HAL open science

On boundary detection

Catherine Aaron, Alejandro Cholaquidis

▶ **To cite this version:**

Catherine Aaron, Alejandro Cholaquidis. On boundary detection. 2016. hal-01291996v1

HAL Id: hal-01291996

https://hal.science/hal-01291996v1

Preprint submitted on 22 Mar 2016 (v1), last revised 4 Jul 2019 (v3)

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

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

On boundary detection

Catherine Aaron[a] and Alejandro Cholaquidis[b]
[a] Université Blaise-Pascal Clermont II, France
[b] Centro de Matemática, Universidad de la República, Uruguay

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. A heuristic for choosing a convenient value for the k_n parameter is also given. Finally we provide a simulation study of the test.

1 Introduction

Set estimation deals with the problem of estimating the support of an unknown distribution $\mathbb{P}_X$ on $\mathbb{R}^d$, from an i.i.d. sample $X_1, \ldots, X_n$ of X. It has been extensively studied when $\mathbb{P}_X$ is uniformly continuous with respect to the Lebesgue measure, that is, the full dimensional problem. A very intuitive estimator (the union of balls of radii r_n, centered at the sample points), was proposed initially by Chevalier (1976). Universal consistency was proved in Devroye and Wise (1980), whenever $r_n \to 0$ and $nr_n \to \infty$. To obtain rates of convergence for the estimators it is necessary to impose shape restrictions on the support. On one hand it is possible to find estimators that converge with an explicit (but slow) convergence rate with few shape restriction; on the other hand when shape conditions are more restrictive support estimators with faster convergence rates can be found. Some results in this respect are summarized in Cuevas and Fraiman (2010). Set estimation also tackles some related problems, such as the estimation of the boundary (see Cuevas and Rodriguez-Casal (2004)), the estimation of functional of the sets (as the length of the boundary (Cuevas et al. (2007)), the integrated mean curvature (Berrendero et al. (2014)), among others), or the recognition of topological properties having support estimator homeomorphic to the support (Aaron and Bodart (2016)).

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 it connection with nonlinear dimensionality reduction techniques (also known as *manifold learning*), as well as *persistent homology*. See for instance Fefferman, et al (2013), Niyogi, Smale and Weinberger (2008), Niyogi, Smale and Weinberger (2011). It would be natural to think that some of the proposed estimators (in the full dimensional framework) are still suitable, but, when considering the problem of estimation of the the boundary of the support it is not possible to directly adapt the methods. Indeed, the

proposed way to estimate the boundary of the support in the full dimensional case is to consider the boundary of the support estimator. Unfortunately, when the support estimator is full dimensional (which is typically the case) this idea is hopeless. To illustrate this, let us study the Devroye-Wise support estimator. It is defined as follows:

$$\hat{S}_{r_n} = \bigcup_{i=1}^{n} \mathcal{B}(X_i, r_n),$$

where $\mathcal{B}(x, r)$ denotes the closed ball (in $\mathbb{R}^d$) of radius r centered at x and $X_1, \ldots, X_n$ is the sample. In the "full dimensinal case", under some reasonable shape restriction on S as well as the condition $r_n = \mathcal{O}((\log(n)/n)^{1/d})$, it can be proved (see Cuevas and Rodriguez-Casal (2004)) that the boundary of $\hat{S}_n$ is a consistent estimator of the boundary of S. However, when the support is a d'-dimensional manifold with $d' < d$, is is easy to see that the boundary of $\hat{S}_{r_n}$ is never a consistent estimator. This is illustrated in Figure 1, in the case $d' = 1$, where the boundary of S has only two points (the extremes of the curve) while the boundary of the Devroye-Wise estimator is a one dimensional manifold.

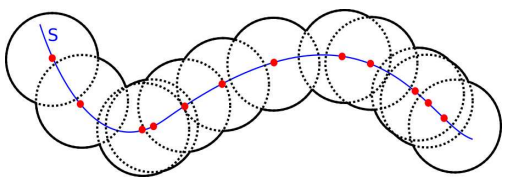

Figure 1: The solid line shows the manifold as well as the boundary of the Devroye-Wise estimator, in the one dimensional case.

Assuming d' is known there exist some support estimators which are $d'-$dimensional based on some restriction on the Delaunay complex but, to our knowledge they had only been studied recently in the case of support without boundary (see Aamari and Levrard (2016))

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 main goal of this paper is to propose a test to address that problem. In order to explain the idea of the test, let us assume that the support M, is a d'-dimensional $\mathcal{C}^2$ manifold, that, the distribution $\mathbb{P}_X$ has a density which is Lipschitz-continuous. Given a point $x \in M$, let us denote by $X_{i(x)}$ the ith observation the closest to x. We put $r_{x,n} = \|x - X_{k_n(x)}\|$ and $\bar{X}_{x,k_n} = \frac{1}{k_n}\sum_{i=2}^{k_n} X_i$. Assume that $k_n \to +\infty$ slowly enough to also have $\max_{x\in S} r_{x,n} \overset{a.s.}{\to} 0$.

Heuristically, first suppose that $\partial M = \emptyset$. Then, for all x, the smoothness of the manifold, the continuity of the density and the k_n conditions ensures that the "rescaled local sample points": $\left\{(X_{1(x)} - x)/r_{x,n} \ldots (X_{k_n(x)} - x)/r_{x,n}\right\}$ is "close" to a sample uniformly drawn on a d' dimensional unit ball. Then, as $k_n \to \infty$ we expect to have $\|\overline{X}_{x,k_n} - x\| r_{x,n} \overset{a.s.}{\to} 0$. Thus we can naturally expect that $\max_i \|\overline{X}_{X_i,k_n} - x\| r_{x,n} \overset{a.s.}{\to} 0$.

Second suppose that ∂M is a $\mathcal{C}^2$ manifold then, if $x \in \partial M$, with the same kind of argument, the "rescaled local sample points" are close to observations uniformly drawn on a half unit ball and $\|\overline{X}_{x,k_n} - x\|/r_{x,n} \to \alpha_{d'}$ with $\alpha_{d'}$ a positive constant. Thus one can naturally expect that we expect to have $\|\overline{X}_{x,k_n} - x\| r_{x,n} \overset{a.s.}{\to} \alpha_{d'}$ and that $\max_i \|\overline{X}_{X_i,k_n} - x\| r_{x,n} \overset{a.s.}{\to} \alpha_{d'}$

The proposed test statistic is based on this idea with a slightly different test statistic, built by introducing local PCA, to estimate the tangent planes in order to improve practical results.

This manuscript is organized as follows. In Section 2 we introduce the notation used throughout the paper and detail the different shape and density hypotheses. In Section 3 we present the test statistic, the associated theoretical results and a way to select suitable values for the parameter k_n. Section 4 is devoted to the proofs. Finally, in section 5 a simulation study is performed.

2 Notations, geometric framework and hypotheses

2.1 Notations

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 closed ball of radius ε centred at x will be denoted by $\mathcal{B}_k(x, \varepsilon) \subset \mathbb{R}^d$ (when $k = d$ the index will be removed) and its Lebesgue measure will be denoted as $\sigma_k = |\mathcal{B}_k(x, 1)|$. When $A = (a_{ij})$, $(i = 1, \ldots, m\ ,\ j = 1, \ldots, n)$ 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 $F_{d'}(t) = 1 - \Psi_{d'}(t)$.

