $100\%$ | $k=50$ $100\%$ |
| $S_3$ | $k=6$ $1,2\%$ | $k=15$ $1,9\%$ | $k=17$ $1,35\%$ | $k=25$ $1,85\%$ | $S_3^+$ | $k=6$ $2,35\%$ | $k=10$ $5,55\%$ | $k=15$ $34,45\%$ | $k=25$ $99,95\%$ |
| $S_4$ | $k=5$ $0,75\%$ | $k=10$ $2,3\%$ | $k=17$ $1,15\%$ | $k=17$ $3,15\%$ | $S_4^+$ | $k=5$ $1\%$ | $k=10$ $10,8\%$ | $k=80$ $100\%$ | $k=80$ $100\%$ |
| Trefoil Knot | $k=8$ $4,7\%$ | $k=15$ $2,4\%$ | $k=25$ $2,15\%$ | $k=30$ $1,45\%$ | Spire | $k=15$ $55,5\%$ | $k=25$ $92,4\%$ | $k=25$ $83,9\%$ | $k=40$ $100\%$ |
| Torus | $k=8$ $5,6\%$ | $k=15$ $5\%$ | $k=17$ $2,65\%$ | $k=20$ $1,75\%$ | Möebius ring | $k=8$ $12,2\%$ | $k=15$ $68,75\%$ | $k=20$ $98,65\%$ | $k=40$ $100\%$ |

Table 1: For different samples, the chosen $k_n$ value and the % of times where $H_0$ is rejected (on 2000 replications).

## 4.1 Detection of "boundary observations"

From Theorem 1, the natural idea is to select $\{X_i : \delta_{i,k_n} \geq F_{d'}^{-1}(9\alpha/(2ne^3))\}$ as "boundary observation". However, as it is illustrated in Figure 4, sometimes it gives "too many" boundary observations (as in the half sphere) and sometimes "too few"(as in the Möebius ring). To overcome this, we will adapt, using tangent spaces, the method given in Aaron et al. (2017), to detect "boundary balls".

Introduce $\phi_x$ is the orthogonal projection on the tangent plane and choose $r_x > 0$ small enough to ensure that $\varphi_x$ is one to one on $\mathcal{B}(x, r_x) \cap M$. As $\partial\varphi_x(M \cap \mathcal{B}(x, r_x)) = \varphi_x(\partial M \cap \mathcal{B}(x, r_x)) \cup \varphi_x(M \cap \partial\mathcal{B}(x, r_x))$ we have:

$$x \in \partial M \Leftrightarrow 0 \in \partial\varphi_x(M \cap \mathcal{B}(x, r_x)). \tag{3}$$

This suggest the following extension of the definition of boundary ball introduced in Aaron et al. (2017) using the notations introduced in Definition 1.

**Definition 2.** *$X_i$ is the centre of a $(k_n, \varepsilon_n)$-tangential boundary ball if*

$$r_i = \max\{\|x\| : \|x\| \leq \|x - X^*_{j(i)}\|,\ \forall\ 1 \leq j \leq k_n\} \geq \varepsilon_n.$$

Indeed, recall first that $X^*_{j(i)}$ is a PCA estimator of $\varphi_{X_i}(X_{j(i)})$ and that $X^*_{1(i)} = 0$ so that, by a plug-in of (3) we decide that $X_i$ is a boundary point of $M$ if 0 is a boundary point of an estimator of $\varphi_{X_i}(X_{j(i)})$ that is if $0 = X^*_{1(i)}$ is the centre of a boundary ball of $\{X^*_{1(i)}, \dots, X^*_{k_n(i)}\}$. The choice of $k_n$ in section 3.2 is still suitable since it allows the local PCA procedure to converge. We can also propose to chose the $\varepsilon_n = 2\max_i \min_j \|X_i - X_j\|$ as proposed in Aaron et al. (2017), then to identify boundary points as the center of $(k_n, \varepsilon_n)$-tangential boundary balls.

## 4.2 Building a "boundary graph"

Let us introduce $\mathcal{Y}_m = \{Y_1, \dots, Y_m\}$ the set of the centers of the $(k_n, \varepsilon_n)$-tangential boundary balls. We aim to construct a graph with vertices $\mathcal{Y}_m$, building edges between the vertexes such that the obtained graph capture the shape of the boundary. To do that we are going to "connect" each $Y_i$ to the $Y_j$ such that $\|Y_i - Y_j\| \leq r_i$, as usual the choice of $r_i$ depends on a balance, $r_i$ should be small enough to connect a point only with its neighbors but also large enough to allows to capture the global structure. In our case we are going to use the fact that that, under our hypotheses, if $\partial M \neq \emptyset$ then it is a $\mathcal{C}^2$, $(d'-1)$-dimensional manifold without boundary. In other terms for any point $Y_i$, $\{Y_j, \|Y_i - Y_j\| \leq r_i\}$ should look like an uniform drawn on the $d'-1$ dimensional ball $\mathcal{B}_{d'-1}(Y_i, r_i)$ and as a consequence $Y_i$ should be "surrounded" by the points of $\{Y_j, 0 < \|Y_i - Y_j\| \leq r_i\}$.

We propose to say that $Y_i$ is "surrounded" by $\{Y_j, 0 < \|Y_i - Y_j\| \leq r_i\}$ if $\pi_{i,r_i}(Y_i)$ belong to the interior of the convex hull of $\{\pi_{i,r_i}(Y_j), 0 < \|Y_i - Y_j\| \leq r_i\}$, where $\pi_{i,r_i}$ is the normal projection on the $(d'-1)$ first axis of a PCA computed on $\{Y_j, \|Y_i - Y_j\| \leq r_i\}$. Then we propose to chose $r_i$ as the smallest value such that all $Y_i$ is "surrounded" by $\{Y_j, 0 < \|Y_i - Y_j\| \leq r_i\}$.

## 4.3 Some experiments

To illustrate the procedures introduced before we considered the Möebius ring and the truncated cylinder with a hole in a cap, (see Figure 4.3). Both are 2-dimensional sub-manifolds of $\mathbb{R}^3$. The boundary of the first one has 1 connected component while the boundary of the second one has 3. The parameter $k$ is chosen using the method proposed in Section 3.1 and as proposed in previous section we choose $\varepsilon = 2\max_i \min_j \|X_i - X_j\|$ for the tangential boundary ball detection.. As expected, in the cylinder the sample size required to have a "coherent" graph is higher.

Second we consider uniform draws of sizes $n \in \{500, 1000, 2000, 4000, 8000, 16000\}$, on the $(d-1)$-dimensional half sphere $\{x_1^2 + \ldots + x_d^2 = 1, x_d \geq 0\} \subset \mathbb{R}^d$ for $d = \{3, 4, 5\}$. Let us 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 Monte-Carlo method drawing 50000 points on $\partial M$. For each value of $n$ and $d$, the box-plot over 50 repetitions of the $p$-values of the test and

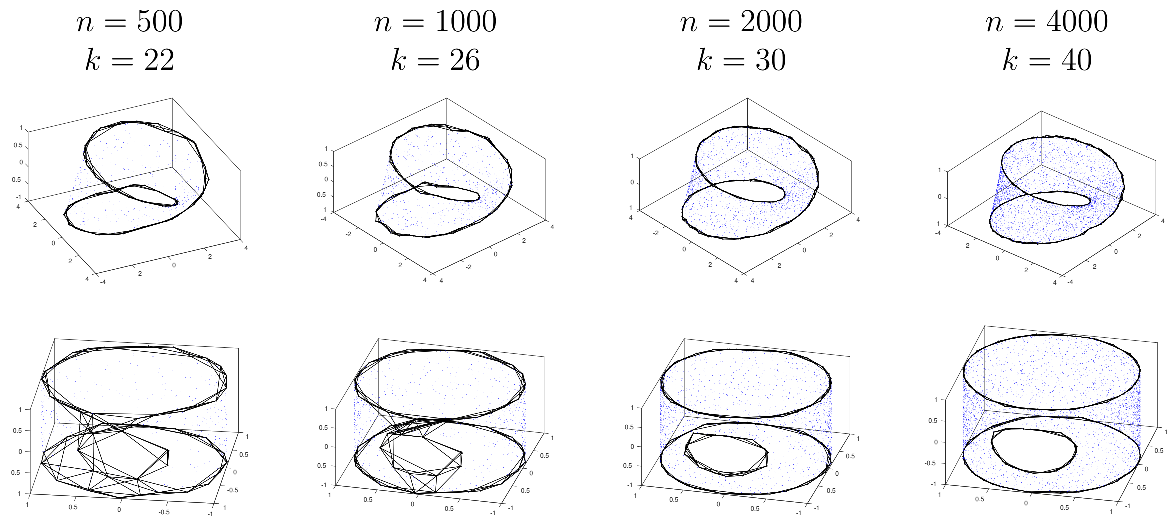


Figure 5: Boundary ball detection and associated graph for different sample sizes. In the first row the Möebius Ring and in the second the truncated cylinder with a hole in a cap. Observations are represented as blue dots while boundary centers are large black dots, the graph is represented as black lines

the estimations of $d_1$ and $d_2$ are shown in Figures 6, 7 and 8, for $d = 3, 4$ and 5 respectively.

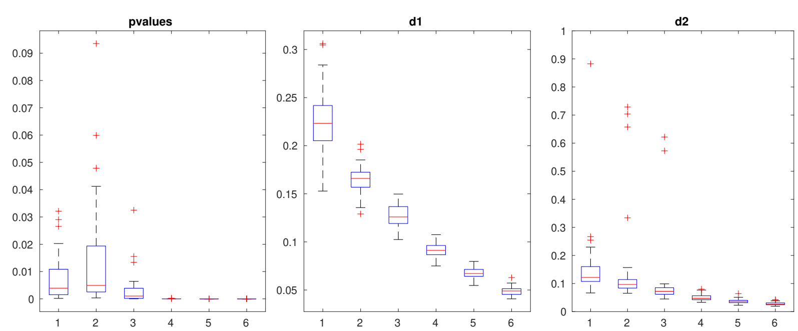


Figure 6: $d = 3$, in abscissa $1 : (n = 500, k = 25)$, $2 : (n = 1000, k = 25)$, $3 : (n = 2000, k = 30)$,$4 : (n = 4000, k = 40)$, $5 : (n = 8000, k = 50)$, $6 : (n = 16000, k = 50)$

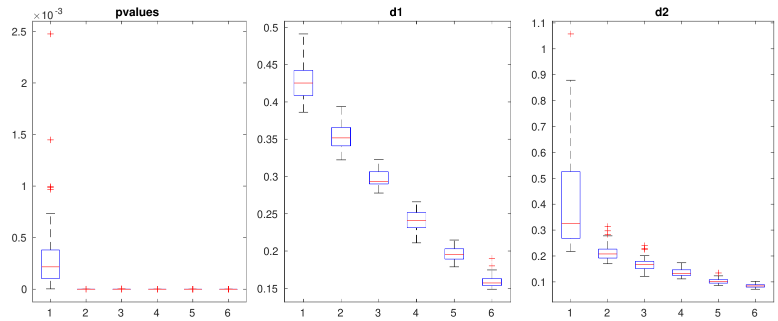


Figure 7: $d = 4$, in abscissa $1 : (n = 500, k = 30)$, $2 : (n = 1000, k = 50)$, $3 : (n = 2000, k = 50)$,$4 : (n = 4000, k = 60)$, $5 : (n = 8000, k = 70)$, $6 : (n = 16000, k = 70)$

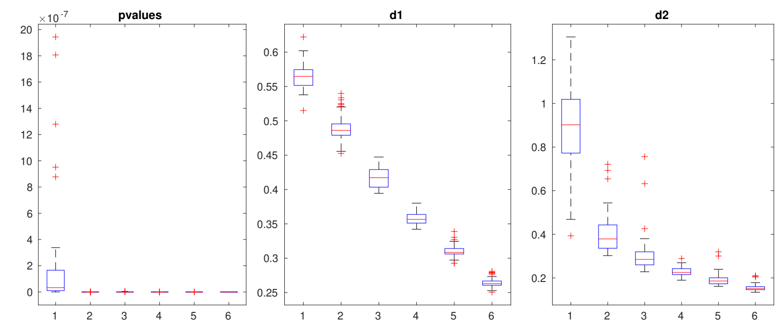


Figure 8: $d = 5$, in abs ice $1 : (n = 500, k = 50)$, $2 : (n = 1000, k = 70)$, $3 : (n = 2000, k = 80)$,$4 : (n = 4000, k = 90)$, $5 : (n = 8000, k = 100)$, $6 : (n = 16000, k = 100)$

# 5 Proofs

## 5.1 Preliminary results

In this section we settle some geometric definitions, notation and properties of compact and smooth enough manifolds that will be used in the rest of the paper. Even though some of them are well known we will give the proofs in the appendix, for the sake of completeness.

