Université Clermont-Auvergne

Laboratoire de Mathématiques Blaise Pascal

Mémoire d'habilitation à diriger les recherches

Some contributions and perspectives in geometric and topological inference

---

Author: Catherine Aaron

Paul Doukhan	*Examinateur*	Université Cergy-Paris
Elisabeth Gassiat	*Présidente*	Université Paris-Saclay
Arnaud Guillin	*Tuteur*	Université Clermont Auvergne
Bertand Michel	*Rapporteur*	Ecole Centrale de Nantes
Bruno Pelletier	*Rapporteur*	Université Rennes II
Alberto Rodríguez-Casal	*Rapporteur*	Universidad de Santiago de Compostela
Anne-Francoise Yao	*Examinatrice*	Université Clermont Auvergne

April 12, 2022

Remerciements

Je souhaite remercier tout particulièrement, Bertrand Michel, Bruno Pelletier et Alberto Rodríguez Casal pour avoir accepter de rapporter ce mémoire. Un merci particulier à Bertrand pour s'être interressé depuis longtemps à mes travaux, m'avoir impliquée dans plusieurs conférences, m'avoir embarquée dans l'aventure IHP et (re)mis en contact avec Eddie et Clément. Un grand merci aussi à Paul Doukhan, Elisabeth Gassiat et Anne-Francoise Yao d'avoir accepté de faire partie du Jury.

Je remercie bien sûr Arnaud Guillin pour avoir accepté d'être mon tuteur mais aussi (et surtout) pour l'aide et la bienveillance dont il a fait preuve quand les choses n'allaient pas si bien.

De manière générale je remercie les membres du LMBP pour supporter mes fréquents "hommages a Backri" et accepter de boire un café suffisament fort pour me reveiller (tâche relativement peu aisée), mais surtout pour leur aide, leur écoute et leur bienveillance. Un merci tout particulier à l'équipe de crise/rapatriement de ces jours pré-Hdr mouvementés...

Pour rester sur le campus, merci aux étudiants qui ont fait de mon metier un plaisir, certains groupes et certains individus m'ont réconfortés dans la beauté de l'enseignement... Merci aussi aux autres de nous fournir un sujet de discussion et d'anecdotes infini : plus de silence génés en soirée grace à vous.

Mes co-auteurs méritent quelques années de congés pour avoir supportée ma petite personne livrée avec un handicap redactionnel flagrant. A tous ce fut vraiment un plaisir de travailler avec vous. On recommence quand vous voulez !

Bien sûr tout n'est pas que travail et je voudrais remercier l'ensemble de mes amis, les parisiens, les toulousains, les clermontois, les vieux, les jeunes, les très jeunes (non petite : travailler n'est pas "s'assoir et rien faire"), les sportifs, les festifs, les joueurs..., ma "fausse famille" de Tursac (le centre du monde) et bien sur ma vraie famille. Merci à tous, votre présence à mes coté est mon bien le plus précieux. Enfin, pour ne pas sombrer dans la sensiblerie, Je tiens à préciser qu'aucun pangolin ni aucune licorne n'a été maltraité durant la rédaction.

Contents

Notations v

1 Introduction 1

2 Full dimensional context (set estimation and related topics) 7
2.1 Introduction and main hypotheses . . . 7
2.2 Support (and boundary) estimation . . . 10
2.2.1 Convexity hypothesis . . . 10
2.2.2 Non convex support estimation . . . 15
2.3 Perspectives in density and level set estimation . . . 21
2.3.1 Density estimation . . . 21
2.4 Estimation of the volume and surface . . . 25
2.4.1 Perspective in Volume estimation . . . 25
2.4.2 Surface estimation . . . 25

3 Lower dimensional context (Manifold learning) 31
3.1 Introduction . . . 31
3.2 Geometric Setting . . . 32
3.2.1 Sub-manifolds with (possible) Boundary . . . 32
3.2.2 Tangent and Normal Structures . . . 33
3.2.3 Geometric Assumptions . . . 33
3.3 True lower dimensional context . . . 35
3.3.1 Testing the true lower dimensional case . . . 36
3.3.2 Testing the boundary existence . . . 36
3.3.3 Boundary estimation and estimation with boundary . . . 38
3.3.4 Volume Estimation . . . 45
3.3.5 Perspectives . . . 46
3.4 Noisy lower dimensional context . . . 47
3.4.1 Estimation of the amount of noise . . . 48
3.4.2 Denoising with use of reflexion on the boundary . . . 49
3.4.3 Denoising with use of the medial axis . . . 49
3.4.4 Perspectives . . . 52
3.5 General perspective: sub-manifold of a known manifold . . . 53