2.2 Geometric framework and hypotheses

In what follows $M \subset \mathbb{R}^d$ is a d'-dimensional compact manifold of class $\mathcal{C}^2$ (in general with $d' < d$). We will consider the Riemannian metric on M inherited from $\mathbb{R}^d$. When M has boundary, as a manifold, it will be denoted by ∂M. For $x \in M$, $T_x M$ denotes the tangent space at x and φ_x the orthogonal projection on the affine tangent space $x + T_x M$. When M is orientable it has a unique associated volume form ω such that $\omega(e_1, \ldots, e_{d'}) = 1$ for all oriented orthonormal basis $e_1, \ldots, e_{d'}$ of $T_x M$. Then if $g : M \to \mathbb{R}$ is a density function, we can define a new measure $\mu(B) = \int_B g\omega$, where $B \subset M$ is a Borel set.

Since we will only be interested in measures μ, 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.

In what follows we will suppose that $X_1, \ldots, X_n$ is an i.i.d. sample of a random variable X, drawn according to a density f supported on M. We will assume that for all $x \in M$, $f(x) \geq f_0 > 0$ and that f is Lipschitz continuous, i.e. there exists K_f such that for all $(x, y) \in M^2$, $\|f(x) - f(y)\| \leq K_f \|x - y\|$.

3 The test

3.1 The test statistics

Definition 1. *Given an i.i.d. sample $X_1, \ldots, X_n$ 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:*

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

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 $i = 1, \ldots, n$. Then the proposed test statistic is:

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

Even if the definition of the test statistic involves many steps, the underlying idea is quite simple: when M is smooth the projected sample $X^*_{k(i)}$ can be seen as a sample on an estimation of the tangent plane. Then, on one hand, when X_i is far enough from the boundary, since we will suppose that the density is continuous, the $X^*_{k(i)}$ samples "look like" a uniform sample drawn on a $d'-$dimensional ball of radius r_{i,k_n} centered at the origin. Then (δ_{i,k_n}) should converge a random variable with distribution $\chi^2(d')$ (see Lemma 4 below). On the other hand when X_i is close to the boundary, the $X^*_{k(i)}$ samples "look like" a uniform sample drawn on a $d'-$dimensional half-ball of radius r_{i,k_n} centered at the origin, so that we should have $\delta_{i,k_n}/k_n \geq \alpha_d > 0$ (see Proposition 4 below). Observe that if $\partial M = \emptyset$ all the observations are far from the boundary, which allows us to control the asymptotic behavior of Δ_{n,k_n} using Equation (7) in Bertail, Gautherat and Harari-Kermadec (2008) and adding extra assumptions on k_n. If $\partial M \neq \emptyset$, we are going to prove that there is at least one of the observations which is "close enough" to the boundary (this also requires extra assumptions on k_n). The extra assumptions on k_n and the general assumptions on the distribution and manifold are summarized in the following.

Hypotheses

Thoughout this work we will assume that the underling manifold M is compact, of class $\mathcal{C}^2$, and that its boundary is either empty or of class $\mathcal{C}^2$.

K. $k_n/(\ln(n))^4 \to \infty$ and $(\ln(n))k_n^{1+d'}/n \to 0$.

P. A probability distribution $\mathbb{P}_X$ on M fulfills condition P if the density f of X (w.r.t. the volume form) is Lipschitz and for all $x \in M$, $f(x) \geq f_0 > 0$. In the following $f_1 = \max_{x \in M} f(x)$.

3.2 Theoretical Results

The first theorem presented here 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 Δ_{n,k_n} introduced in Definition 1 and rejection region $\{\Delta_{n,k_n} \geq t\}$.

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

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

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

The second theorem states that, under H_0 and the same assumptions on k_n, the empirical distribution of δ_{i,k_n} converges in mean square towards a $\chi^2(d')$ distribution. In practice we will use this result to choose the parameter k_n by selecting the value that provides the empirical cumulative distribution closest to the $\chi^2(d')$ distribution.

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 $x \in M$,

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

Now concerning the power we have:

Theorem 3. *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* (1) *with rejection zone* (2) *has power* 1 *for n large enough.*

Finally we also have a consistent decision rule :

Theorem 4. *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. *Then, with probability one, the decision rule: $\partial M = \emptyset$ if and only if $\Delta_{n,k_n} \leq \beta_n$ with*

$$\lambda \ln n \leq \beta_n \leq \mu k_n \text{ with } \lambda > 4 \text{ and } \mu \leq (d'+2) \left(\frac{\Gamma\left(\frac{d'+2}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{d'+3}{2}\right)} \right)^2$$

is consistent.

3.3 Automatic choice for k_n

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

i. If the estimated $p-$value (using $k-$nearest neigbors) is greater than 5% (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 5%, first identify the points "far from the boundary" as the observations $i \in I_k = \{F_{d'}(\delta_{i,k}) \geq 0.05\}$. Then, if we define

$$\hat{\psi}_{0.05,n,k}(x) = \frac{1}{\#I_k} \sum_{i \in I_k} \mathbb{I}_{\{\delta_{i,k} \leq x\}},$$

compute

$$d_{\chi^2}(k) = \frac{1}{\#I_k} \sum_{i \in I_k} \left| \hat{\Psi}_{0.05,n,k}(\delta_{i,k}) - \Psi_{0.05,d'}(\delta_{i,k}) \right|,$$

where $\Psi_{0.05,d'}(x) = (0.95)^{-1} \Psi_{d'}(x) \mathbb{I}_{\{\Psi_{d'}(x) \leq 0.95\}}$.

Finally choose $k = \text{argmin}_k d_{\chi^2}(k)$.

3.4 Discussion on the hypotheses

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}. Thinking of a polyhedron, when a corner becomes acute the local PCA fails to estimate a "tangent" plane at the corner, see Figure 2 right.

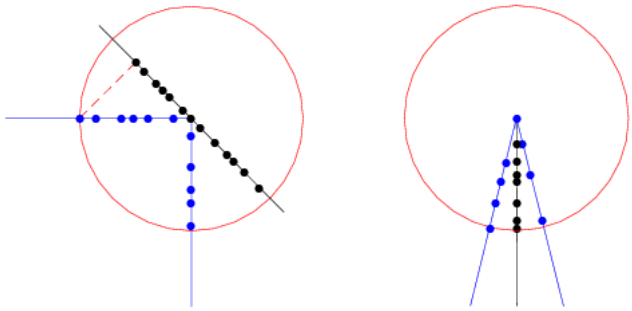

Figure 2: Behavior for polyhedron. When the angle allows us to estimate the "tangent" plane the normalization is not suitable. When the angle is too acute the projection is not accurate. The manifold, and sample points ar in blue, the estimated tangent plane and projected observations are in black.

On the other hand, 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.

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). In order to illustrate the issues discussed in this section numerically, we will present in Section 5 results for some samples that do not fulfill the hypothesis,

4 Proofs

4.1 Preliminary results

As explained above, the idea is to project the k_n-nearest neighbor of a sample points X_i orthogonally onto the d'-dimensional space spanned by the eigenvectors corresponding to the highest d' eigenvalues, and prove that by taking k_n large enough (which will entail that r_{i,k_n} goes to zero, where r_{i,k_n} is as in Definition 1), the projected sample behaves like a uniform sample on a ball. In order to do that, we will prove some technical lemmas.

4.1.1 Preliminary geometric lemmas on compact uniform $\mathcal{C}^2$ $d'-$ manifold with bounded curvature

The aim of this section is to present some geometric lemmas that quantify the local aspect of the support. They are all direct consequences of the $\mathcal{C}^2$ smoothness of M (and smoothness of ∂M when $\partial M \neq \emptyset$) and compactness.