### 5.1.1 Geometric Background

Let $M \subset \mathbb{R}^d$ be a compact $\mathcal{C}^2$ $d'$-manifold with either $\partial M = \emptyset$ or $\partial M$ is a $\mathcal{C}^2$ $(d'-1)$-manifold. For $x \in M$ we denote $N_x M$ the normal plane of $M$ at $x$. For $x \in \partial M$ we denote $u_x$ the unit normal outer vector to $\partial M$. Let us denote $\varphi_x : M \to x + T_x M$ the orthogonal projection onto the tangent affine plane.

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

1. *For all $x \in M$, $\varphi_x$ is a $\mathcal{C}^2$ bijection from $M \cap \mathcal{B}(x,r)$ to $\varphi_x\big(M \cap \mathcal{B}(x,r)\big)$ for all $r \leq r_M$.*
2. *For all $\|x-y\| \leq r_M$ ($x \in M$ and $y \in x + T_x M$), let $J_x(y)$ be the Jacobian matrix of $\varphi_x^{-1}$ and $G_x(y) = \sqrt{\det(J_x'(y)J_x(y))}$, then $|G_x(y) - 1| \leq c_M \|x-y\|$*
3. *For all $x, y \in M$, $\|x-y\| \leq r_M$ then $\|\varphi_x(y) - y\| \leq c_M \|x - \varphi_x(y)\|^2 \leq c_M \|x-y\|^2$*
4. *For all $x \in M$, if $d(x, \partial M) \geq r$:*
$$\mathcal{B}(x, r - c_M r^2) \cap (x + T_x M) \subset \varphi_x(\mathcal{B}(x,r) \cap M) \subset \mathcal{B}(x,r) \cap (x + T_x M). \quad (4)$$
5. *For all $x \in \partial M$, if $d(x, \partial M) < r$, let us 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,*
$$H_x^- \cap \mathcal{B}(x, r - c_M r^2) \cap (x + T_x M) \subset \varphi_x(\mathcal{B}(x,r) \cap M) \subset H_x^+ \cap \mathcal{B}(x,r) \cap (x + T_x M). \quad (5)$$

Let us recall the change of variable formula :

$$V \subset \mathcal{B}(x, r_{0,M}) \Rightarrow \mu(V) = \int_V f dw = \int_{\varphi_x(V)} f(\varphi_x^{-1}(y)) \sqrt{\det G_x(y)} dy. \quad (6)$$

From (6) and Proposition 1 we will prove (see Section 6.2):

**Corollary 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. *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 $d(x, \partial M) \geq r$, $\left|\mathbb{P}_X(\mathcal{B}(x,r)) - f(x)\sigma_{d'} r^{d'}\right| \leq C r^{d'+1}$.*

That in turns entails 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. *Let $k_n$ be a sequence of positive integers such that $k_n \to +\infty$ and $(\ln(n))k_n^{1+d}/n \to 0$. Then, $k_n r_n \overset{a.s.}{\to} 0$, where $r_n$ was introduced in Definition 1.*

### 5.1.2 Local PCA process

The following result, whose proof is given in Section 6.3, useful to obtain the uniform convergence rate of the local $PCA$ process to the tangent planes. Let us denote $\mathcal{M}_d(\mathbb{R})$ the $d \times d$ 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 $S \in \mathcal{M}_d(\mathbb{R})$ let us denote $S = Q_S \Delta_S Q_S'$, $\Delta_S$ being diagonal with $(\Delta_S)_{1,1} \geq (\Delta_S)_{2,2} \geq \ldots \geq (\Delta_S)_{d,d}$ and $Q_S$ is the matrix containing (in column) an orthonormalized basis of eigenvectors of $S$. Introduce now $P_{S,d'} = Q_S I_{d',d} Q_S'$ that is the matrix of the the orthogonal projection on the plane spanned by the $d'$ eigenvectors associated to the $d'$ largest eigenvalues of $S$. Notice that $P_{I_{d',d},d'} = I_{d',d}$

**Proposition 2.** *Let $\Delta \in \mathcal{M}_{d'}(\mathbb{R})$ be a diagonal matrix whose eigenvalues, $\lambda$, fulfills that there exists $\lambda_0 > 0$, such that $\lambda \geq \lambda_0$. Let $D = \begin{pmatrix} \Delta & 0 \\ 0 & 0 \end{pmatrix} \in M_d(\mathbb{R}^d)$. Let us define $c_0 = 3d^{3/2}/(2\lambda_0)$. There exists $\varepsilon_0$ (depending only on $\lambda_0$ and $d$) such that for all $\varepsilon \leq \varepsilon_0$, and all symmetric matrix $S$ fulfilling $\|S - D\|_\infty \leq \varepsilon$ we have: $\|(P_{S,d'} - I_{d',d})X\|_2 \leq c_0 \varepsilon \|X\|_2$, for all $X \in \mathbb{R}^d$.*

## 5.2 Proof of Theorems 1 and 2

In order to state now two probabilistic results we will introduce the following functions, for $\varepsilon > 0$ and $k, d \in \mathbb{N}$,

$$H_k(\varepsilon) = \exp\left(-\frac{k\varepsilon^{\frac{2}{3}}(d+2)^{-\frac{4}{3}}}{d^2\left(k^{\frac{1}{3}} + (d+2)^{\frac{1}{3}}\varepsilon^{\frac{1}{3}}\right)^2}\right), \qquad R_k(\varepsilon) = \exp\left(-\frac{k^{\frac{1}{3}}\varepsilon^{\frac{2}{3}}}{d^2(d+2)^{\frac{4}{3}}}\right),$$
$$G_k(t) = \min_{\varepsilon \in [0,t]} \left(\frac{2e^3}{9} F_d(t-\varepsilon) + (d^2+d)H_k(\varepsilon) + 2dR_k(\varepsilon)\right).$$

**Proposition 3.** *Let $k_n$ be a sequence such that $k_n \gg (\ln n)^4$. Then*

*i.* *For all $\lambda > 2$, $nG_{k_n}(\lambda \ln(n)) \to 0$.*

*ii.* *If we define $t_n(\alpha) = F^{-1}(9\alpha/(2e^3 n))$, then $nG_{k_n}(t_n(\alpha) + o(1)) \leq \alpha + o(1)$.*

*iii.* *For all $\lambda > 4$, $\sum_n nG_{k_n}(\lambda \ln n) < +\infty$.*

*Proof.* If we use a standard expansion of the incomplete Gamma function we get $F_d(x) \sim e^{-x/2}(1 + x/2)^{d/2-1}/\Gamma(d/2)$. By definition, for any sequence $\varepsilon_n \in [0, t_n(\alpha)]$;

$$G_{k_n}(t_n(\alpha)) \leq \left( \frac{2e^3}{9} F_d(t_n(\alpha) - \varepsilon_n) + (d^2 + d)H_{k_n}(\varepsilon_n) + 2dR_{k_n}(\varepsilon_n) \right).$$

Finally *i.* and *ii.* follow by taking the sequence $\varepsilon_n = \varepsilon$ for all $n$, and *iii.* follows from $\varepsilon_n = \frac{\lambda-4}{2} \ln(n)$. □

**Lemma 2.** *Let $X_1, \ldots, X_n$ be an i.i.d. sample uniformly drawn on $\mathcal{B}(x, r) \subset \mathbb{R}^d$ and let us denote $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$. We have:*

$$\frac{(d+2)n\|\overline{X}_n - x\|^2}{r^2} \xrightarrow{\mathcal{L}} \chi^2(d), \tag{7}$$

*and, for all $n > d$*

$$\mathbb{P}\left( \frac{(d+2)n\|\overline{X}_n - x\|^2}{r^2} \geq t \right) \leq G_n(t). \tag{8}$$

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

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

$$f(x) = \frac{1}{\sigma_d} \sigma_{d-1}(1 - x^2)^{(d-1)/2} \mathbb{I}_{[-1,1]}(x), \tag{9}$$

and then

$$\mathrm{Var}(X_{.,i}) = \int_{-1}^{1} x^2 \frac{1}{\sigma_d} \sigma_{d-1}(1 - x^2)^{(d-1)/2} dx = \frac{\sigma_{d-1}}{\sigma_d} B\big(3/2, (d+1)/2\big),$$

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

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

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

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

Now, to prove (7) observe that

$$(d+2)n\|\overline{X}_n\|^2 = \left(\sqrt{n(d+2)}\frac{1}{n}\sum_{i=1}^{n} X_{i,1}\right)^2 + \cdots + \left(\sqrt{n(d+2)}\frac{1}{n}\sum_{i=1}^{n} X_{i,d}\right)^2.$$

For all $k = 1, \ldots, d$, by the Central Limit Theorem, $\left(\sqrt{n(d+2)}\frac{1}{n}\sum_{i=1}^{n} X_{i,k}\right)^2 \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)^2$. This, together with the independence of the $Y_k = \left(\sqrt{n(d+2)}\frac{1}{n}\sum_{i=1}^{n} X_{i,k}\right)^2$ concludes the proof of (7).

In order to prove (8), let us denote by $\hat{S}_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$ the empirical covariance matrix of the observations and by $\Sigma^2 = \frac{1}{d+2} I_d$ the real covariance matrix. We can express our statistic as : $n\overline{X}_n' \Sigma^{-2} \overline{X}_n$. Now if we use equation (7) in Bertail et al. (2008), for all $n > d$

$$\mathbb{P}\Big(n\overline{X}_n' \hat{S}_n^{-2} \overline{X}_n > t\Big) \leq \frac{2e^3}{9} F_d(t). \tag{10}$$

Let us denote $\Gamma_n = \Sigma^{-2} - \hat{S}_n^{-2}$. We have

$$\mathbb{P}\Big(n\overline{X}_n' \Sigma^{-2} \overline{X}_n > t\Big) = \mathbb{P}\Big(n\overline{X}_n' \hat{S}_n^{-2} \overline{X}_n + n\overline{X}_n' \Gamma_n \overline{X}_n > t\Big),$$

then,

$$\mathbb{P}\Big(n\overline{X}_n' \hat{S}_n^{-2} \overline{X}_n > t\Big) \leq \min_{\varepsilon \in [0,t]} \Big(\mathbb{P}\Big(n\overline{X}_n' \hat{S}_n^{-2} \overline{X}_n \geq t - \varepsilon\Big) + \mathbb{P}\Big(n\overline{X}_n' \Gamma_n \overline{X}_n > \varepsilon\Big)\Big)$$

and applying (10),

$$\mathbb{P}\Big(n\overline{X}_n' \Sigma^{-2} \overline{X}_n > t\Big) \leq \min_{\varepsilon \in [0,t]} \left(\frac{2e^3}{9} F_d(t-\varepsilon) + \mathbb{P}\Big(n\overline{X}_n' \Gamma_n \overline{X}_n > \varepsilon\Big)\right). \tag{11}$$

In order to prove (8), it remains to bound $\mathbb{P}(n\overline{X}_n' \Gamma_n \overline{X}_n > \varepsilon)$. First with a rough bound we get $n\overline{X}_n' \Gamma_n \overline{X}_n \leq d^2 n \|\Gamma_n\|_\infty \|\overline{X}_n\|_\infty^2$. Thus

$$\mathbb{P}\Big(n\overline{X}_n' \Gamma_n \overline{X}_n > \varepsilon\Big) \leq \mathbb{P}(d^2 n \|\Gamma_n\|_\infty \|\overline{X}_n\|_\infty^2 > \varepsilon),$$

and then,

$$\mathbb{P}\big(n\overline{X}_n'\Gamma_n\overline{X}_n > \varepsilon\big) \leq \min_{a>0}\left(\mathbb{P}\left(\|\Gamma_n\|_\infty > a\right) + \mathbb{P}\left(\|\overline{X}_n\|_\infty^2 > \frac{\varepsilon}{nd^2a}\right)\right). \tag{12}$$