4 Application to statistical learning 55
4.1 Dimension reduction and some statistics in lower dimension . . . 55
4.1.1 Convergences rates for the geodesic distances . . . 56
4.1.2 Perspectives and open questions . . . 57
4.1.3 Application to Frechet mean estimation . . . 58
4.2 Clustering . . . 59

4.3 Classification : SVM with no Kernel Trick 60
4.4 Robust Fusion for big data 61

Bibliography **63**

Notations

1. Throughout, $D \geq 1$ is referred to as the ambient dimension and $\mathbb{R}^D$ is endowed with the Euclidean inner product $\langle \cdot, \cdot \rangle$ and the associated norm $\|\cdot\|$. The closed Euclidean ball of center x and radius r is denoted by $\mathrm{B}(x, r)$, and its open counterpart by $\mathring{\mathrm{B}}(x, r)$. The sphere of center x and radius r is denoted by $\mathcal{S}(x, r)$. Given two vectors u and v there angle $\angle u, v = \sin^{-1}\left(\frac{\langle u,v \rangle}{\|u\|\|v\|}\right)$. If T is a linear subspace then $\mathrm{B}_T(O, r) = B(O, r) \cap T$.

2. Classically for a set S we denote by S^c, $\overline{S}$, $\mathring{S}$ and ∂S its complement, closure, interior and boundary.

3. $\mathcal{H}(S)$ denotes the convex hull of S.

4. $S \oplus \varepsilon B = \{x, d(x, S) \leq \varepsilon\}$, $S \ominus \varepsilon B = \{x, B(x, \varepsilon) \subset S\}$.

5. $d_H(S_1, S_2) = \min\{t, S_1 \subset S_2 \oplus tB \text{ and } S_2 \subset S_1 \oplus tB\}$.

6. $S_1 \Delta S_2 = (S_1 \setminus S_2) \cup (S_2 \setminus S_1)$ and for any measure μ, $d_\mu(S_1, S_2) = |S_1 \Delta S_2|$ is called measure of the symmetric difference.

7. When S is compact $d(x, S) = \min\{||x - s||, s \in S\}$ and, when define $\pi_S(x) = \arg\min\{||x - s||, s \in S\}$.

8. Let $N_S(\varepsilon)$ be the covering number of S by ball of radius ε, when exists the Minkowski dimension is defined by $\mathrm{Dim}_{\mathrm{Mink}} = -\lim_{\varepsilon \to 0} \frac{\ln N(\varepsilon)}{\ln(\varepsilon)}$.

9. Let $E = \{e_1, \dots e_n\} \subset \mathbb{R}^D$ is a finite set $\#E$ denotes its cardinality, for $x \in E$ $\mathrm{Vor}_E(x)$ is the Voronoi cell of x in E i.e. $\mathrm{Vor}_E(x) = \{y \in \mathbb{R}^D, ||y - x|| \leq \min_i ||y - e_i||\}$. In the classical case where we consider the Voronoi cells of points within our set the index will be omitted i.e. $\mathrm{Vor}_E(e_i) = \mathrm{Vor}(e_i)$.

10. Let M be a d-dimensional sub-manifold of $\mathbb{R}^D$, $|M|_d$ denotes its d-dimensional Hausdorff measure on $\mathbb{R}^D$. Here again, when no ambiguity the index will be omitted. $\omega_d = |B_d(0, 1)|_d$ where $B_d(0, 1)$ is the d-dimensional unit ball. When define, $T_x M$ denotes the tangent (to M) linear space at x.

Chapter 1

Introduction

Geometric and topological inference consists in estimating sets S (or quantities related to S such as ∂S, $|S|$, $|\partial S|$, Betti number or homology groups...), and testing geometrical or topological properties based on random samples. In this document we will only focus on inferring knowledge on S from sample points $\mathbb{X}_n \subset \mathbb{R}^D$ drawn according to a distribution supported by S (the unknown set of interest) or "near" S.

There is many way of estimating S. Historically ([Rényi & Slanke 1963] and [Rényi & Slanke 1964]) the first works on support estimation concern the estimation of a convex S with $\mathcal{H}(\mathbb{X}_n)$ the convex hull of the sample (see Figure 1.1). It is a fully data driven method that has minimax convergence rate when the support is convex. This was a motivation to build a convexity test in [J6].