Lemma 1. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional $\mathcal{C}^2$ manifold (with $d' < d$). Let $\varphi_x : M \to x + T_x M$ be the orthogonal projection onto the tangent affine plane. Then, there exists $\rho_{M,0} > 0$, such that: for all $x \in M$, φ_x is a bijection from $M \cap \mathcal{B}(x, \rho_{M,0})$ to $\varphi_x\big(M \cap \mathcal{B}(x, \rho_{M,0})\big)$.*

Proof. Proceeding by contradiction, we take a sequence $r_n \to 0$ and three sequences of points, 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) $x_n \to x \in M$, (which implies $y_n \to x$ and $z_n \to x$) and $w_n \doteq \frac{y_n - z_n}{\|y_n - z_n\|} \to w$. 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 γ_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 $r_0 > 0$. Therefor (see Proposition 88 in Berger (2003)), if we take n large enough that $r_n \leq r_0/2$, we may take γ_n to be the (unique) geodesic which is the image, by the exponential map, of a vector $v_n \in T_{y_n} M$. The Taylor expansion of the exponential map shows that $w_n = \frac{v_n}{\|y_n - z_n\|} + o(1)$. Then, taking the limit as $n \to \infty$ we get $w \in T_x M$ which contradicts the fact that $w \in (T_x M)^\perp$. $\square$

Corollary 1. *If M is a d'-dimensional compact manifold of class $\mathcal{C}^2$ then there exists $r_{0,M} > 0$ such that: for all $x \in M$ $\varphi_x : M \cap \mathcal{B}(x, r_{M,0}) \to \varphi_x(M \cap \mathcal{B}(x, r_{M,0}))$ is a bijective function of class $\mathcal{C}^2$. Moreover, let $e_{x,1}, \ldots, e_{x,d}$ be an orthonormal basis of $\mathbb{R}^d$ such that $e_{x,1}, \ldots, e_{x,d'}$ is an orthonormal basis of $T_x M$. Then there exist $\mathcal{C}^2$ functions:*

$$\Phi_{x,k} : \varphi_x\big(M \cap \mathcal{B}(x, r_{M,0})\big) - x \subset T_x M \to \mathbb{R}, \quad k = d' + 1, \ldots, d,$$

such that:

$$\varphi_x^{-1} : \varphi_x\big(M \cap \mathcal{B}(x, r_{M,0})\big) \to M \cap \mathcal{B}(x, r_{M,0})$$

$$x + \begin{pmatrix} y_1 \\ \vdots \\ y_{d'} \\ 0_{d-d'} \end{pmatrix} \mapsto x + \begin{pmatrix} y \\ \Phi_{x,d'+1}(y) \\ \vdots \\ \Phi_{x,d}(y) \end{pmatrix} \text{ with } y = \begin{pmatrix} y_1 \\ \vdots \\ y_{d'} \end{pmatrix} \text{ and } 0_{d-d'} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^{d-d'},$$

where the vectors are written in the above-mentioned bases. These functions satisfy: $\vec{\nabla}\Phi_{x,k}(0) = 0$ and there exists λ_M such that for every eigenvalue $\lambda_{x,k}$ of $H_{\Phi(x,k)}(0)$ we have $|\lambda_{x,k}| \leq \lambda_M$.

Sketch of proof.
In view of Lemma 1 there exist functions:

$$\Phi_{x,k} : \varphi_x\big(M \cap \mathcal{B}(x, \rho_{M,0})\big) - x \to \mathbb{R}$$

such that:

$$\varphi_x^{-1}: \quad \varphi_x\big(M \cap \mathcal{B}(x, \rho_{M,0})\big) \to M \cap \mathcal{B}(x, \rho_{M,0}) \tag{3}$$
$$x + \begin{pmatrix} y_1 \\ \vdots \\ y_{d'} \\ 0_{d-d'} \end{pmatrix} \mapsto x + \begin{pmatrix} y \\ \Phi_{x,d'+1}(y) \\ \vdots \\ \Phi_{x,d}(y) \end{pmatrix}$$

The $\mathcal{C}^2$ regularity of M implies that φ_x^{-1} restricted to some $\varphi_x\big(M \cap \mathcal{B}(x, r_{x,0})\big)$ is of class $\mathcal{C}^2$ and compactness allows us to find a uniform $r_{0,M}$. The gradient condition just reflects the fact that φ_x is the projection on the tangent space. The Hessian condition, which is a direct consequence of the smoothness and compactness of M, gives a bound for the maximal curvature. □

With the notation introduced in Corollary 1 one can compute $\mu(V)$ by integrating the density function f with respect to the volume form, by integration on the tangent space when $V \subset M$ is a set of diameter inferior to $r_{0,M}/2$, so that there exists $x \in M$ with $V \subset \mathcal{B}(x, r_{0,M})$.

Let us denote:

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

We have

$$V \subset \mathcal{B}(x, r_{0,M}) \Rightarrow \mu(V) = \int_V f dw = \int_{\varphi_x(V)} f(\varphi_x^{-1}(y)) \sqrt{\det G_x(y)} dy. \tag{4}$$

Corollary 2. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional $\mathcal{C}^2$ manifold (with $d' < d$) . Let $\varphi_x : M \to x + T_x M$ be the orthogonal projection onto the tangent affine plane.*

a) There exists $r_{M,1} > 0$ and c_1, such that, for any $x \in M$ and $y \in T_x M$, with $\|y\| \leq r_{M,1}$ we have $|\sqrt{\det G_x(y)} - 1| \leq c_{M,1}\|y\|$.

b) There exists $c_2 > 0$ and $r_{M,2}$, such that, if $\|x - x'\| \leq r_{M,2}$ then $\|\varphi_x(x') - x'\| \leq c_{M,2}\|x - \varphi_x(x')\|^2$.

Sketch of proof.

a) First let us introduce the matrix

$$E_x(y) = \begin{pmatrix} \vec{\nabla}\Phi_{x,d'+1}(y) \\ \vdots \\ \vec{\nabla}\Phi_{x,d}(y) \end{pmatrix}.$$

Observe that $G_x(y) = I_{d'} + E_x(y)'E_x(y)$. Then, the Taylor expansion of the determinant gives $\det(G_x(y)) = 1+\mathrm{tr}(E_x(y)'E_x(y))+o(\|E_x(y)'E_x(y)\|_\infty)$. So that $|\sqrt{\det G_x(y)}-1| \leq \frac{1}{2}|\mathrm{tr}(E_x(y)'E_x(y))|+o(\|E_x(y)'E_x(y)\|_\infty)$. Then using Taylor expansions of $\{\frac{\partial}{\partial e_i}\Phi_{x,k}\}_{i=1,\dots,d'}$ around 0, we get $\|E_x(y)\|_\infty \leq \lambda_M\|y\| + o(\|y\|)$. Finally, compactness allows us to find $r_{M,1}$ associated to any $c_1 > d(d-d')\lambda_M^2$.

b) By Pythagoras $\|\varphi_x(x')-x'\|^2 = \|x-x'\|^2 - \|\varphi_x(x')-x\|^2$. Introducing $y = \varphi_x(x')-x$ we obtain $\|\varphi_x(x')-x'\|^2 \leq (d-d')\lambda_M^2\|y\|^4 + o(\|y\|^4)$ and, as $\|x-x'\| > \|y\|$ we have: $\|\varphi_x(x')-x'\| \leq \sqrt{d-d'}\lambda_M\|x-x'\|^2 + o(\|x-x'\|^2)$. Once again we use a compactness argument to conclude the proof with any $c_2 > \sqrt{d-d'}\lambda_M$. □