Now, let us bound $\mathbb{P}(\|\Gamma_n\|_\infty > a)$. If we denote $E_n = \Sigma^2 - \hat{S}_n^2$, then, applying Hoeffding's inequality for all $i, j$ we get that, for all $a' > 0$, $\mathbb{P}(|E_{i,j}| > a') \leq 2\exp(-na'^2)$ and so:

$$\mathbb{P}(\|E_n\|_\infty > a) \leq d(d+1)\exp(-na^2), \tag{13}$$

where we have used that $E_n$ is symmetric and the maximum value of the $d(d+1)/2$ terms is considered in the norm. Notice now that, if $\|E_n\|_\infty < (d(d+2))^{-1}$, then:

$$\hat{S}_n^2 = \frac{1}{d+2}\big(I_d - (d+2)E_n\big) \implies \hat{S}_n^{-2} = (d+2)\sum_{k=0}^{+\infty}(d+2)^k E_n^k.$$

Finally, using that $\|E_n^k\|_\infty \leq d^k\|E_n\|_\infty^k$, we get

$$\|\Gamma_n\|_\infty \leq \frac{d(d+2)^2\|E_n\|_\infty}{1 - d(d+2)\|E_n\|_\infty}. \tag{14}$$

Therefore, for all $a > 0$,

$$\|\Gamma_n\| > a \quad \text{if and only if} \quad \|E_n\|_\infty > \frac{a}{d(d+2)(a+d+2)}. \tag{15}$$

Since $a > 0$ we have $\frac{a}{d(d+2)(a+d+2)} \leq \frac{1}{d(d+2)}$. Combining (13) and (14) we obtain:

$$\mathbb{P}(\|\Gamma_n\|_\infty > a) \leq d(d+1)\exp\left(-\frac{na^2(d+2)^{-2}}{d^2(a+d+2)^2}\right). \tag{16}$$

To finish, we perform the same kind of calculus on $\mathbb{P}(\|\overline{X}_n\|_\infty^2 > \varepsilon/(nd^2a))$. By Hoeffding's inequality, for all $i$: $\mathbb{P}(\overline{X}_{.,i} > b) \leq 2\exp(-nb^2)$. Now taking $b = \sqrt{\varepsilon/(nd^2a)}$ we obtain $\mathbb{P}(\overline{X}_{.,i}^2 > \varepsilon/(nd^2a)) \leq 2\exp(-\varepsilon/(d^2a))$. Finally, we get $\mathbb{P}(\|\overline{X}_n\|_\infty^2 > \varepsilon/(nda) \leq 2d\exp(-\varepsilon/(d^2a))$. This and (16) changes (12) into:

$$\mathbb{P}(n\overline{X}_n'\Gamma_n\overline{X}_n > \varepsilon) \leq \min_{a>0}\left(d(d+1)\exp\left(-\frac{na^2(d+2)^{-2}}{d^2(a+d+2)^2}\right) + 2d\exp\left(\frac{-\varepsilon}{d^2a}\right)\right).$$

Taking $a = ((d+2)^4\varepsilon/n)^{1/3}$, we get $\mathbb{P}(n\overline{X}_n'\Gamma_n\overline{X}_n > \varepsilon) \leq d(d+1)H_n(\varepsilon) + 2dR_n(\varepsilon)$. Combining this and (11), this concludes the proof. □

**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 $\partial M = \emptyset$. Then there exists a constant $A_d$ such that*

$$X^*_{k_n(i)} = (I_d + E_{i,n})\varphi_{X_i}(X_{k_n(i)}) - X_i \textit{ and } \max_i \|E_{i,n}\|_\infty \le A_d \sqrt{\frac{\ln(n)}{k_n}} \textit{ e.a.s.}$$

*Proof.* By Hoeffding's inequality we have that, for all $i$:

$$\mathbb{P}\big(\|r_{i,k_n}^{-2}\hat{S}_{i,k_n} - r_{i,k_n}^{-2}S_i\|_\infty \ge a\big) \le 2d^2 \exp(-2a^2 k_n),$$

where $S_i = \mathbb{E}(Y'Y \mid \|Y\| \le r_{i,k_n})$ with $Y = X - X_i$ and $\hat{S}_{i,k_n}$ as in Definition 1. Then

$$\mathbb{P}\big(\exists i : \|r_{i,k_n}^{-2}\hat{S}_{i,k_n} - r_{i,k_n}^{-2}S_i\|_\infty \ge a\big) \le n2d^2 \exp(-2a^2 k_n).$$

Now if we apply the Borel-Cantelli Lemma with $a = \sqrt{\frac{3\ln(n)}{2k_n}}$ we get that, with probability one, for $n$ large enough,

$$\|r_{i,k_n}^{-2}\hat{S}_{i,k_n} - r_{i,k_n}^{-2}S_i\|_\infty \le \sqrt{\frac{3\ln(n)}{2k_n}} \quad \text{for all } i = 1, \dots, n. \tag{17}$$

Let us denote by $P_i$ the matrix whose first $d'$ columns form an orthonormal base of $T_{X_i}M$, completed to obtain an orthonormal base of $\mathbb{R}^d$. By Lemma 1 $r_n \to 0$. For $n$ large enough, combining Proposition 1 points 3. and 4. and (6), there exists $c$ such that with probability one, for $n$ large enough,

$$\text{for all } i : \left\|r_{i,k_n}^{-2}S_i - \frac{1}{d'+2}P_i' J_{d'} P_i\right\|_\infty \le cr_n \quad \text{, where } J_{d'} = \begin{pmatrix} I_{d'} & 0 \\ 0 & 0 \end{pmatrix}. \tag{18}$$

Now, (17) and (18) give that, with probability one, for $n$ large enough and for all $i = 1, \dots, n$.

$$\left\|r_{i,k_n}^{-2}\hat{S}_{i,k_n} - \frac{1}{d'+2}P' J_{d'} P\right\|_\infty \le \sqrt{\frac{3\ln(n)}{2k_n}} + cr_n = \sqrt{\frac{3\ln(n)}{2k_n}}(1 + o(1)). \tag{19}$$

In what follows we consider $n$ large enough to ensure (19), and $\varepsilon_n = \sqrt{\frac{3\ln(n)}{2k_n}} + cr_n \le \frac{1}{4\sqrt{2d}(d'+2)}$. Since (19) holds for all $i$, we can remove the index $i$ in the matrices and vectors and assume that $i$ is fixed. For ease of writing (up to a change of base) we can assume that $P = I_d$, then

$$\left\|r_{k_n}^{-2}\hat{S}_{k_n} - \frac{1}{d'+2}J_{d'}\right\|_\infty \le \varepsilon_n.$$

It only remains to apply Proposition 2. □

## 5.3 Proof of Theorems 1 and 2

Theorems 1 and 2 follows from the following Lemma.

**Lemma 4.** *Let* $(k_n)$ *be a sequence which fulfills condition* K *and* $X_1, \dots, X_n$ *an i.i.d. sample drawn according to a distribution* $\mathbb{P}_X$ *which fulfills condition* P, *with* $\partial M = \emptyset$. *If* $r_n$ *is as in Definition 1, then for* $i = 1, \dots, n$, *we can built* $\delta^*_{i,k_n}$ *such that:*

*i.* $\delta_{i,k_n} = \delta^*_{i,k_n} + \varepsilon_{i,n}$,

*ii.* $\mathbb{P}(\delta^*_{i,k_n} \leq t | r_n < 1/k_n) = \Psi_n(t) \to 1 - F_{d'}(t)$,

*iii.* $\mathbb{P}(\delta^*_{i,k_n} > t | r_n < 1/k_n) \leq G_{k_n}(t)$,

*iv.* $\sqrt{\ln(n)} \max_i |\varepsilon_{i,k_n}| \xrightarrow{a.s.} 0$.

*Proof.* In what follows we consider $n$ large enough to have $1/k_n < r_M$.

For a given $i$ consider the sample $X^i_1, \dots, X^i_{k_n}$ with $X^i_j = X_{j(i)}$. Introduce $Y^i_j = \varphi_{X_i}(X^i_j)$ and

$$\delta^Y_{i,k_n} = \frac{k_n(d'+2)\|\overline{Y^i} - X_i\|^2}{r^2_{i,k_n}}.$$

First we are going to prove that $\delta^Y_{i,k_n} = \delta^*_{i,k_n} + e_{i,k_n}$, with $\delta^*_{i,k_n}$ satisfying points *ii.*, *iii.*, and *iv*, and with $\sqrt{\ln(n)} \max_i e_{i,k_n} \xrightarrow{a.s.} 0$.

Conditionally to $X_i$ and $r_{i,k_n}$ the sample $X^i_1, \dots X^i_{k_n}$ is drawn with the density $f^i(x) = \frac{f(x)}{\mathbb{P}_X(\mathcal{B}(X_i, r_{i,k_n}))} \mathbb{I}_{M \cap \mathcal{B}(X_i, r_{i,k_n})}$. So that the sample $Y^i_1, \dots Y^i_{k_n}$ is drawn with the density $g^i(x) = f^i(\varphi^{-1}_{X_i}(x)) \sqrt{\det(G_{X_i}(x))} \mathbb{I}_{B^i_n}$ (where $B^i_n = \varphi_{X_i}\big(M \cap \mathcal{B}(X_i, r_{i,k_n})\big)$).

By Proposition 1, for $n$ large enough,

$$f^i(x) \geq \frac{f(x)}{f(x)\sigma_{d'} r^{d'}_{i,k_n} \left( \frac{c_M r_{i,k_n}}{f_0 \sigma_{d'}} + 1 \right)}.$$

Again by Proposition 1, $\sqrt{\det(G_{X_i}(x))} > 1 - c_M r_{i,k_n}$. Observe that by Lemma 1 we can take $n$ large enough such that, for all $x \in B^i_n$:

$$g^i(x) \geq \frac{1 - c_M r^2_{i,k_n}}{\sigma_{d'} r^{d'}_{i,k_n} \left( \frac{c_M r_{i,k_n}}{f_0 \sigma_{d'}} + 1 \right)} \geq 0\,; \tag{20}$$

Notice that, by Proposition 1 we have:

$$\mathcal{B}\Big(X_i, r_{i,k_n}\big(1-c_M r_{i,k_n}\big)\Big) \cap (X_i + T_{X_i}M) \subset B_n^i \subset \mathcal{B}\big(X_i, r_{i,k_n}\big) \cap (X_i + T_{X_i}M). \quad (21)$$

Let us denote $B^-(X_i, r_{i,k_n}) = \mathcal{B}\big(X_i, r_{i,k_n}(1-c_M r_{i,k_n})\big) \cap (X_i + T_{X_i}M)$, and define $p_n = (1-c_M/k_n)^{d'+1}(\frac{c_M}{f_0 \sigma_{d'} k_n}+1)^{-1}$. Observe that $q_n = 1-p_n$ fulfills the conditions of Lemma 7. Equations (20), (21) and the assumptions on $r_n$ and $n$ allows us to claim that $\mathcal{Y}^i = \{Y_1^i, \dots Y_{k_n}^i\}$ has the same law as $\mathcal{Z}^i = \{Z_1, \dots Z_{k_n}\}$, where $Z_i$ is drawn as the mixture of a uniform law on $B^-(X_i, r_{i,k_n})$ with probability $p_n$ and a residual law of density $h_n^i$ with a probability $1-p_n$.

Let us denote by $K_n^i$ the number of points drawn with the uniform part of the mixture. Up to a re-indexing let us suppose that $Z_1, \dots, Z_{K_n^i}$ is the part of the sample drawn according to the uniform part of the mixture and that $Z_{K_n^i+1}, \dots, Z_{k_n}$ is the "residual" part of the sample.