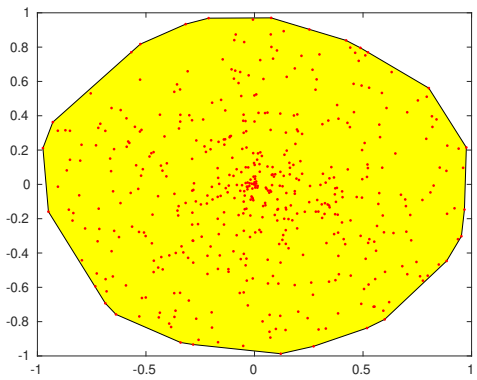

Figure 1.1: The convex hull of a sample drawn on a $B(0,1)$

When S is no longer supposed to be convex, $DW_r(\mathbb{X}_n) = \bigcup_i B(X_i, r)$ (see Figure 1.2 top left graphic), the union of balls centered at the observations has been introduced in [Devroye & Wise 1980] (or [Chevalier 1976]). It has been proved to be universally consistent and has been has been extensively studied. See for instance [Biau *et al.* 2008] and [Biau *et al.* 2009] for more precise results on convergence rates and central limit theorem or [Baíllo *et al.* 2000] for heuristics about tuning the r parameter.

If the Devroye-Wise support estimator $DW_{r_n}(\mathbb{X}_n)$ is universally consistent it has not as good rates as the convex hull of the sample when the support is convex. This was at the origin of various convex hull extensions from [Edelsbrunner *et al.* 1983] where the r-shape $Sh_r(\mathbb{X}_n)$ and the r-convex hull $C_r(\mathbb{X}_n)$ was introduced when $D = 2$. The r-convex hull

(see Figure 1.2 top right graphic) is defined as follows.

$$C_r(\mathbb{X}_n) = \left(\bigcup_{x, \mathring{B}(x,r)\cap\mathbb{X}_n=\emptyset} \mathring{B}(x,r) \right)^c . \tag{1.1}$$

The definition of the r-shape (see Figure 1.2 bottom left graphic) is a bit more complicated, it requires that we preliminary define $\mathcal{D}(\mathbb{X}_n)$ the set of Delaunay simplices $\sigma = \mathcal{H}(X_{i_1}, \dots X_{i_{D+1}})$ such that $\mathring{B}(O_\sigma, r_\sigma) \cap \mathbb{X}_n = \emptyset$ where $\mathcal{S}(O_\sigma, r_\sigma)$ is the circumhypersphere of σ.

$$Sh_r(\mathbb{X}_n) = \bigcup_{\sigma\in\mathcal{D}(\mathbb{X}_n), r_\sigma \le r} \sigma. \tag{1.2}$$

More recently, in [Getz & Wilmers 2004] the Local Convex hull estimator (see Figure 1.2 bottom right graphic)

$$LcH_r(\mathbb{X}_n) = \bigcup_i \mathcal{H}\left(B(X_i, r) \cap \mathbb{X}_n\right), \tag{1.3}$$

was introduced for application in ecography and home range estimation.

These three estimators are generalizations of the convex hull since $Sh_{+\infty}(\mathbb{X}_n) = C_{+\infty}(\mathbb{X}_n) = LcH_{+\infty}(\mathbb{X}_n) = \mathcal{H}(\mathbb{X}_n)$. When the support is "full dimensional", "smooth enough" and the distribution "uniform enough" (see Figure 1.2) this estimators have nice asymptotic properties. See [Rodríguez Casal 2007] for results on $C_r(\mathbb{X}_n)$, [J7] for results on $LcH_r(\mathbb{X}_n)$. Also see [Arias-Castro & Rodríguez-Casal 2017] for the use of $Sh_r(\mathbb{X}_n)$ for perimeter estimation when $D = 2$.

There exists other generalizations of the convex hull as the one introduced in [Cholaquidis *et al.* 2014] or [Cholaquidis & Cuevas 2020] more adapted to non smooth support which is not the purpose of this document.

Chapter 2 is dedicated to the presentation of our main contributions under such a context (full dimensionality, support smooth enough and density uniform enough) which are:

1. a convexity test based on a generalization of the maximal spacing (see [J6]);
2. asymptotic study of the $LcH_r(\mathbb{X}_n)$ estimator (see [J7]) that is proved to have minimax rates and topological guarantees;
3. additional results on the $C_r(\mathbb{X}_n)$ estimator whose boundary is proved to be homeomorphic to the boundary of S (with probability one for n large enough) and Estimation of $|\partial S|_{D-1}$ (see [P2]).

In this chapter we also present perspectives on density estimation, and on the use of Sh.