Lemma 2. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional (with $d' < d$) $\mathcal{C}^2$ manifold without boundary. For all $x \in M$ and for all $r \leq \min(r_{M,0}, r_{M,2})$,*

$$\mathcal{B}\big(x, r(1-c_{M,2}r)\big) \cap (x + T_xM) \subset \varphi_x(\mathcal{B}(x,r) \cap M).$$

Proof. Proceeding by contradiction, suppose that there exists $0 < r \leq \min(r_{M,0}, r_{M,2})$, $x \in M$, $y \in \mathcal{B}\big(x, r(1-c_{M,2}r)\big) \cap T_xM$ such that $y \notin \varphi_x\big(\mathcal{B}(x,r)\cap M\big)$. As $x \in \varphi_x(\mathcal{B}(x,r)\cap M)$ and φ_x is continuous, the line segment $[x,y]$ intersects $\partial\varphi_x\big(\mathcal{B}(x,r)\cap M\big)$. Let $z \in [x,y] \cap \partial\varphi_x\big(\mathcal{B}(x,r)\cap M\big)$. Clearly we have $\|x-z\| < \|x-y\| < r(1-c_2r)$. Since $r \leq r_{M,0}$, there exists z_0 such that $z = \varphi_x(z_0)$. Again using that φ_x is a continuous function, $z_0 \in \partial\big(M \cap \mathcal{B}(x,r)\big)$. Since $\partial M = \emptyset$, $\partial\big(M \cap \mathcal{B}(x,r)\big) = \big(M \cap \partial\mathcal{B}(x,r)\big) \cup \big(\partial M \cap \mathcal{B}(x,r)\big) = M \cap \partial\mathcal{B}(x,r)$, we have $z_0 \in \partial\mathcal{B}(x,r)$, so $\|x-z_0\| = r$. Finally we have $r \leq r_{M,2}$, $\|x-z_0\| = r$ and $\|x-\varphi_x(z_0)\| < r(1-c_{M,2}r)$, so by Corollary 2 part b),

$$r^2 = \|x-z\|^2 + \|z-z_0\|^2 < r^2(1-c_{M,2}r)^2 + \big(c_{M,2}r^2(1-c_{M,2}r)^2\big)^2 \leq r^2$$

which is impossible. □

Lemma 3. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional $\mathcal{C}^2$ manifold with a $\mathcal{C}^2$ boundary (with $d' < d$). Then for all $x \in \partial M$ there exists a unit vector u_x such that, for all $r \leq \min(r_{M,0}, r_{M,1}, r_{M,2}, r_{\partial M,0}, r_{\partial M,1}, r_{\partial M,2})$, :*

$$(x + T_xM) \cap \mathcal{B}(x, r - c_{M,2}r^2) \cap \big\{y : \langle y-x, u_x\rangle \geq (c_{M,2} + c_{\partial M,2})r^2\big\} \subset \varphi_x\big(\mathcal{B}(x,r)\cap M\big), \tag{5}$$

and

$$\varphi_x\big(\mathcal{B}(x,r)\cap M\big) \subset (x + T_xM) \cap \mathcal{B}(x,r) \cap \big\{y, \langle y-x, u_x\rangle \geq -(c_{M,2} + c_{\partial M,2})r^2\big\}. \tag{6}$$

Sketch of proof.
Let us take a unit vector $v_x \in T_xM \cap (T_x\partial M)^\perp$ so that an application of Corollary 2 part b) on ∂M ensures that, for all $y \in \mathcal{B}(x,r)\cap\partial M$, $|\langle y-x, v_x\rangle| \leq c_{\partial M,2}r^2$. Applying now Corollary 2 part b) on M we see that: for all $y \in \mathcal{B}(x,r)\cap\partial M$, $|\langle \varphi_x(y)-x, v_x\rangle| \leq (c_{M,2} + c_{\partial M,2})r^2$. This, in addition to $\varphi_x\big(\mathcal{B}(x,r)\cap M\big) \subset \mathcal{B}(x,r)\cap(x+T_xM)$, implies that:

i. $\varphi_x\big(\mathcal{B}(x,r)\cap M\big) \subset T_xM \cap \mathcal{B}(x,r) \cap \big\{y \in \mathbb{R}^d : \langle y-x, v_x\rangle \geq -(c_{M,2} + c_{\partial M,2})r^2\big\}$ in which case take $u_x = v_x$;

ii or : $\varphi_x\big(\mathcal{B}(x,r)\cap M\big) \subset T_xM \cap \mathcal{B}(x,r) \cap \big\{y \in \mathbb{R}^d : \langle y-x, v_x\rangle \leq (c_{M,2}+c_{\partial M,2})r^2\big\}$ in which case take $u_x = -v_x$;

this is (6). The inclusion (5) is obtained by combining this kind of argument and those of Lemma 2. □

Proposition 1. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional (with $d' < d$) $\mathcal{C}^2$ manifold without boundary. Let X be a random variable whose distribution, $\mathbb{P}_X$, is supported by M, and whose density f is Lipschitz continuous. Then, there exist positive constants R_1 and C_1 such that: if $r \leq R_1$, then $\big|\mathbb{P}_X(\mathcal{B}(x,r)) - f(x)\sigma_{d'}r^{d'}\big| \leq C_1 r^{d'+1}$.*

Proof. Let us take $r < \min(r_{M,0}, r_{M,1}, r_{M_2})$,

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

Applying first the Lipschitz hypothesis on f we get,

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

Now by formula (4):

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

By Corollary 2 a):

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

Finally applying Lemma 2:

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

This implies:

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

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

Proofs of the following proposition is similar to the previous one and are left to the reader.

Proposition 2. *Let $M \subset \mathbb{R}^d$ be a compact, d'-dimensional (with $d' < d$) $\mathcal{C}^2$ manifold with a $\mathcal{C}^2$ boundary. Let X be a random variable whose distribution, $\mathbb{P}_X$, fulfills condition* P. *Then, there exists constants C_2 and R_2 such that for all $r \leq R_2$ and all $x \in \partial M$, we have, $\frac{f_0 \sigma_{d'}}{2} r^{d'} - C_2 r^{d'+1} \leq \mathbb{P}_X(\mathcal{B}(x,r)) \leq \frac{f_1}{2}\sigma_{d'} r^{d'} + C_2 r^{d'+1}$, where $f_1 = max_{x \in M} f(x)$.*

In the sequel the radius r_M and the constant c_M are defined as follows:

Definition 2. *If M is a $\mathcal{C}^2$ manifold without boundary, and f is a Lipschitz continuous function on M bounded bellow by $f_0 > 0$, we define $r_M = \min(r_{M,0}, r_{M,1}, r_{M,2}, R_1)$ and $c_M = \max(c_{M,1}, c_{M,2}, C_1)$.*

If M is a $\mathcal{C}^2$ manifold with a $\mathcal{C}^2$ boundary and $0 < f_0 \leq f \leq f_1 < +\infty$ on M, we define $r_M = \min(r_{M,0}, r_{M,1}, r_{M,2}, r_{\partial M,0}, r_{\partial M,1}, r_{\partial M,2}, R_2)$. and $c_M = \max(c_{M,1}, c_{M,2} + c_{\partial M,2}, C_2)$