Let us now draw a new artificial sample $Z'_{K_n^i+1}, \dots Z'_{k_n}$, i.i.d. and uniformly drawn in $B^-(X_i, r_{i,k_n})$. Let us define $Z_j^* = Z_j^i$ when $j \le K_n^i$ and $Z_j^* = Z'_j$ when $j > K_n^i$. Let us also define $e_j = Z_j - Z'_j$ for $j \in \{K_n^i + 1, \dots k_n\}$. We have:

$$\overline{Z^i} \stackrel{d}{=} \frac{1}{k_n}\sum_{j=1}^{k_n} Z_j^* + \frac{1}{k_n}\sum_{j=K_n^i+1}^{k_n} e_j.$$

Thus

$$\delta_{i,k_n}^Y \stackrel{d}{=} \frac{(d'+2)k_n}{r_{i,k_n}^2}\left\|\frac{1}{k_n}\sum_{j=1}^{k_n} Z_j^* - X_i + \frac{1}{k_n}\sum_{j=K_n^i+1}^{k_n} e_j\right\|^2.$$

Let us introduce:

$$\delta_{i,k_n}^* = (1-c_M r_{i,k_n})^2 \frac{(d'+2)k_n}{(r_{i,k_n} - c_M r_{i,k_n})^2}\left\|\frac{1}{k_n}\sum_{j=1}^{k_n} Z_j^* - X_i\right\|^2$$

and:

$$e_{i,k_n} = (\delta_{i,k_n}^Y - \delta_{i,k_n}^*).$$

First, the condition $r_n \leq 1/k_n$ gives that:

$$\left(1 - \frac{c_M}{k_n}\right)^2 \frac{(d'+2)k_n}{(r_{i,k_n} - c_M r_{i,k_n})^2} \left\| \frac{1}{k_n} \sum_{j=1}^{k_n} Z_j^* - X_i \right\|^2 \leq \delta_{i,k_n}^*$$
$$\leq \frac{(d'+2)k_n}{(r_{i,k_n} - c_M r_{i,k_n})^2} \left\| \frac{1}{k_n} \sum_{j=1}^{k_n} Z_j^* - X_i \right\|^2.$$

Therefore, applying Lemma 2 to $\frac{(d'+2)k_n}{(r_{i,k_n} - c_M r_{i,k_n})^2} \left\| \frac{1}{k_n} \sum_{j=1}^{k_n} Z_j^* - X_i \right\|^2$ it directly comes that $\delta_{i,k_n}^*$ fulfills conditions *ii.* and *iii.*

Let us now prove that $\max_i |e_{i,k_n}|$ fulfills *iv*. Denoting $E_{i,k_n} = \frac{1}{k_n} \sum_{K_n^i+1}^{k_n} e_j$, we have that $\|E_{i,k_n}\| \leq \frac{k_n - K_n^i}{k_n} r_{i,k_n}$. Then, applying the Cauchy-Schwartz inequality, we get

$$|e_{i,k_n}| = 2\frac{(d'+2)k_n}{r_{i,k_n}^2} \left\langle \frac{1}{k_n} \sum_{j=1}^{k_n} Z_j^* - X_i \ , \frac{1}{k_n} \sum_{j=K_n^i+1}^{k_n} e_j \right\rangle$$
$$+ \frac{(d'+2)k_n}{r_{i,k_n}^2} \|E_{i,k_n}\|^2$$
$$\leq 2\sqrt{d'+2} \sqrt{\delta_{i,k_n}^*} \frac{k_n - K_n^i}{\sqrt{k_n}} + 2(d'+2) \frac{(k_n - K_n^i)^2}{k_n},$$

where $K_n^i \rightsquigarrow Binom(k_n, p_n)$ and so $k_n - K_n^i \rightsquigarrow Binom(k_n, 1 - p_n)$. By direct application of Lemma 7 and Borel-Cantelli we obtain that $\ln(n) \max_i \left| \frac{k_n - K_n^i}{\sqrt{k_n}} \right| \overset{a.s.}{\to} 0$. Now, by Lemma 2 and Proposition 3 point *iii*, $\max_i \sqrt{\delta_{i,k_n}^*} \leq \sqrt{5 \ln(n)}$ e.a.s. Thus

$$\sqrt{\ln(n)} \max_i |e_{i,k_n}| \overset{a.s.}{\longrightarrow} 0. \tag{22}$$

Now, by Lemma 3 we have, for all $i$: $\delta_{i,k_n} = \delta_{i,k_n}^Y + e'_{i,k_n}$ with $|e'_{i,k_n}| \leq A_d \sqrt{\frac{\ln(n)}{k_n}} (2\sqrt{d} + d)\delta_{i,k_n}^Y$ e.a.s. Let us introduce $B_d = A_d(2\sqrt{d} + d)$. Then, with probability 1, for $n$ large enough,

$$\sqrt{\ln(n)} \max_i |e'_{i,k_n}| \leq B_d \sqrt{\frac{(\ln(n))^4}{k_n}} \frac{1}{\ln(n)} \max \delta_{i,k_n}^* + B_d \sqrt{\frac{\ln(n)}{k_n}} \sqrt{\ln(n)} \max |e_{i,k_n}|.$$

As (22) holds and $\ln(n)/k_n \to 0$ it only remains to prove that

$$B_d\sqrt{\frac{(\ln(n))^4}{k_n}}\frac{1}{\ln(n)}\max \delta^*_{i,k_n} \xrightarrow{a.s.} 0$$

to conclude the proof. This last point follows directly from Proposition 3 point *iii* and the condition $(\ln(n))^4/k_n \to 0$

□

We can now prove Theorem 1, which basically says that, under the assumptions of Lemma 4, $P\big(\Delta_{n,k_n} \geq t_n(\alpha)\big) \leq \alpha + o(1)$.

*Proof of Theorem 1.* Theorem 1 It is a direct consequence of Lemma 1 and 4. Indeed:

$$\mathbb{P}_{H_0}\big(\Delta_{n,k_n} \geq t_n(\alpha)\big) \leq \mathbb{P}_{H_0}\big(\Delta_{n,k_n} \geq t_n(\alpha)|r_n < 1/k_n\big) + \mathbb{P}_{H_0}(r_n > 1/k_n).$$

By Lemma 1 $\mathbb{P}_{H_0}(r_n > 1/k_n) \to 0$. On the other hand,

$$\begin{aligned}\mathbb{P}_{H_0}\big(\Delta_{n,k_n} \geq t_n(\alpha)|r_n < 1/k_n\big) &\leq \mathbb{P}_{H_0}\left(\max_i \delta^*_{i,k_n} + \max|\varepsilon_{i,n}| \geq t_n(\alpha)\middle| r_n < 1/k_n\right)\\ &= \mathbb{P}_{H_0}\left(\max_i \delta^*_{i,k_n} \geq t_n(\alpha) - 1/\sqrt{n}\middle| r_n < 1/k_n\right) +\\ &\qquad \mathbb{P}_{H_0}\left(\max|\varepsilon_{i,n}| \geq 1/\sqrt{n}\middle| r_n < 1/k_n\right)\\ &\leq \alpha + o(1).\end{aligned}$$

□

Now, we prove Theorem 2 which says that, under the assumptions of Lemma 4 we have $\hat{\Psi}_n(x) \overset{L^2}{\to} \Psi_{d'}(x)$.

*Proof of Theorem 2.* A direct consequence of Lemma 4 is that $\mathbb{E}(\hat{\Psi}_n(x)) \to \Psi_{d'}(x)$. Therefore, we only have to prove $\mathbb{V}(\hat{\Psi}_n(x)) \to 0$.

Let us consider a sequence $\varepsilon_n$ such that $\varepsilon_n \in [0,1]$ and $\varepsilon_n \to 0$. Let us denote $p_{x,n} = \mathbb{P}_X\big(\mathcal{B}(x,(2+\varepsilon_n)/k_n)\big)$. Since $f$ is Lipschitz, if we denote $K_f$ the constant, we get

$$\begin{aligned}p_{x,n} &\leq \sigma_{d'}\big((2+\varepsilon_n)/k_n\big)^{d'} f(x)\big(1+(2+\varepsilon_n)K_f/k_n\big)\\ &\leq \sigma_{d'}(3/k_n)^{d'} f(x)\big(1+3K_f/k_n\big).\end{aligned} \tag{23}$$

In the same way, $p_{x,n} \geq \sigma_{d'}(2/k_n)^{d'} f(x)\big(1-3K_f/k_n\big)$.

Let $N_{x,n}$ denote the number of observation belonging to $\mathcal{B}\big(x,(2+\varepsilon_n)/k_n\big)$. Applying Hoeffding's inequality we get, for all $\lambda_n > 1$:

$$\mathbb{P}\big(N_{x,n} \geq \lambda_n p_{n,x} n\big) = \mathbb{P}\left(\frac{N_{x,n}}{n} - p_{n,x} \geq (\lambda_n - 1)p_{n,x}\right) \leq \exp\Big(-2\big((\lambda_n-1)p_{n,x}\big)^2 n\Big).$$

Taking, $\lambda_n = \mu k_n^d \sqrt{\frac{\ln(n)}{n}}$ with $\mu > 0$,

$$\mathbb{P}\left(N_{x,n} \geq p_{n,x} k_n^d \sqrt{n\ln(n)}\right) \leq \exp\left(-\mu^2\sigma_{d'}^2 2^{2d'} f(x)^2 \ln(n)(1+o(1))\right),$$

so that:

$$\mathbb{P}\left(N_{x,n} \geq p_{n,x} k_n^d \sqrt{n\ln(n)}\right) \leq \exp\left(-\mu^2\sigma_{d'}^2 2^{2d'} f_0^2 \ln(n)(1+o(1))\right).$$

Now, by (23),

$$\begin{aligned}\mathbb{P}\left(N_{x,n} \geq \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}\right) &\leq \mathbb{P}\left(N_{x,n} \geq p_{n,x} k_n^d \sqrt{n\ln(n)}\right)\\ &\leq \exp\Big(-(\mu\sigma_{d'}2^{d'}f_0)^2\ln(n)(1+o(1))\Big).\end{aligned}$$

Let us cover $M$ with $x_1,\ldots,x_{\nu_n}$ (deterministic) balls of radius $\varepsilon_n/k_n$. Observe that we can take $\nu_n \leq \theta_M(k_n/\varepsilon_n)^d$. If we denote $\mathcal{X}_n = \{X_1,\ldots,X_n\}$, then,

$$\begin{aligned}&\mathbb{P}\left(\cup_{i=1}^n\Big\{\#\big(\mathcal{B}(X_i,2/k_n)\cap\mathcal{X}_n\big) \geq \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}\Big\}\right) \leq\\ &\mathbb{P}\left(\cup_{i=1}^{\nu_n}\Big\{\#\big(\mathcal{B}(x_i,(2-\varepsilon_n)/k_n)\cap\mathcal{X}_n\big) \geq \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}\Big\}\right) \leq\\ &\qquad\theta_M k_n^d \varepsilon_n^{-d} n^{-(\mu\sigma_{d'}2^{d'}f_0)^2(1+o(1))}.\end{aligned}$$

If we choose $\varepsilon_n = \min((\ln(n))^{-1/d},1)$ and $\mu > (\sigma_{d'}2^{d'}f_0)^{-1}$, the condition $(\ln(n))k_n^{1+d}/n \to 0$ implies that

$$\mathbb{P}\left(\cup_{i=1}^n\Big\{\#\big(\mathcal{B}(X_i,2/k_n)\cap\mathcal{X}_n\big) \geq \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}\Big\}\right) \to 0.$$

Now, let

$$\mathcal{A}_n = \cap_{i=1}^n\Big\{\#\big(\mathcal{B}(X_i,2/k_n)\cap\mathcal{X}_n\big) < \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}\Big\}\cap\{r_n < 1/k_n\}.$$

Observe that the random variables $\delta_{i,k_n}$ are not independent in general. However, if $\|X_i - X_j\| > 2r_n$, $\delta_{i,k_n}$ and $\delta_{j,k_n}$ are independent. Therefore

$$\begin{aligned}\mathbb{V}\left(\hat{\Psi}_n(x)\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\{j:\|X_i-X_j\|<2r_n\}} cov(\mathbb{I}_{\{\delta_i\geq x\}}, \mathbb{I}_{\{\delta_j\geq x\}}) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\{j:\|X_i-X_j\|<2/k_n\}} cov(\mathbb{I}_{\{\delta_i\geq x\}}, \mathbb{I}_{\{\delta_j\geq x\}})\end{aligned}$$

Thus, conditioned to $\mathcal{A}_n$, since $cov(\mathbb{I}_{\{\delta_i\geq x\}}, \mathbb{I}_{\{\delta_j\geq x\}}) \leq 1$ we get

$$\sum_{\{j:\|X_i-X_j\|<2/k_n\}} cov(\mathbb{I}_{\{\delta_i\geq x\}}, \mathbb{I}_{\{\delta_j\geq x\}}) \leq \mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)}.$$

Finally, conditioned to $\mathcal{A}_n$, the variance of $\mathbb{V}_{\mathcal{A}_n}\left(\hat{\Psi}_n(x)\right)$ fulfills

$$\mathbb{V}_{\mathcal{A}_n}\left(\hat{\Psi}_n(x)\right) \leq \frac{1}{n}\mu\sigma_{d'} f_1 3^{d'}(1+3K_f/k_n)\sqrt{n\ln(n)} \to 0.$$

As $\mathbb{P}(\mathcal{A}_n) \to 1$ and $\mathbb{P}(r_n < 1/k_n) \to 1$, we finally obtain $\mathbb{V}\left(\hat{\Psi}_n(x)\right) \to 0$ which concludes the proof. □

## 5.4 Proof of Theorems 3 and 4

Theorems 3 and 4 are direct consequences of the following lemma.