4.1.2 Preliminary probabilistic results

In order to state 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 4. *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

$$\begin{aligned}\mathrm{Var}(X_{.,i}) =& \int_{-1}^{1} x^2 \frac{1}{\sigma_d}\sigma_{d-1}(1-x^2)^{(d-1)/2}dx \\ =& \frac{\sigma_{d-1}}{\sigma_d}\int_0^1 u^{1/2}(1-u)^{(d-1)/2}du \\ =& \frac{\sigma_{d-1}}{\sigma_d}B\big(3/2,(d+1)/2\big),\end{aligned}$$

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, Gautherat and Harari-Kermadec (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,

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

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:

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

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. □

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.*

$$\mathbb{E}\left(\frac{\langle X-x, u\rangle}{r}\right) = \alpha_d, \tag{17}$$

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 $(1, 0, \ldots, 0)$ (with $r > 0$). Then Y has uniform distribution on $\mathcal{B}_{e_1}(0,1)$ and so (17) holds. □

Lemma 5. *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 ε_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 Proposition 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

$$\mathbb{P}\left(r_n \geq \frac{a}{k_n}\right) \leq A_M \varepsilon_n^{-d} k_n^{1+d} \binom{k_n}{n} \left(1 - \frac{f_0 \sigma_d (a-\varepsilon_n)^d}{k_n^d}(1+o(1))\right)^{n-k_n}. \tag{18}$$

Applying Stirling's formula to the right hand side of (18), 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,

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

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

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

and then

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

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

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

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 6. *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,

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

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,

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

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

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

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

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

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

□

Lemma 7. *Let $X_1, \ldots, X_n$ 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*

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

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

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

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

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

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,

$$\|r^{-2}_{i,k_n}\hat{S}_{i,k_n} - r^{-2}_{i,k_n}S_i\|_\infty \leq \sqrt{\frac{3\ln(n)}{2k_n}} \quad \text{for all } i = 1, \ldots, n. \tag{19}$$

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 $\mathbb{R}^d$. Corollary 2, Lemma 2 and the Lipschitz continuity of the density ensures that $r^{-2}_{i,k_n}S_i \to \Sigma^d_{X_i}$ where $\Sigma^d_{X_i} = P'_iJ_{d'}P_i$ is

the covariance matrix of a uniform variable drawn on $\mathcal{B}(0,1)\cap T_{X_i}M$. More precisely it can be proved that there exists r and c such that: when $r_n \leq r$,

$$\text{for all } i : \left\| r_{i,k_n}^{-2} S_i - \frac{1}{d'+2} P_i' J_{d'} P_i \right\|_\infty \leq c r_n \quad \text{, where } J_{d'} = \begin{pmatrix} I_{d'} & 0 \\ 0 & 0 \end{pmatrix}. \tag{20}$$

Now, (19) and (20) give that, with probability one, for n large enough and for all $i = 1, \ldots, n$.

$$\left\| r_{i,k_n}^{-2} \hat{S}_{i,k_n} - \frac{1}{d'+2} P' J_{d'} P \right\|_\infty \leq \sqrt{\frac{3\ln(n)}{2k_n}} + c r_n = \sqrt{\frac{3\ln(n)}{2k_n}}(1+o(1)). \tag{21}$$

In what follows we consider n large enough to ensure (21), and $\varepsilon_n = \sqrt{\frac{3\ln(n)}{2k_n}} + c r_n \leq \frac{1}{4\sqrt{2d}(d'+2)}$.

Since (21) holds for all i, from now on we will 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 also assume that $P = I_d$ (the tangent space is spanned by the d' first vectors of the canonical basis of $\mathbb{R}^d$).

The simplified version of (21) is thus:

$$\left\| r_{k_n}^{-2} \hat{S}_{k_n} - \frac{1}{d'+2} J_{d'} \right\|_\infty \leq \varepsilon_n. \tag{22}$$

Let U_i be an eigenvector of $r_{k_n}^{-2}\hat{S}_{k_n}$ with $\|U_i\|_2 = 1$, associated to an eigenvalue λ_i. If we denote $U_i = (U_{i,1}', U_{i,2}') \in \mathbb{R}^{d'} \times \mathbb{R}^{d-d'}$ then from (21) we have:

$$\max\left(\left| \frac{1}{d'+2} - \lambda_i \right| \|U_{i,1}\|_\infty, |\lambda_i| \|U_{i,2}\|_\infty \right) \leq \varepsilon_n \max(\|U_{i,1}\|_\infty, \|U_{i,2}\|_\infty).$$

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

$$\max\left(\left| \frac{1}{d'+2} - \lambda_i \right| \|U_{i,1}\|_2, |\lambda_i| \|U_{i,2}\|_2 \right) \leq \sqrt{d}\varepsilon_n. \tag{23}$$

Suppose that $\|U_{i,1}\|_2 \geq \|U_{i,2}\|_2$ then $\|U_{i,1}\|_2 \geq 1/\sqrt{2}$. Then (23) successively implies $\left|\frac{1}{d'+2} - \lambda_i\right| \leq \sqrt{2d}\varepsilon_n$ and $|U_{i,2}\|_2 \leq \frac{\sqrt{d}\varepsilon_n}{(d'+2)^{-1} - \sqrt{2d}\varepsilon_n}$. Finally, the condition on n provides $\|U_{i,2}\|_2 \leq \frac{4(d'+2)\sqrt{d}}{3}\varepsilon_n$. Let us introduce $\varepsilon_n' = \frac{16(d'+2)^2 d}{9}\varepsilon_n^2$. We have (the proof of (25) being similar to the proof of (24)):

$$\|U_{i,1}\|_2 \geq \|U_{i,2}\|_2 \Rightarrow \left| \frac{1}{d'+2} - \lambda_i \right| \leq \sqrt{2d}\varepsilon_n \Rightarrow \|U_{i,2}\|_2 \leq \varepsilon_n', \tag{24}$$

$$\|U_{i,2}\|_2 \geq \|U_{i,1}\|_2 \Rightarrow |\lambda_i| \leq \sqrt{2d}\varepsilon_n \Rightarrow \|U_{i,1}\|_2 \leq \varepsilon_n'. \tag{25}$$

Suppose now that n is large enough to have: $\varepsilon_n < (2\sqrt{2d}(d'+2))^{-1}$ (that is $\left|\frac{1}{d'+2} - \lambda_i\right| \leq \sqrt{2d}\varepsilon_n \Rightarrow |\lambda_i| > \sqrt{2d}\varepsilon_n$ and $|\lambda_i| \leq \sqrt{2d} \Rightarrow \left|\frac{1}{d'+2} - \lambda_i\right| > \sqrt{2d}\varepsilon_n$); $d\varepsilon'_n \leq 10^{-1}$ and $d^2\left(\frac{10d\varepsilon'_n}{9}\right)^2 d < 1 - \varepsilon'_n$.

Suppose that the eigenvalues are sorted so that $\lambda_1 \geq \lambda_2 \geq \ldots, \geq \lambda_d$ and denote by U_k the eigenvector associated to λ_k. Denote by l the last index such that $\left|\frac{1}{d'+2} - \lambda_i\right| \leq \sqrt{2d}\varepsilon_n$. We are going to prove that for n large enough, $l = d'$.

First notice that for all $1 \leq j < k \leq l$: $|\langle U_{j,1}, U_{k,1}\rangle| \leq \varepsilon'_n$ (because $\langle U_j, U_k\rangle = 0$, so, by (24) and Cauchy Schwartz inequality $|\langle U_{j,1}, U_{k,1}\rangle| \leq \varepsilon'_n$). We also have $|\|U_{j,1}\|^2 - 1| \leq \varepsilon'_n$ (similarly using $||U_j||^2 = 1$ and (24)).

Proceeding by contradiction, if $l \geq d'+1$ then since for all j, $U_{j,1} \in \mathbb{R}^{d'}$, the projections $U_{j,1}$ $j = 1, \ldots, l$ are linearly dependent, and then there exists $k \in \{1, \ldots, l\}$ such that $U_{k,1} = \sum_{k' \in K} \alpha_{k'} U_{k',1}$, where $K = \{k' \in \mathbb{N}, k' \leq l, k' \neq k\}$. Now, for all $l \in K$, on one hand: $|\langle U_{k,1}, U_{k',1}\rangle| \leq \varepsilon'_n$ while on the other hand: $|\langle U_{k,1}, U_{k',1}\rangle| \geq |\alpha_l| - \varepsilon'_n \sum_{k' \in K} |\alpha_{k'}|$ so that $\varepsilon'_n \geq |\alpha_l| - \varepsilon'_n \sum_{k' \in K} |\alpha_{k'}|$ and, summing this inequalities gives $\sum_{k' \in K} |\alpha_{k'}| \leq \frac{d'\varepsilon'_n}{1 - d'\varepsilon'_n}$. Finally, conditions on n gives $\sum_{k' \in K} |\alpha_{k'}| \leq \frac{10}{9} d\varepsilon'_n$ so that, for all $k' \in K$, $|\alpha_{k'}| \leq \frac{10}{9} d\varepsilon'_n$. This implies that $||U_{k,1}||^2 = \sum_{(k_1,k_2) \in K^2} \alpha_{k_1}\alpha_{k_2}\langle U_{k_1,1}, U_{k_2,2}\rangle \leq d^2\left(\frac{10d\varepsilon'_n}{9}\right)^2$ which contradicts $|\|U_{j,1}\|^2 - 1| \leq \varepsilon'_n$ for the given conditions on n.

One can obtain that $d - l \leq d - d'$ by a similar proof (reasoning on the second component of the eigenvector), so that we can conclude that for n large enough, $l = d'$ which implies that the d' largest eigenvalues of $r_{i,k_n}^{-2}\hat{S}_{i,k_n}$ are associated to d' eigenvectors U_i such that $\|U_{i,2}\|_2 \leq \sqrt{\varepsilon'_n}$. For any $V = (V'_1, V'_2)'$, let us denote by V^* its projection onto the plane spanned by $U_1, \ldots, U_{d'}$. We have:

$$V^* - (V'_1, 0)' = \sum_{i=1}^{d'} \langle V_2, U_{i,2}\rangle U_i - \sum_{i=d'+1}^{d} \langle V_1, U_{i,1}\rangle U_i.$$

Thus:

$$\|V^* - (V'_1, 0)'\|_2 \leq d\|V\|_2\sqrt{\varepsilon'_n} = \frac{4(d'+2)d^{3/2}}{3}\sqrt{\frac{3\ln(n)}{2n}}(1 + o(1))\|V\|_2.$$

This concludes the proof for any constant $A_d > \frac{2\sqrt{2}(d+2)d^{3/2}}{\sqrt{3}}$. □

4.2 Proof of Theorems 1 and 2

Theorems 1 and 2 are corollaries of the following Lemma.

Lemma 8. *Let (k_n) be a sequence which fulfills condition* K *and $X_1, \ldots, X_n$ 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 $i = 1, \ldots, n$, we can built δ^*_{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 $X^i_1, \ldots, X^i_{k_n}$ 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 δ^*_{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 $X^i_1, \ldots X^i_{k_n}$ 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 $Y^i_1, \ldots Y^i_{k_n}$ 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, (using the constant introduced in Definition 2),

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

By Corollary 2 part a), $\sqrt{\det(G_{X_i}(x))} > 1 - c_M r_{i,k_n}$. Observe that by Lemma 5 we can take n large enough such that, for all $x \in B^i_n$:

$$g^i(x) \geq \frac{1 - c_M r^2_{i,k_n}}{\sigma_{d'} r^{d'}_{i,k_n} \left(\frac{c_M r_{i,k_n}}{f_0 \sigma_{d'}} + 1\right)} \geq 0\,; \tag{26}$$

Notice that, by Lemma 2 we have:

$$\mathcal{B}\Big(X_i, r_{i,k_n}\big(1 - c_M r_{i,k_n}\big)\Big) \cap (X_i + T_{X_i}M) \subset B^i_n \subset \mathcal{B}\big(X_i, r_{i,k_n}\big) \cap (X_i + T_{X_i}M). \tag{27}$$

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 6. Equations (26), (27) and the assumptions on r_n and n allows us to claim that $\mathcal{Y}^i = \{Y^i_1, \ldots Y^i_{k_n}\}$ has the same law as $\mathcal{Z}^i = \{Z_1, \ldots Z_{k_n}\}$, 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^i_n with a probability $1 - p_n$.

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

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

Therefore, applying Lemma 4 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 δ_{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

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

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 6 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 4 and Proposition 3 point *iii*, $\max_i \sqrt{\delta^*_{i,k_n}} \leq \sqrt{5\ln(n)}$ e.a.s. Thus

$$\sqrt{\ln(n)}\max_i |e_{i,k_n}| \xrightarrow{a.s.} 0. \tag{28}$$

Now, by Lemma 7 we have, for all i: $\delta_{i,k_n} = \delta^Y_{i,k_n} + e'_{i,k_n}$ with $|e'_{i,k_n}| \leq A_d\sqrt{\frac{\ln(n)}{k_n}}(2\sqrt{d}+d)\delta^Y_{i,k_n}$ 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 (28) 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 8, $P\big(\Delta_{n,k_n} \geq t_n(\alpha)\big) \leq \alpha + o(1)$.

Proof of Theorem 1. Theorem 1 It is a direct consequence of Lemma 5 and 8. 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 5 $\mathbb{P}_{H_0}(r_n > 1/k_n) \to 0$. 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 8 we have $\hat{\Psi}_n(x) \overset{L^2}{\to} \Psi_{d'}(x)$.

Proof of Theorem 2. A direct consequence of Lemma 8 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 ε_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

$$\begin{aligned} p_{x,n} &\leq \sigma_{d'}\big((2+\varepsilon_n)/k_n\big)^{d'} f(x)\big(1+(2+\varepsilon_n)K_f/k_n\big) \\ &\leq \sigma_{d'}(3/k_n)^{d'} f(x)\big(1+3K_f/k_n\big). \end{aligned} \tag{29}$$

In the same way,

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

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 $\lambda_n > 1$:

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

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 (29),

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

Let us cover M with $x_1,\ldots,x_{\nu_n}$ (deterministic) balls of radius ε_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,

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

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

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

Now, let

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

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

4.3 Proof of Theorems 3 and 4

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

Lemma 9. *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 *where M has boundary. 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 ∂M is $\mathcal{C}^2$, its inner packing number $\nu(\varepsilon)$ (the maximal number of balls, centered in ∂M, of radius ε 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 $i \in \{1, \ldots, v(\varepsilon)\}$, 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 2 for the uniform density on M, there exist 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:

$$\big|\partial M \oplus \mathcal{B}(0,\varepsilon)\big|_{d'} \geq B\sigma_{d'}\frac{\varepsilon}{2} - BC\varepsilon^2. \tag{30}$$

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 ∂M such that $\|X_{i_0} - x_0\| \leq c_{\partial M}\ln(n)/n$, and by u_{x_0} the unit vector defined in Lemma 3. Let us introduce $Y_{k(i_0)} = \varphi_{x_0}(X_{k(i_0)})$.