**Proposition 4.** *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.*

$$\mathbb{E}\left(\frac{\langle X - x, u\rangle}{r}\right) = \alpha_d, \tag{24}$$

*where* $\alpha_d = \left(\frac{\Gamma(\frac{d+2}{2})}{\sqrt{\pi}\Gamma(\frac{d+3}{2})}\right)$.

*Proof.* Let us first assume that $r = 1$, $x = 0$ and $u = e_1 = (1, 0, \ldots, 0)$. The marginal density of $X_1$ is

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

so

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

For a general value of $r$, $x$ and $u$ let us 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 $Y$ has uniform distribution on $\mathcal{B}_{e_1}(0,1)$ and so (24) holds. $\square$

**Lemma 5.** *Let $k_n$ be a sequence fulfilling condition* K. *Let us 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 *with $\partial M \neq \emptyset$. Then, there exists a sequence $\lambda_n \xrightarrow{a.s.} \alpha_{d'}^2$ such that: $\Delta_{n,k_n}/k_n \geq (d'+2)\lambda_n$, where $\alpha_{d'}$ was defined in Proposition 4.*

*Proof.* We will divide the proof into two steps. In the first one we are going to prove that there exists a constant $c_{\partial M}$ such that, with probability one, there exists $X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M}\ln(n)/n)$ for $n$ large enough. In the second step we are going to prove that, eventually almost surely, for all $X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M}\ln(n)/n)$ it holds that $\delta_{i_0,k_n}/k_n \geq (d'+2)\alpha_{d'}^2(1+o(1))$.

In order to prove the first step, observe that as $\partial M$ is $\mathcal{C}^2$, its inner packing number $\nu(\varepsilon)$ (the maximal number of balls, centered in $\partial M$, of radius $\varepsilon$ that are all pairwise disjoint) satisfies $\nu(\varepsilon) \geq B\varepsilon^{-d'+1}$ for some constant $B > 0$. Let us denote by $x_i$, for $i \in \{1,\dots,v(\varepsilon)\}$, the centers of these balls. Then $|\partial M \oplus \mathcal{B}(0,\varepsilon)|_{d'} \geq \sum_i |\mathcal{B}(x_i,\varepsilon)\cap M|_{d'}$. Now, as a direct consequence of Proposition 1 point 5, there exists $R$ and $C$ such that, for all $\varepsilon \leq R$: $|\partial M \oplus \mathcal{B}(0,\varepsilon)|_{d'} \geq B\varepsilon^{-d'+1}(\sigma_{d'}\varepsilon^{d'}/2 - C\varepsilon^{d'+1})$. That is:

$$\big|\partial M \oplus \mathcal{B}(0,\varepsilon)\big|_{d'} \geq B\sigma_{d'}\frac{\varepsilon}{2} - BC\varepsilon^2. \tag{25}$$

Thus, the probability that there is no sample point in $\partial M \oplus \mathcal{B}(0, \frac{3\ln(n)}{f_0 B\sigma_{d'} n})$ can be bounded as follows:

$$\mathbb{P}\left(\left(\partial M \oplus \frac{3\ln(n)}{f_0 B\sigma_{d'} n}\mathcal{B}(0,1)\right)\cap \mathcal{X}_n = \emptyset\right) \leq \left(1 - \frac{3\ln(n)}{2n}\left(1 - \frac{6C\ln(n)}{f_0 B\sigma_{d'} n}\right)\right)^n = n^{-3/2+o(1)}.$$

Finally, the first step follows as a direct application of the Borel-Cantelli Lemma, with $c_{\partial M} = 3/(B\sigma_{d'})$.

For an observation $X_{i_0}$ such that $d(X_{i_0}, \partial M) \leq c_{\partial M} \ln(n)/n$, let us denote by $x_0$ a point of $\partial M$ such that $\|X_{i_0} - x_0\| \leq c_{\partial M} \ln(n)/n$, and recall that $u_{x_0}$ denotes the unit vector tangent to $M$ and normal to $\partial M$ pointing outward $M$. Let us introduce $Y_{k(i_0)} = \varphi_{x_0}(X_{k(i_0)})$.

In what follows we will prove that for all $X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M} \ln(n)/n\big)$:

$$\frac{\frac{1}{k_n}\sum_{k=1}^{k_n} \langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle}{r_{i_0,k_n}} \xrightarrow{a.s.} \alpha_{d'}. \tag{26}$$

Let us define $\rho_{n,-} = r_{i_0,k_n} - c_{\partial M} \ln(n)/n$ and $\rho_{n,+} = r_{i_0,k_n} + c_{\partial M} \ln(n)/n$.

Observe that, according to Proposition 1, $\langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle \in [-c_M \rho_{n,+}^2, \rho_{n+}]$, so that applying Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{k_n\rho_{n,+}(1+c_M\rho_{n,+})}\sum_{k=1}^{k_n}\langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle - \frac{\mathbb{E}(\langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle)}{\rho_{n,+}(1+c_M\rho_{n,+})}\right| \geq t\right) \leq 2\exp(-2t^2 k_n). \tag{27}$$

Then, to prove (26) it only remains to prove that, for all $X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M} \ln(n)/n\big)$:

$$\text{(a)}\ \frac{\ln(n)}{n r_{i_0,k_n}} \xrightarrow{a.s.} 0, \quad \text{(b)}\ \frac{\mathbb{E}\langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle}{(\rho_{n,+} + c_M \rho_{n,+}^2)} \longrightarrow \alpha_{d'}.$$

Indeed:

i. From (b) and (27) we obtain

$$\frac{1}{k_n(\rho_{n,+} + c_M\rho_{n,+}^2)}\sum_{k=1}^{k_n}\langle Y_{k(i_0)} - x_0, -u_{x_0}\rangle \xrightarrow{a.s.} \alpha_{d'}, \tag{28}$$

from a direct application of the Borel-Cantelli Lemma, by noticing that $k_n/(\ln n)^4 \to \infty$ implies that $\sum_n \exp(-2t^2 \ln(k_n)) < +\infty$.

ii. From (28) and (a) we get (26).

First assume that $r_{i_0,k_n} \xrightarrow{a.s.} 0$ (the proof is is similar to the proof of Lemma 1, using a covering of $\partial M$ instead of $M$, and bounding the probability according to Proposition 1 point 5. instead instead of point 4.. Then, from now to the end of the proof, we suppose that $n$ is large enough to have $r_{i_0,k_n} \leq r_M$.

Let us now prove (a). First we cover $\partial M$ with $\nu_n \leq B'(n/\ln(n))^{d'-1}$ balls, centered at $x_i \in \partial M$ with a radius $c_{\partial M}\ln(n)/n$. Let us denote $R_n^- = (\ln(n) - 2c_{\partial M})\ln(n)/n$ and $R_n^+ = (\ln(n) + 2c_{\partial M})\ln(n)/n$. We have:

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M}\ln(n)/n), r_{i_0,k_n} \leq R_n^-\Big) \leq \sum_{i=1}^{\nu_n} \mathbb{P}\Big(\#\big\{\mathcal{B}\big(x_i, R_n^- + 2c_{\partial M}\ln(n)/n\big) \cap \mathcal{X}_n\big\} \geq k_n\Big). \quad (29)$$

Since $R_n^- = (\ln(n) - 2c_{\partial M})\ln(n)/n$, if we apply Proposition 1 point 5. and $f \leq f_1$ we can bound the right hand side of (29) by:

$$\mathbb{P}\Big(\#\big\{\mathcal{B}\big(x_i, R_n^- + 2c_{\partial M}\ln(n)/n\big) \cap \mathcal{X}_n\big\} \geq k_n\Big) \leq \sum_{j=k_n}^{n} \binom{n}{j} \left(\frac{f_1\sigma_{d'}(\ln(n))^{2d'}}{2n^{d'}}(1+o(1))\right)^j.$$

Now from the bound $n!/(n-j)! \leq n^j$, we get

$$\mathbb{P}\Big(\#\big\{\mathcal{B}\big(x_i, R_n^- + 2c_{\partial M}\ln(n)/n\big) \cap \mathcal{X}_n\big\} \geq k_n\Big) \leq \sum_{j=k_n}^{n} \frac{1}{j!} \left(\frac{f_1\sigma_{d'}(\ln(n))^{2d'}}{2n^{d'-1}}(1+o(1))\right)^j. \quad (30)$$

Finally, using $\sum_{j=k}^{n} x^j/j! \leq x^k e^x/k!$ for $x \geq 0$ to bound the right hand side of (30) we obtain:

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M}\ln(n)/n), r_{i_0,k_n} \leq R_n^-\Big) \leq B'\left(\frac{n}{\ln n}\right)^{d'-1} \frac{\left(\frac{f_1\sigma_{d'}(\ln(n))^{2d'}}{2n^{d'-1}}(1+o(1))\right)^{k_n}}{k_n!} \exp\left(\frac{f_1\sigma_{d'}(\ln(n))^{2d'}}{2n^{d'-1}}(1+o(1))\right). \quad (31)$$

Now we will consider two cases: $d' = 1$ and $d' > 1$. For the first one ($d' = 1$), using Stirling's formula we can bound the right hand side of (31) from above by

$$\frac{B'}{\sqrt{2\pi k_n}} \exp\left(-k_n \ln\left(\frac{k_n}{e}\right) + k_n \ln\left(\frac{f_1\sigma_{d'}(\ln(n))^2(1+o(1))}{2}\right) + (\ln(n))^2 \frac{f_1\sigma_{d'}(1+o(1))}{2}\right)(1+o(1))$$

Then, the condition $k_n \gg (\ln(n))^4$ ensures that

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M}\ln(n)/n\big), r_{i_0,k_n} \leq R_n^-\Big) \leq \frac{1}{\sqrt{2\pi k_n}} \exp\left(-k_n \ln\left(\frac{k_n}{e}\right)(1+o(1))\right).$$

Second, if $d' > 1$ then from (31) we directly obtain

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M} \ln(n)/n\big), r_{i_0,k_n} \leq R_n^-\Big) = o((k_n!)^{-1}).$$

In both cases $k_n \gg (\ln(n))^4$ ensures that :

$$\sum_n \mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M} \ln(n)/n\big), r_{i_0,k_n} \leq R_n^-\Big) < +\infty.$$

The proof of (a) follows by a direct application of the Borel-Cantelli Lemma.

Let us now prove (b).