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

$$\frac{\frac{1}{k_n}\sum_{k=1}^{k_n}\langle Y_{k(i_0)} - x_0, u_{x_0}\rangle}{r_{i_0,k_n}} \xrightarrow{a.s.} \alpha_{d'}. \tag{31}$$

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 Lemma 3, $\langle Y_{k(i_0)} - x_0, u_{x_0}\rangle \in [-c_M\rho_{n,+}^2, \rho_{n+}]$, so that applying Hoeffding's inequality,

$$\mathbb{P}\left(\left|\frac{1}{k_n\rho_{n,+}(1+c_M\rho_{n,+})}\sum_{k=1}^{k_n}\langle Y_{k(i_0)} - x_0, u_{x_0}\rangle - \frac{\mathbb{E}(\langle Y_{k(i_0)} - x_0, u_{x_0}\rangle)}{\rho_{n,+}(1+c_M\rho_{n,+})}\right| \geq t\right) \leq 2\exp(-2t^2k_n). \tag{32}$$

Then, to prove (31) it only remains to prove that, for all $X_{i_0} \in \partial M \oplus \mathcal{B}\big(0, c_{\partial M}\ln(n)/n\big)$:

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

Indeed:

i. From (b) and (32) we obtain

$$\frac{1}{k_n(\rho_{n,+} + c_M\rho_{n,+}^2)}\sum_{k=1}^{k_n}\langle Y_{k(i_0)} - x_0, u_{x_0}\rangle \xrightarrow{a.s.} \alpha_{d'}, \tag{33}$$

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 (33) and (a) we get (31).

First assume that $r_{i_0,k_n} \xrightarrow{a.s.} 0$ (the proof is is similar to the proof of Lemma 5, using a covering of ∂M instead of M, and bounding the probability according to Proposition 2 instead of Corollary 1). 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 ∂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:

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M} \ln(n)/n), r_{i_0,k_n} \leq R_n^-\Big) \leq \sum_{i=1}^{\nu_n} \mathbb{P}\Big(\#\big\{\mathcal{B}\big(x_i, R_n^- + 2c_{\partial M} \ln(n)/n\big) \cap \mathcal{X}_n\big\} \geq k_n\Big). \quad (34)$$

Since $R_n^- = (\ln(n) - 2c_{\partial M}) \ln(n)/n$, if we apply Proposition **??** we can bound the right hand side of (34) 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

$$\mathbb{P}\Big(\#\big\{\mathcal{B}\big(x_i, R_n^- + 2c_{\partial M} \ln(n)/n\big) \cap \mathcal{X}_n\big\} \geq k_n\Big) \leq \sum_{j=k_n}^{n} \frac{1}{j!} \left(\frac{f_1 \sigma_{d'} (\ln(n))^{2d'}}{2n^{d'-1}}(1 + o(1))\right)^j. \quad (35)$$

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 (35) we obtain:

$$\mathbb{P}\Big(\exists X_{i_0} \in \partial M \oplus \mathcal{B}(0, c_{\partial M} \ln(n)/n), r_{i_0,k_n} \leq R_n^-\Big) \leq B' \left(\frac{n}{\ln n}\right)^{d'-1} \frac{\left(\frac{f_1 \sigma_{d'} (\ln(n))^{2d'}}{2n^{d'-1}}(1 + o(1))\right)^{k_n}}{k_n!} \exp\left(\frac{f_1 \sigma_{d'} (\ln(n))^{2d'}}{2n^{d'-1}}(1 + o(1))\right). \quad (36)$$

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 (36) 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 (36) 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 8, the Lipschitz continuity of f, Corollary 2 part a) and Lemma 3 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,

$$\left| \frac{\sigma_{d'} r_{i_0,k_n}^{d'}}{2} \mathbb{E}\big(\langle Y - x_0, u_{x_0}\rangle | r_{i_0,k_n}\big) - \int_{B_0} \langle x - x_0, u_{x_0}\rangle dx \right| \leq$$
$$\varepsilon_n \int_{B_0} \|x\| dx \leq \varepsilon_n \int_{\mathcal{B}(x_0,\rho_{n,+})} \|x\| dx \leq \varepsilon_n \frac{\sigma_{d'-1}}{d'+1} \rho_{n,+}^{d'+1}. \quad (37)$$

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 3, we get,

$$\mathcal{B}(x_0, \rho_{n-}) \cap \big\{y : \langle y - x_0, u_{x_0}\rangle \geq c_M \rho_{n,+}^2\big\} \subset B_0$$
$$\subset \mathcal{B}(x_0, \rho_{n,+}) \cap \big\{y, \langle y - x_0, u_{x_0}\rangle \geq -c_M \rho_{n,+}^2\big\} \quad (38)$$

From (38) 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:

$$\left| \int_{B_0} \langle x - x_0, u_{x_0}\rangle dx - \int_{\mathcal{B}_{u_{x_0}}(x_0,r_{i_0})} \langle x - x_0, u_{x_0}\rangle dx \right| \leq \sigma_{d'}(\rho_{n,+}^{d'+1} - \rho_{n,-}^{d'+1}) + 2c_{\partial M}\sigma_{d'-1}\rho_{n,+}^{d'+2}. \quad (39)$$

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 (37) and (39) provides the existence of C and C' such that

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

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 (31) is now proved.

Now, in order to finish the proof of the Lemma, notice that, reasoning similarly to what has been done in Lemma 7 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

$$\frac{\|\sum_{k=1}^{k_n}X^*_{k(i_0)}\|}{k_nr_{i_0,k_n}}\geq(1-\|F_{n,i_0}\|_\infty)\frac{\frac{1}{k_n}\sum_{k=1}^{k_n}\langle Y_{k(i_0)}-x_0,u_{x_0}\rangle}{r_{i_0,k_n}}-(1+\|F_{n,i_0}\|_\infty)\frac{c_{\partial M}\ln(n)}{nr_{i_0,k_n}}. \tag{40}$$

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 the conditions $k_n\gg(\ln(n))^4$ ensures 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 (40).

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 8, Proposition 3 and the Borel-Cantelli Lemma ensure that $\mathbb{P}_{H_0}(\Delta_{n,k_n}<t_n)=1$ for n large enough. □

5 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 Moebius ring, and a torus (for these two last examples the samples are not uniform). We also study the test for samples that do not fulfill the hypotheses as $S_{2,+,+}$ the quarter of a 2 dimensional sphere (the boundary is not $\mathcal{C}^2$), a drawn according to a not continuous density on a circle and a uniform drawn on a square (the manifold is not $\mathcal{C}^2$).

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_{χ^2} (blue). In order to have comparable curves d_{χ^2} has been artificially rnormalized 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_{χ^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 where $\frac{2e^3}{9} F_{d'}(\delta_{i,k}) \leq 5\%$ that is the one that are expected be located "near to" the boundary.