Let us denote by $g_{r_{i_0,k_n}}$ the density of $Y = \varphi_{x_0}(X)$ conditioned by $r_{i_0,k_n}$ and $\|X - X_{i_0}\| \leq r_{i_0,k_n}$. Let us introduce the set $B_0 = \varphi_{x_0}\big(\mathcal{B}(X_{i,0}, r_{i_0,k_n}) \cap M\big)$. Reasoning as we did at the beginning of Lemma 4, the Lipschitz continuity of $f$, Proposition 1 part 3. and Lemma 5 ensure that there exists a sequence $\varepsilon_n = O(r_{i_0,k_n})$ such that, for all $x \in B_0$:

$$\left| g_{r_{i_0,k_n}}(x) \frac{\sigma_{d'} r_{i_0,k_n}^{d'}}{2} - 1 \right| \leq \varepsilon_n.$$

Thus,

$$\left| \frac{\sigma_{d'} r_{i_0,k_n}^{d'}}{2} \mathbb{E}\big(\langle Y - x_0, u_{x_0}\rangle | r_{i_0,k_n}\big) - \int_{B_0} \langle x - x_0, u_{x_0}\rangle dx \right| \leq$$
$$\varepsilon_n \int_{B_0} \|x\| dx \leq \varepsilon_n \int_{\mathcal{B}(x_0,\rho_{n,+})} \|x\| dx \leq \varepsilon_n \frac{\sigma_{d'-1}}{d'+1} \rho_{n,+}^{d'+1}. \quad (32)$$

Observe that $(\mathcal{B}(X_{i,0}, \rho_{n,-}) \cap M) \subset (\mathcal{B}(X_{i,0}, r_{i_0,k_n}) \cap M) \subset (\mathcal{B}(X_{i,0}, \rho_{n,+}) \cap M)$. Therefore, by Lemma 5, we get,

$$\mathcal{B}(x_0, \rho_{n-}) \cap \big\{y : \langle y - x_0, u_{x_0}\rangle \geq c_M \rho_{n,+}^2\big\} \subset B_0$$
$$\subset \mathcal{B}(x_0, \rho_{n,+}) \cap \big\{y, \langle y - x_0, u_{x_0}\rangle \geq -c_M \rho_{n,+}^2\big\} \quad (33)$$

From (33) we obtain (using a very rough upper bound) that:

$$\big|B_0 \Delta \mathcal{B}_{u_{x_0}}(x_0, r_{i_0})\big| \leq \sigma_{d'}(\rho_{n,+}^{d'} - \rho_{n,-}^{d'}) + 2c_M \sigma_{d'-1} \rho_{n,+}^{d'+1}.$$

Thus:

$$\left| \int_{B_0} \langle x - x_0, u_{x_0} \rangle dx - \int_{\mathcal{B}_{u_{x_0}}(x_0, r_{i_0})} \langle x - x_0, u_{x_0} \rangle dx \right| \leq \sigma_{d'} (\rho_{n,+}^{d'+1} - \rho_{n,-}^{d'+1}) + 2 c_{\partial M} \sigma_{d'-1} \rho_{n,+}^{d'+2}. \tag{34}$$

Proposition 4 shows that $\int_{\mathcal{B}_{u_{x_0}}(x_0,r_{i_0})} \langle x - x_0, u_{x_0} \rangle dx = \alpha_{d'} r_{i_0}$. Thus (32) and (34) provides the existence of $C$ and $C'$ such that

$$\left| \mathbb{E} \left( \frac{\langle Y - x_0, u_{x_0} \rangle}{r_{i_0,k_n}} \middle| r_{i_0,k_n} \right) - \alpha_{d'} \right| \leq 2 \frac{\rho_{n,+}^{d'+1} - \rho_{n,-}^{d'+1}}{r_{i_0,k_n}^{d'+1}} + (C \rho_{n,+} + C' \varepsilon_n) \frac{\rho_{n,+}^{d'+1}}{r_{i_0,k_n}^{d'+1}}.$$

Therefore (a) gives:

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

Applying (a) again $\frac{\mathbb{E}\langle Y - x_0, u_{x_0} \rangle}{(\rho_{n,+} + c'_{M,4} \rho_{n,+}^2)} \to \alpha_{d'}$ , we get (b). As a consequence (26) is now proved.

Now, in order to finish the proof of the Lemma, notice that, reasoning similarly to what has been done in Lemma 3 and using $(a)$ and $(b)$ it can be proved that $X_{k(i)}^* = (I_d + F_{n,i_0})(Y_{k(i)} - x_0 + x_0 - X_{i_0})$ with $\|F_{n,i_0}\|_\infty \xrightarrow{a.s.} 0$. Then

$$\frac{\| \sum_{k=1}^{k_n} X_{k(i_0)}^* \|}{k_n r_{i_0,k_n}} \geq (1 - \|F_{n,i_0}\|_\infty) \frac{\frac{1}{k_n} \sum_{k=1}^{k_n} \langle Y_{k(i_0)} - x_0, u_{x_0} \rangle}{r_{i_0,k_n}} - (1 + \|F_{n,i_0}\|_\infty) \frac{c_{\partial M} \ln(n)}{n r_{i_0,k_n}}. \tag{35}$$

Thus, there exists a sequence $\lambda_n \xrightarrow{a.s.} \alpha_{d'}^2$ such that $\frac{\delta_{i_0,k_n}}{(d'+2)k_n} \geq \lambda_n$, which concludes the proof. □

*Proof.* Proof of Theorems 3 and 4

To prove Theorem 3 observe that $k_n \gg (\ln(n))^4$ ensure the existence of $n_1$ such that for all $n \geq n_1$, $\frac{k_n}{2}(d'+2)\alpha_{d'}^2 \geq t_n(\alpha)$. The proof follows from equation (35).

Regarding Theorem 4, if $t_n \leq \mu k_n$ with $\mu < (d'+2)\alpha_{d'}^2$ then, reasoning exactly as previously, $\mathbb{P}_{H_1}(\Delta_{n,k_n} \geq t_n) = 1$ for $n$ large enough. On the other hand if $t_n \geq \lambda \ln(n)$ for some $\lambda > 4$ then Lemma 4, Proposition 3 and Borel-Cantelli's Lemma ensure that $\mathbb{P}_{H_0}(\Delta_{n,k_n} < t_n) = 1$ for $n$ large enough. □

# 6 Appendix

Proofs of preliminary results

### 6.1 Proof of Proposition 1

*Proof.* 1. Proceeding by contradiction, let $r_n \to 0$, $x_n$, $y_n$ and $z_n$ 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) $x_n \to x \in M$. Let us denote $w_n \doteq \frac{y_n - z_n}{\|y_n - z_n\|} \to w$. Since $\varphi_{x_n}(y_n) = \varphi_{x_n}(z_n)$ we have $w_n \in (T_{x_n}M)^\perp$. As $M$ is of class $\mathcal{C}^2$, we have $w \in (T_x M)^\perp$. Let $\gamma_n$ be a geodesic curve on $M$ that joins $y_n$ to $z_n$ (there exists at least one since $M$ is compact). As $M$ is compact and $\mathcal{C}^2$ it has an injectivity radius $r_{inj} > 0$. Therefore (see Proposition 88 in Berger (2003)), if we take $n$ large enough that $r_n \le r_{inj}/2$, we may take $\gamma_n$ 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 $w_n = \frac{v_n}{\|y_n - z_n\|} + o(1)$. 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 $r_0$ such that, for all $x \in M$ $\varphi_x$ is one to one from $M \cap \mathcal{B}(x, r)$ to $\varphi_x\big(M \cap \mathcal{B}(x, r)\big)$ (then the existence of $r_1$ such that for all $x \in M$ and $r \le r_1$ $\varphi_x$ is one to one and $\mathcal{C}^2$ is easily to obtained)

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

$$\begin{aligned} \varphi_x^{-1} : \quad & \varphi_x\big(M \cap \mathcal{B}(x, r_1)\big) \to M \cap \mathcal{B}(x, r_1) \\ & x + \begin{pmatrix} y_1 \\ \vdots \\ y_{d'} \\ 0_{d-d'} \end{pmatrix} \mapsto x + \begin{pmatrix} y \\ \Phi_{x,d'+1}(y) \\ \vdots \\ \Phi_{x,d}(y) \end{pmatrix} \end{aligned} \tag{36}$$

The $\mathcal{C}^2$ regularity and compactness of $M$ allow 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 $\varphi_x$ is the orthogonal projection we have, for all $x$ and $k$: $\nabla \Phi_k(0)$. Once again smoothness and compactness assumptions guarantee that the Hessian matrices $H(\Phi_{x,k})(0)$ has there eigen values uniformly bounded by a $\lambda_M$.

Thus, first

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

and then, there exists $c_3$ and $r_3$ such that, for all $(x, y) \in M$ such that $\|x-y\| \leq r_3$,

$$\|\varphi_x(y) - y\| \leq c_3\|x - y\|^2. \tag{37}$$

Second :

$$J_x(y) = \begin{pmatrix} I_{d'} \\ \vec{\nabla}\Phi_{x,d'+1}(y) \\ \vdots \\ \vec{\nabla}\Phi_{x,d}(y) \end{pmatrix} = \begin{pmatrix} I_{d'} \\ O(\|y\|) \\ \vdots \\ O(\|y\|) \end{pmatrix} \text{ and } G_x(y) = W_x(y)'W_x(y) = I_{d'} + O(||y||).$$

This, together with the differentiability of the determinant entails that there exists $c_4$ and $r_4$ such that for all $(x, y) \in M$ such that $\|x - y\| \leq r_4$,

$$|G_x(y) - 1| \leq c_4\|x - y\|. \tag{38}$$

4. First notice that only the first inclusion has to be proved, the second one is obvious. Let us introduce $\tilde{r} = \min\{r_1, r_2, r_3, 1/c_3\}$. Proceeding by contradiction, suppose that there exists $r, x$ and $y$ such that: $0 < r \leq \tilde{r}$, $x \in M$, $d(x, \partial M) > r$, $y \in \mathcal{B}\big(x, r(1 - c_3 r)\big) \cap T_x M$ and $y \notin \varphi_x\big(\mathcal{B}(x, r) \cap M\big)$. As $x \in \varphi_x(\mathcal{B}(x, r) \cap M)$ the line segment $[x, y]$ intersects $\partial(\varphi_x\big(\mathcal{B}(x, r) \cap M)\big)$. Let $z \in [x, y] \cap \partial\varphi_x\big(\mathcal{B}(x, r) \cap M\big)$. On one hand we clearly have $\|x - z\| < \|x - y\| \leq r(1 - c_3 r)$. On the other hand, since $\varphi_x^{-1}$ is a continuous function, $\partial\varphi_x\big(\mathcal{B}(x, r) \cap M\big) = \varphi_x\big(\partial(\mathcal{B}(x, r) \cap M\big))$, and, because $d(x, \partial M) > r$ it comes that $\partial\varphi_x\big(\mathcal{B}(x, r) \cap M\big) = \varphi_x(M \cap \partial\mathcal{B}(x, r)))$ then, there exists $z_0$, $||x - z_0|| = r$, $\varphi_x(z_0) = z$. Then by (37)

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

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

$$\mathcal{B}(x, r - c_5 r^2) \cap (x + T_x M) \subset \varphi_x(\mathcal{B}(x, r) \cap M) \subset \mathcal{B}(x, r) \cap (x + T_x M). \tag{39}$$

5. Sketch of proof. Suppose that $\partial M \neq \emptyset$, for all $x \in \partial M$ introduce $\varphi_x^*$ the affine projection on $x + T_x\partial M$. First notice that, for all $y \in \partial M$ we have $\varphi_x^*(y) = \varphi_x(y) - \langle y - x, u_x\rangle u_x$ thus $|\langle y - x, u_x\rangle| \leq \|\varphi_x^*(y) - y\| + \|\varphi_x(y) - y\|$. Recall that $\partial M$ is of class $\mathcal{C}^2$ so that, by application of (39) (on $M$ and $\partial M$) we have there exists $r_6$ 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:

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

and

$$\varphi_x(\partial M \cap \mathcal{B}(x,r)) \subset \mathcal{B}(x,r) \cap (x+T_xM) \cap \big\{y : |\langle y-x, u_x\rangle| \leq c_6\|x-y\|^2\big\} \quad (40)$$

Let us introduce $A^- = \mathcal{B}(x,r) \cap T_xM \cap \{y : \langle y-x, u_x\rangle \leq -2c_6\|x-y\|^2\}$. Notice that $A^-$ is convex. By definition of $u_x$ there exists a path $\gamma$ in $M$ that links $x$ to $x \in \mathring{M}$ with $\gamma'(0) = -u_x$ and $\gamma \cap \partial M = \{x\}$ that quickly implies that, for all $\varepsilon > 0$ exists $x_\varepsilon \in A^- \cap \varphi_x(\mathcal{B}(x,r) \cap M)$ and $\|x - x_\varepsilon\| \leq \varepsilon$.

Suppose now that, as previously there exists $0 < r < \min(r_3, 1/c_3)$, $x \in \partial M$ and $y \in \mathcal{B}\big(x, r(1-2c_3r)\big) \cap A^-$ such that $y \notin \varphi_x\big(\mathcal{B}(x,r) \cap M\big)$. Fix now $\varepsilon = c_3r^2$. As previously the line segment $[x_\varepsilon, y]$ intersects $\partial\varphi_x\big(\mathcal{B}(x,r) \cap M\big)$ at a point $z \in A^-$. Clearly we have $\|x - z\| \leq \varepsilon + \|x-y\| < r(1-c_Mr)$. Again $z = \varphi_x(z_0)$ with $z_0 \in \partial\big(M \cap \mathcal{B}(x,r)\big) = \big(M \cap \partial\mathcal{B}(x,r)\big) \cup \big(\partial M \cap \mathcal{B}(x,r)\big)$. As $\partial\big(M \cap \mathcal{B}(x,r)\big) = \big(M \cap \partial\mathcal{B}(x,r)\big) \cup \big(\partial M \cap \mathcal{B}(x,r)\big)$, $\varphi_x(z_0) \in A^-$ and (40) we necessary have $z_0 \in \partial\mathcal{B}(x,r)$, so $\|x - z_0\| = r$. Finally we have $r \leq r_M$, $\|x - z_0\| = r$ and $\|x - \varphi_x(z_0)\| < r(1-c_Mr)$. By point 3.

$$r^2 = \|x-z\|^2 + \|z-z_0\|^2 < (r(1-c_3r))^2 + c_3^2r^4 \leq r^2,$$

that is a contradiction. Then we proved that there exists $c_7$ and $r_7$ such that, for all $x \in \partial M$, for all $r \leq r_7$ we have

$$\mathcal{B}(x, r(1-c_7)) \cap (x + T_xM) \cap \{y : \langle y-x, u_x\rangle \leq -c_7r^2\} \subset \varphi_x(\mathcal{B}(x,r)).$$

The proof of,

$$\varphi_x(\mathcal{B}(x,r)) \subset \mathcal{B}(x,r) \cap (x+T_xM) \cap \{y : \langle y-x, u_x\rangle \leq c_7r^2\},$$

is easier and it is left to the reader.

□

## 6.2 Proof of Corollary 1

*Proof.* For any $r \leq r_M$ and any $x$

$$\mathbb{P}_X(\mathcal{B}(x,r)) \geq f_1 \int_{\varphi_x(\mathcal{B}(x,r)\cap M)} \sqrt{\det G_x(y)}dy$$

Thus by Proposition 1 point 2 we have:

$$\mathbb{P}_X(\mathcal{B}(x,r)) \leq f_1\sigma_{d'}r^{d'}(1+c_Mr) \quad (41)$$

For any $r$ consider first the points $x$ such that $d(x, \partial M) \geq r/2$, we have:

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

Now, since $r \leq 2r_M$ applying Property 1 point 2 and 4 we obtain:

$$\mathbb{P}_X(\mathcal{B}(x,r)) \geq f_0 \sigma_{d'}(r - c_M r^2)^{d'}(1 - c_M r) \tag{42}$$

Now if we consider points $x$ such that $d(x, \partial M) \leq r/2$, let $x^*$ be the projection of $x$ on $\partial M$ we have

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

since $r \leq 2r_M$ applying Property 1 point 2 and 5 we obtain:

$$\mathbb{P}_X(\mathcal{B}(x,r)) \geq f_0 \left( \frac{\sigma_{d'}}{2}(r)^{d'} - c_M \sigma_{d'-1} r^{d'+1} \right)(1 - c_M r) \tag{43}$$

Point 1 is a direct consequence of (41),(42) and (43).

To prove point 2 consider $r \leq r_M$.

$$\mathbb{P}_X(\mathcal{B}(x,r)) = \int_{\mathcal{B}(x,r)\cap M} f(y)\omega(y).$$

Applying first the Lipschitz hypothesis on $f$ we get,

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

Now by formula (6):

$$\int_{\mathcal{B}(x,r)\cap M} \omega(y) = \int_{\varphi_x(\mathcal{B}(x,r)\cap M)} \sqrt{\det G_x(y)}dy.$$

Applying Proposition 1 point 2:

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

Finally applying Proposition 1 point 4:

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

This implies:

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

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 $R_1$. □

**Lemma 6.** *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*, where $M$ is a manifold without boundary. Let $k_n$ be a sequence of positive integers such that $k_n \to +\infty$ and $(\ln(n))k_n^{1+d}/n \to 0$. Then, $k_n r_n \overset{a.s.}{\to} 0$, where $r_n$ was introduced in Definition 1.*

*Proof.* Let $\varepsilon_n \to 0$ be a sequence of positive real numbers. Let us first cover $M$ with $\nu_n \leq A_M \varepsilon_n^{-d} k_n^d$ balls of radius $\varepsilon_n/k_n$ centered in some $x_i \in M$. If we denote $\mathcal{X}_n = X_1, \ldots, X_n$, we have that

$$\mathbb{P}(r_n \geq a/k_n) \leq \mathbb{P}\Big(\exists i = 1, \ldots, \nu_n : \#\big\{\mathcal{B}(x_i, (a-\varepsilon_n)/k_n) \cap \mathcal{X}_n\big\} < k_n\Big).$$

If we use Corollary 1 and $\binom{j}{n} p^j (1-p)^{n-j} \leq \binom{j}{n}(1-p)^{n-j}$, we get

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq A_M \varepsilon_n^{-d} k_n^d \sum_{j=0}^{k_n} \binom{j}{n} \left(1 - \frac{f_0 \sigma_d (a-\varepsilon_n)^d}{k_n^d}(1+o(1))\right)^{n-j}.$$

Now, if we take take $n$ large enough so that $k_n/n < 0.5$ we get $\binom{j}{n} \leq \binom{k_n}{n}$, and then

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq A_M \varepsilon_n^{-d} k_n^{1+d} \binom{k_n}{n} \left(1 - \frac{f_0 \sigma_d (a-\varepsilon_n)^d}{k_n^d}(1+o(1))\right)^{n-k_n}. \tag{44}$$

Applying Stirling's formula to the right hand side of (44), we get

$$\frac{A_M \varepsilon_n^{-d}}{\sqrt{2\pi}} k_n^{1+d} \left(1 - \frac{k_n}{n}\right)^{-n+k_n} \left(\frac{n}{k_n}\right)^{k_n} \left(1 - \frac{f_0 \sigma_{d'} (a-\varepsilon_n)^d}{k_n^d}(1+o(1))\right)^{n-k_n}.$$

With the usual Taylor expansions,

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq \frac{A_M \varepsilon_n^{-d}}{\sqrt{2\pi}} \left(\frac{n}{k_n}\right)^{k_n} k_n^{1+d} \exp\left(k_n - \frac{n f_0 \sigma_d a^d (1+o(1))}{k_n^d}\right)(1+o(1)).$$

Since $k_n^{1+d}/n \to 0$, for $n$ large enough,

$$k_n - \frac{nf_0\sigma_d a^d(1+o(1))}{k_n^d} = -\frac{n}{k_n^d}\left(f_0\sigma_d(1+o(1)) - \frac{k_n^{d+1}}{n}\right) \leq -\frac{n}{2k_n^d} f_0\sigma_d a^d,$$

So, for $n$ large enough

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq \sqrt{2}\frac{A_M\varepsilon_n^{-d}}{\sqrt{\pi}}\left(\frac{n}{k_n}\right)^{k_n} k_n^{1+d}\exp\Big(-\frac{n}{2k_n^d}f_0\sigma_d a^d\Big).$$

Therefore

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq \sqrt{2}\frac{A_M\varepsilon_n^{-d}}{\sqrt{\pi}}\exp\left(-\frac{nf_0\sigma_d a^d}{2k_n^d} + k_n\ln(n) - k_n\ln(k_n) + (1+d)\ln(k_n)\right),$$

and then

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq \sqrt{2}\frac{A_M\varepsilon_n^{-d}}{\sqrt{\pi}}\exp\left(-\frac{nf_0\sigma_d a^d}{2k_n^d} + k_n\ln(n)(1+o(1))\right).$$

As $\ln(n)k_n^{1+d}/n \to 0$ we have:

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq \sqrt{2}\frac{A_M\varepsilon_n^{-d}}{\sqrt{\pi}}\exp\left(-\frac{nf_0\sigma_d a^d}{2k_n^d}(1+o(1))\right).$$

Applying again that $(\ln(n))k_n^{1+d}/n \to 0$ we get

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \ll \sqrt{2}\frac{A_M\varepsilon_n^{-d}}{\sqrt{\pi}}\exp\Big(-\frac{f_0\sigma_{d'}a^d}{2}k_n^d\ln(n)\Big)$$

If we choose $\varepsilon_n = 1/n$ then since $k_n \to +\infty$, the Lemma follows as a direct consequence of the Borel-Cantelli Lemma. □

**Lemma 7.** *Let $T_n \rightsquigarrow Binom(k'_n, q_n)$ with $q_n\sqrt{k'_n}\ln(n) \to 0$ and $k'_n/(\ln(n))^4 \to +\infty$. Then, for all $\lambda > 0$,*

$$\sum_n n\mathbb{P}\left(\ln(n)T_n/\sqrt{k'_n} > \lambda\right) < +\infty.$$

*Proof.* Let us bound $\mathbb{P}(T_n \geq \lfloor\lambda\sqrt{k'_n}/\ln(n)\rfloor)$. If we denote $j(\lambda,n) = \lfloor\lambda\sqrt{k'_n}/\ln n\rfloor$ then,

$$\mathbb{P}(T_n \geq j(\lambda,n)) = \sum_{j=j(\lambda,n)}^{k'_n}\binom{k'_n}{j}q_n^j(1-q_n)^{n-j}.$$

Notice that when $j \geq q_n(k'_n+1)-1$ and $j' > j$ we have:

$$\binom{k'_n}{j} q_n^j (1-q_n)^{n-j} > \binom{k'_n}{j'} q_n^{j'} (1-q_n)^{n-j'}.$$

Since $q_n \sqrt{k'_n} \ln(n) \to 0$, for $n$ large enough,

$$\mathbb{P}(T_n \geq j(\lambda,n)) \leq (k'_n - j(\lambda,n)) \binom{k'_n}{j(\lambda,n)} q_n^{j(\lambda,n)} (1-q_n)^{k'_n - j(\lambda,n)}.$$

Applying Stirling's formula,

$$\begin{aligned}\binom{k'_n}{j(\lambda,n)} &\sim \frac{1}{\sqrt{2\pi j(\lambda,n)}} \frac{{k'_n}^{k'_n+1/2}}{(k'_n-j)^{k'_n-j(\lambda,n)+1/2} j(\lambda,n)^{j(\lambda,n)}} \\ &\sim \frac{1}{\sqrt{2\pi j(\lambda,n)}} \frac{{k'_n}^{k'_n}}{(k'_n-j(\lambda,n))^{k'_n-j(\lambda,n)} j(\lambda,n)^{j(\lambda,n)}}.\end{aligned}$$

Now if we bound $(1-q_n)^{k'_n-j(\lambda,n)} \leq 1$ we get that, for $n$ large enough, $\mathbb{P}(T_n \geq j(\lambda,n))$ is bounded from above by,

$$\begin{aligned}&\frac{k'_n - j(\lambda,n)}{\sqrt{2\pi j(\lambda,n)}} \left(\frac{q_n k'_n}{j(\lambda,n)}\right)^{j(\lambda,n)} \left(1 - \frac{j(\lambda,n)}{k'_n}\right)^{-(k'_n - j(\lambda,n))} \\ &\quad = \frac{k'_n - j(\lambda,n)}{\sqrt{2\pi j(\lambda,n)}} \left(\frac{q_n k'_n}{j(\lambda,n)}\right)^{j(\lambda,n)} \exp\left(-\Big(k'_n - j(\lambda,n)\Big) \ln\left(1 - \frac{j(\lambda,n)}{k'_n}\right)\right)(1+o(1)).\end{aligned}$$