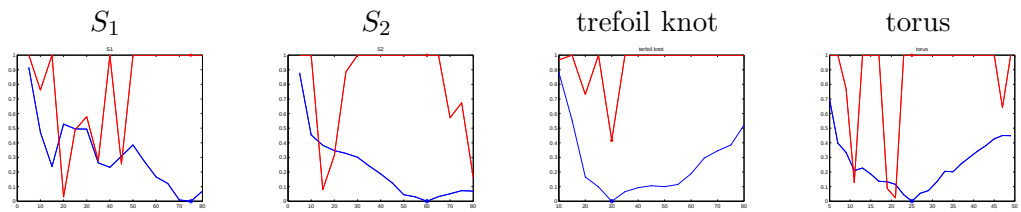

Figure 3: Some examples for support without boundary support. Abscissa: k, blue:$d_{\chi^2}(k)$, red:$\hat{p}_v(k)$.

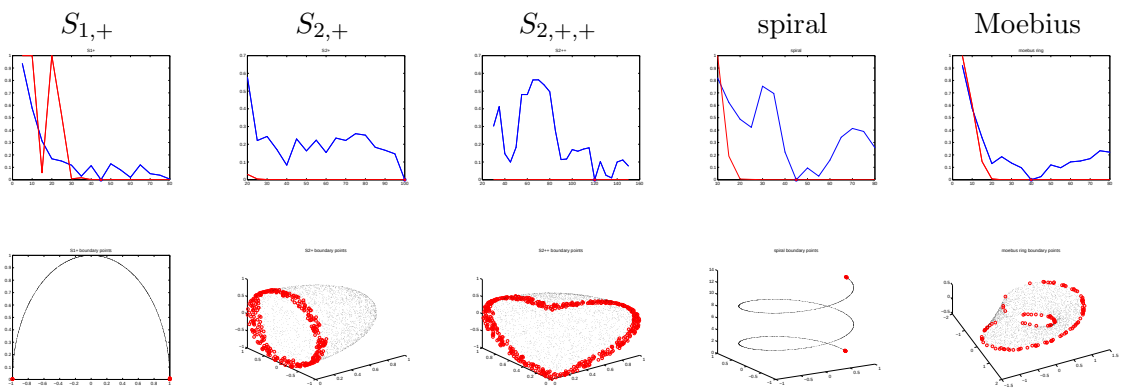

Figure 4: Some examples for support with boundary. First line: Abscissa: k, blue:$d_{\chi^2}(k)$, red:$\hat{p}_v(k)$. Second line: the associated sample and points that are identified as "close to the boundary"

In Figure 5 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 the percent of rejection (i.e. here the level) is less than 5% since $n \geq 500$ for every example associated to a support without boundary which satisfy the hypotheses of our Theorems. When the support has a boundary, the percent of rejection (i.e. here the power) converges

quickly to 100%, even for S_2++ (for which ∂M is not of class $\mathcal{C}^2$). We also present some results when the density is not continuous or when M is not of class $\mathcal{C}^2$ to illustrate the necessity of our hypotheses.

Notice that, to shorten the computational time, we had preliminary chosen the k_n-value by averaging the one obtained with the d_{χ^2} criteria with 50 samples (for each example and each sample size). The selected k_n are presented in the figure.

example		$n=100$	$n=200$	$n=500$	$n=10000$	$=2000$	$n=3000$
S_1	k_n	15	20	20	35	40	40
	% reject	$1,45\%$	$1,15\%$	$1,05\%$	1%	$0,9\%$	$0,85\%$
S_2	k_n	15	17	20	25	30	40
	% reject	3%	$2,55\%$	$1,6\%$	$1,4\%$	$1,35\%$	$1,05\%$
S_3	k_n	6	10	15	17	25	25
	% reject	$1,2\%$	$2,5\%$	$1,9\%$	$1,35\%$	$1,85\%$	$1,15\%$
S_4	k_n	5	5	10	17	17	17
	% reject	$0,75\%$	$0,05\%$	$2,3\%$	$1,15\%$	$3,15\%$	$1,5\%$
S_1+	k_n	15	20	20	35	40	40
	% reject	$89,25\%$	$79,75\%$	$60,7\%$	$97,1\%$	$99,3\%$	$99,05\%$
S_2+	k_n	17	30	30	50	50	50
	% reject	$84,8\%$	100%	100%	100%	100%	100%
S_3+	k_n	6	8	10	15	25	25
	% reject	$2,35\%$	$4,4\%$	$5,55\%$	$34,45\%$	$99,95\%$	$99,95\%$
S_4+	k_n	5	5	10	80	80	80
	% reject	1%	$0,3\%$	$10,8\%$	100%	100%	100%
Trefoil	k_n	8	13	15	25	30	40
Knot	% reject	$4,7\%$	$5,95\%$	$2,4\%$	$2,15\%$	$1,45\%$	$0,8\%$
Spire	k_n	15	202	25	25	40	40
	% reject	$55,5\%$	$81,25\%$	$92,4\%$	$83,9\%$	100%	$99,9\%$
Moebus	k_n	8	13	15	20	40	40
ring	% reject	$12,2\%$	$65,75\%$	$68,75\%$	$98,65\%$	100%	100%
Torus	k_n	8	13	15	17	20	20
	% reject	$5,6\%$	$10,45\%$	5%	$2,65\%$	$1,75\%$	$2,05\%$
S_2++	k_n	17	30	30	50	50	50
	% reject	$99,95\%$	100%	100%	100%	100%	100%
not	k_n	15	17	20	25	30	30
continuous	% reject	$16,8\%$	$14,75\%$	$11,9\%$	$16,25\%$	$17,3\%$	$14,95\%$
square	k_n	10	13	225	30	30	50
(not $\mathcal{C}^2$)	% error	$4,75\%$	$4,5\%$	$5,1\%$	$4,4\%$	$3,25\%$	$9,55\%$

Figure 5: For different samples, the chosen k_n value and the % of times where H_0 is rejected (on 2000 replications).

References

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

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

Berger, M.(2003) A panoramic view of Riemannian geometry. *Springer*

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

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

Cholaquidis, A., Cuevas, A., and Fraiman, R (2014) On Poincaré cone property. *Ann. Statist.* **42**, 255–284.

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

Cuevas, A.; Fraiman, R. and Pateiro-Lopez, B.(2012) On statistical properties of sets fulfilling rolling-type conditions. *Adv. in Appl. Probab.* **44**, 311–239.

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

Delicado, P.; Hernández, A. and Lugosi, G. (2014) Data-based decision rules about the convexity of the support of a distribution. *Electron. J. Statist.* **8**, 96–129.

Fefferman, C., Mitter, S., and Narayanan, H.2013 Testing the manifold hyphotesis.

Fasy, B.T., Lecci, F., Rinaldo, R., Wasserman, L. Balakrishnan, S. and Singh, A. (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42**, 2301–2339.

Genovese, C.; Perone-pacifico, M.; Verdinelli, I. and Wasserman, L.(2012) Minimax Manifold Estimation. *Journal of Machine Learning Research* **13**, 1263–1291

Guillemin, V. and Pollack, A. Differential Topology *Prentice-Hall, Inc., Englewood Cliffs, New Jersey*

Niyogi, P., Smale, S. and Weinberger, S. (2011) A topological view of unsupervised learning from noisy data. *SIAM J. Comput.* 40, no. 3, 646–663.

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.

Pennec, X. (2006). Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements. *Journal of Mathematical Imaging and Vision* **25** 127–154.

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

Ranneby, B. (1984). The maximal spacing method. An estimation method related to maximum likelihood method. *Scand. J. Statist.* **11** 93–112.

Walther, G. (1997) Granulometric smoothing. *Ann. Statist.* **25** 2273–2299