Since $j(\lambda,n)/k'_n \to 0$ and $j(\lambda,n)^2/k'_n \to 0$, we get,

$$\mathbb{P}(T_n \geq j(\lambda,n)) \leq \frac{k'_n - j(\lambda,n)}{\sqrt{2\pi j(\lambda,n)}} \left(\frac{q_n k'_n}{j(\lambda,n)}\right)^{j(\lambda,n)} \exp(j + o(j))(1+o(1)).$$

With $j(\lambda,n) = \lfloor \lambda \sqrt{k'_n}/\ln(n) \rfloor$, $n\mathbb{P}(T_n \geq j(\lambda,n))$ is bounded from above by,

$$\begin{aligned}&\frac{n(\ln(n))^{1/2}(k'_n)^{3/4}}{\sqrt{2\lambda\pi}} \left(\frac{q_n \sqrt{k'_n}\ln(n)}{\lambda}\right)^{\lambda\sqrt{k'_n}/\ln(n)} \exp\left(\frac{\lambda\sqrt{k'_n}}{\ln(n)}(1+o(1))\right)(1+o(1)) \\ &= \frac{n(\ln(n))^{1/2}(k'_n)^{3/4}}{\sqrt{2\lambda\pi}} \exp\left(\frac{\lambda\sqrt{k'_n}}{\ln(n)}\left(1 + \ln\left(\frac{q_n\sqrt{k'_n}\ln(n)}{\lambda}\right) + o(1)\right)\right)(1+o(1)).\end{aligned}$$

Since $q_n\sqrt{k'_n}\ln(n) \to 0$, we can take $n$ large enough such that

$$1 + \ln\left(\frac{q_n\sqrt{k'_n}\ln(n)}{\lambda}\right) + o(1) \le -1.$$

Then, if we bound $1 + o(1) \le 2$,

$$\begin{aligned} n\mathbb{P}(T_n \ge j(\lambda, n)) \le& \frac{\sqrt{2}n(\ln(n))^{1/2}(k'_n)^{3/4}}{\sqrt{\lambda\pi}} \exp\left(-\frac{\lambda\sqrt{k'_n}}{\ln(n)}\right) \\ =& \sqrt{\frac{2}{\lambda\pi}} \exp\left(-\frac{\lambda\sqrt{k'_n}}{\ln(n)} + \frac{3}{4}\ln(k'_n) + \ln(n) + \frac{1}{2}\ln(\ln(n))\right). \end{aligned}$$

Since $k'_n/\ln(n)^4 \to +\infty$

$$-\frac{\lambda\sqrt{k'_n}}{\ln(n)} + \frac{3}{4}\ln(k'_n) + \ln(n) + \frac{1}{2}\ln(\ln(n)) = -A_n\ln(n), \text{ with } A_n \to +\infty,$$

and then $\sum_n n\mathbb{P}(T_n \ge j(\lambda, n)) < +\infty$.

□

## 6.3 Proof of Proposition 2

*Proof.* Let us define,

$$\varepsilon_0 = \min\left\{\frac{\lambda_0}{3\sqrt{2d^3}}, \frac{\lambda_0}{2\sqrt{2d^2}}, \frac{\lambda_0\sqrt{\sqrt{16d^4+1}-1}}{8d^{7/2}}\right\}.$$

Let $u$ be an eigenvector of $S$ with $\|u\|_2 = 1$, associated to an eigenvalue $\mu$. As $Su = \mu u = Du + (S-D)u$ we have : $\|\mu u - Du\|_\infty \le d\varepsilon\|u\|_\infty$, denoting $u = (v, w) \in \mathbb{R}^{d'} \times \mathbb{R}^{d-d'}$ we have:

$$\max\left\{\min_i(|\mu - \lambda_i|)\|v\|_\infty, |\mu|\|w\|_\infty\right\} \le d\varepsilon \max\left\{\|v\|_\infty, \|w\|_\infty\right\}.$$

Since $\|\cdot\|_\infty \le \|\cdot\|_2 \le \sqrt{d}\|\cdot\|_\infty$ and $\|u\|_2 = 1$ we get,

$$\max\left\{\min_i(|\mu - \lambda_i|)\|v\|_2, |\mu|\|w\|_2\right\} \le d^{3/2}\varepsilon. \tag{45}$$

Suppose that $\|v\|_2 \geq \|w\|_2$ then $\|v\|_2 \geq 1/\sqrt{2}$. Then (45) implies, $\min_i(|\mu - \lambda_i|) \leq \sqrt{2d^3}\varepsilon$ and $\|w\|_2 \leq \frac{d^{3/2}\varepsilon}{\lambda_0 - \sqrt{2d^3}\varepsilon} \leq \frac{3d^{3/2}\varepsilon}{2\lambda_0}$ (the last inequality is a consequence of $\varepsilon \leq \varepsilon_0 \leq \frac{\lambda_0}{3\sqrt{2d^3}}$). Let us introduce $\varepsilon' = \frac{9d^3}{4\lambda_0^2}\varepsilon_n^2$. Proceeding as before it can be proved,

$$\|v\|_2 \geq \|w\|_2 \Rightarrow \min_i |\mu - \lambda_i| \leq \sqrt{2d^3}\varepsilon \Rightarrow \|w\|_2 \leq \sqrt{\varepsilon'}, \tag{46}$$

$$\|w\|_2 \geq \|v\|_2 \Rightarrow |\mu| \leq \sqrt{2d^3}\varepsilon \Rightarrow \|v\|_2 \leq \sqrt{\varepsilon'}. \tag{47}$$

Suppose that the eigenvalues of $S$ are sorted so that $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_d$. Let us denote $u_k = (v_k, w_k)$ an associated orthonormal basis of eigenvector. Notice that, with the condition $\varepsilon \leq \varepsilon_0 \leq \frac{\lambda_0}{3\sqrt{2d^3}}$, the $l$ eigenvalues $\mu$ such that $\min_i |\mu - \lambda_i| \leq \sqrt{2d^3}\varepsilon$ are the $l$ largest eigenvalues. We are going to prove that $l = d'$.

Proceeding by contradiction, let us suppose that $l \geq d' + 1$.

First notice that for all $1 \leq i < j \leq l$: $|\langle v_i, v_j\rangle| \leq \varepsilon'$ (because $\langle u_i, u_j\rangle = 0$, (45) and Cauchy Schwartz). We also have $|\|v_i\|^2 - 1| \leq \varepsilon'$ (similarly using $\|u\|^2 = 1$ and (46)).

Now, as $l \geq d' + 1$ the vectors $v_i$ $i = 1, \ldots, l$ are linearly dependent, and then there exists $i \in \{1, \ldots, l\}$ such that $v_i = \sum_{j \neq i} \alpha_j v_j$. Now, for all $k \neq i$, on one hand: $|\langle v_i, v_k\rangle| \leq \varepsilon'$ while on the other hand: $|\langle v_i, v_k\rangle| \geq |\alpha_k| - \varepsilon' \sum_{j \notin \{i,k\}} |\alpha_j|$ so that $\varepsilon' \geq |\alpha_k| - \varepsilon' \sum_{j \notin \{i,k\}} |\alpha_j|$ and, summing this inequalities gives $(l-1)\varepsilon' \geq (1 - (l-2)\varepsilon') \sum_{k \neq i} |\alpha_k|$ so that $\sum_{k \neq i} |\alpha_k| \leq \frac{(l-1)\varepsilon'}{1-(l-2)\varepsilon'} \leq \frac{d\varepsilon'}{1-d\varepsilon'}$ and, for all $j \neq i$ $|\alpha_j| \leq \frac{d\varepsilon'}{1-d\varepsilon'}$. Thus, with very rough bounds: $\|u_i\|^2 \leq \frac{d^4\varepsilon'^2}{(1-d\varepsilon')^2} \leq 4d^4\varepsilon'^2$ (the last inequality comes from $\varepsilon \leq \varepsilon_0 \leq \frac{\lambda_0}{2\sqrt{2}d^2}$) that contradicts $\|u_i\|^2 \geq 1 - \varepsilon'$ because $\varepsilon \leq \varepsilon_0 \leq \frac{\lambda_0\sqrt{\sqrt{16d^4+1}-1}}{8d^{7/2}}$

One can obtain that $d - l \leq d - d'$ by a similar proof (reasoning on the component $w_i$ for $i \in \{l+1, \ldots d\}$), so that we can conclude that $l = d'$. Thus for all $i \leq d'$ $\|w_i\| \leq \sqrt{\varepsilon'}$ and for all $i > d'$ $\|v_i\| \leq \sqrt{\varepsilon'}$. For all $X \in \mathbb{R}^d$, let us write $X = \sum_i \alpha_i u_i$ then $P_{S,d'}X = \sum_{i=1}^{d'} \alpha_i u_i = \sum_i \alpha_i (v_i', w_i')'$ and $I_{d',d}X = \sum_{i=1}^{d} (v_i', 0)'$ so that:

$$(P_{S,d'} - I_{d',d})X = \sum_{i=1}^{d'} \alpha_i \begin{pmatrix} 0 \\ w_i \end{pmatrix} - \sum_{i=d'+1}^{d} \alpha_i \begin{pmatrix} v_i \\ 0 \end{pmatrix}.$$

from where it follows that,

$$\|(P_S - I_{d',d})X\|_2 \leq \sum_1^d |\alpha_i|\sqrt{\varepsilon'} \leq \frac{3d^{3/2}}{2\lambda_0}\varepsilon||X||_2.$$

That concludes the proof. □

## Acknowledgements

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

## References

Aamari, E., Levrard, C.(2016). Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction. *arXiv:1512.02857v1.*

Aaron, C., Cholaquidis, A., Cuevas, A.(2016). Stochastic detection of some topological and geometric feature. *arXiv:1702.05193v1.*

Aaron, C., Bodart, O.(2016). Local convex hull support and boundary estimation. *J. Of Multivariate Analysis.* **147**, 82–101.

Berger, M.(2003) A panoramic view of Riemannian geometry. *Springer-Verlag Berlin Heidelberg*

Berrendero, J.R., Cholaquidis, A., Cuevas, A. and Fraiman, R.(2014). A geometrically motivated parametric model in manifold estimation. *Statistics.* **48**(5).

Bertail, P.; Gautherat, E. and Harari-Kermaded, E.(2008), *Elect. Comm. in Probab.* **13**(1) 628–640.

Bickel, P. and Levina, E. (2005) Maximum likelihood estimation of intrinsic dimension. *Advances in NIPS - MIT Press* **17**

Camastra, F. and Staiano, A. (2016) Intrinsic dimension estimation. *Information Sciences: an International Journal*, **328**(C) 26–41.

Carlsson, G. (2009) Topology and data. *Bulletin of the American Mathematical Society*, **46**(2), 255–308.

Chevalier, J. (1976) Estimation du support et du contour de support d'une loi de probabilité. *Ann. Inst. H. Poincaré B* , 339–364.

Cuevas, A. and Rodriguez-Casal, A.(2004) On boundary estimation. *Adv. in Appl. Probab.* **36**, 340–354.

Cuevas, A. and Fraiman, R. (2009). Set estimation. In *New Perspectives on Stochastic Geometry,* eds W.S. Kendall and I. Molchanov. *Oxford University Press*, 366–389.

Cuevas, A., Fraiman, R. and Rodríguez-Casal, A.(2007) A nonparametric approach to the estimation of lengths and surface areas. *Ann. Statist.*, **35**, 1031–1051.

Devroye, L. and Wise, G. (1980) Detection of abnormal behaviour via nonparametric estimation of the support. *SIAM J. Appl. Math.* **3**, 480–488.

do Carmo, M (1992) Riemannian Geometry. *Birkhäuser*

Fefferman, C., Mitter, S., and Narayanan, H.2016 Testing the manifold hyphotesis. *J. Amer. Math. Soc.* **29**, 983–1049.

Niyogi, P., Smale, S. and Weinberger, S.(2008) Finding the Homology of Submanifolds with High Confidence from Random Samples. *Discrete Comput. Geom.* **39**. 419–441.

Niyogi, P., Smale, S. and Weinberger, S. (2011) A topological view of unsupervised learning from noisy data. *SIAM J. Comput*, **40**(3), 646–663.