When S is a smooth enough d-dimensional manifold with $d < D$, the Devroye wise estimator widely "overestimates" S (see Figure 1.3 top left graphic) but allows the homology recognition see [Niyogi *et al.* 2008]. To overcome the overestimation one can imagine to replace the full dimensional balls $B(X_i, r_n)$ of the Devroye wise estimator by "patches" $B_{X_i+\hat{T}_i}(X_i, r_n)$ where $\hat{T}_i$ is an estimator of $T_{X_i}S$ (see Figure 1.3 top right graphic). In [Aamari & Levrard 2019] it is proven to have minimax rates when S is a manifold without

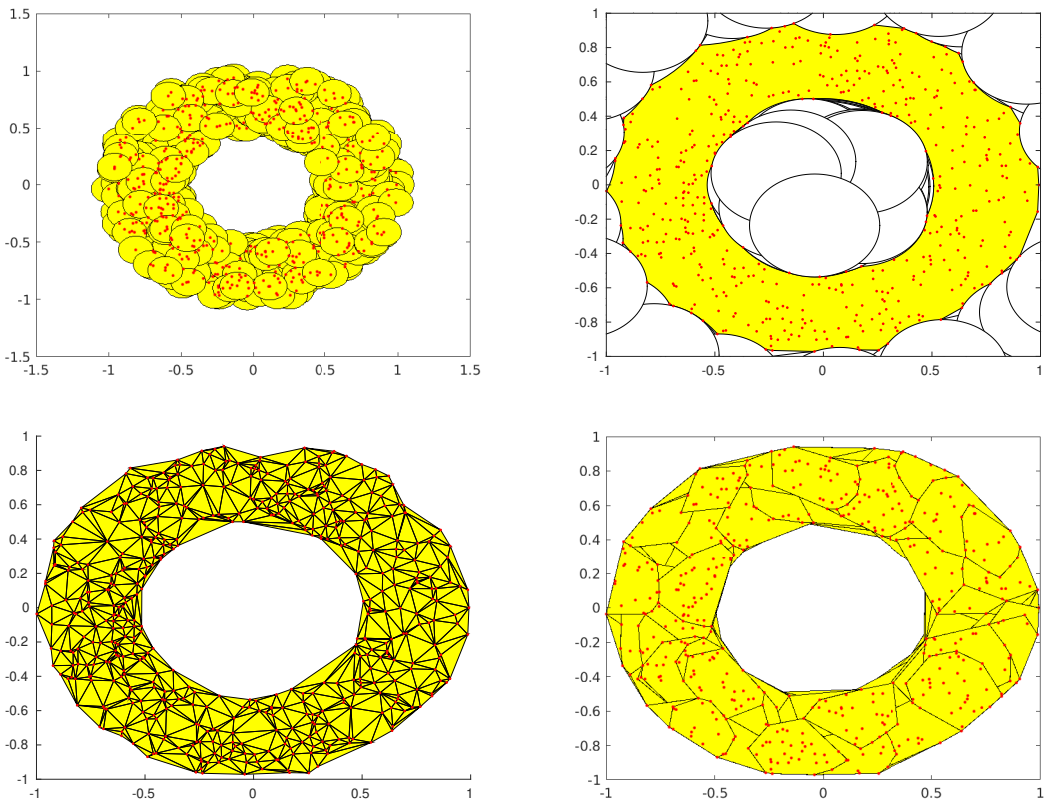

Figure 1.2: 4 support estimators computed on the same sample of size 500 drawn on $B(0,1) \setminus B(0,.5)$. Top and Left: DW. Top and right: C_r. Bottom and Left: Sh_r. Bottom and right: LcH_r.

boundary. The $LcH_r(\mathbb{X}_n)$ (see Figure 1.3 bottom left graphic) seems to have a nice behavior. In [Divol 2020] an estimator close to LcH_r is proven to have minimax rates when S has no boundary. The r-hull and the r-shape have degenerated behavior. Indeed, when r is small enough with regard to the smoothness of S we have that $C_r(\mathbb{X}_n) = \mathbb{X}_n$ and $Sh_r(\mathbb{X}_n) = \emptyset$. A generalization of $Sh_r(\mathbb{X}_n)$ (see Figure 1.3 bottom right graphic) has been proven to have minimax rates and nice topological properties in [Aamari & Levrard 2018].

Our main contributions under such hypotheses consist in extension to the case where the boundary may not be empty that allows an unification of the two main settings in geometric inference : the full dimensional case and the lower dimensional without boundary case. A first part of Chapter 3 concerns the geometric setting of manifold with boundary, then in a second part we present our main results that are:

1. a test of the lower dimensional setting (see [J5]);
2. a test for the boundary emptiness (see [J1]);
3. minimax estimation of the boundary and minimax estimation of the manifold when

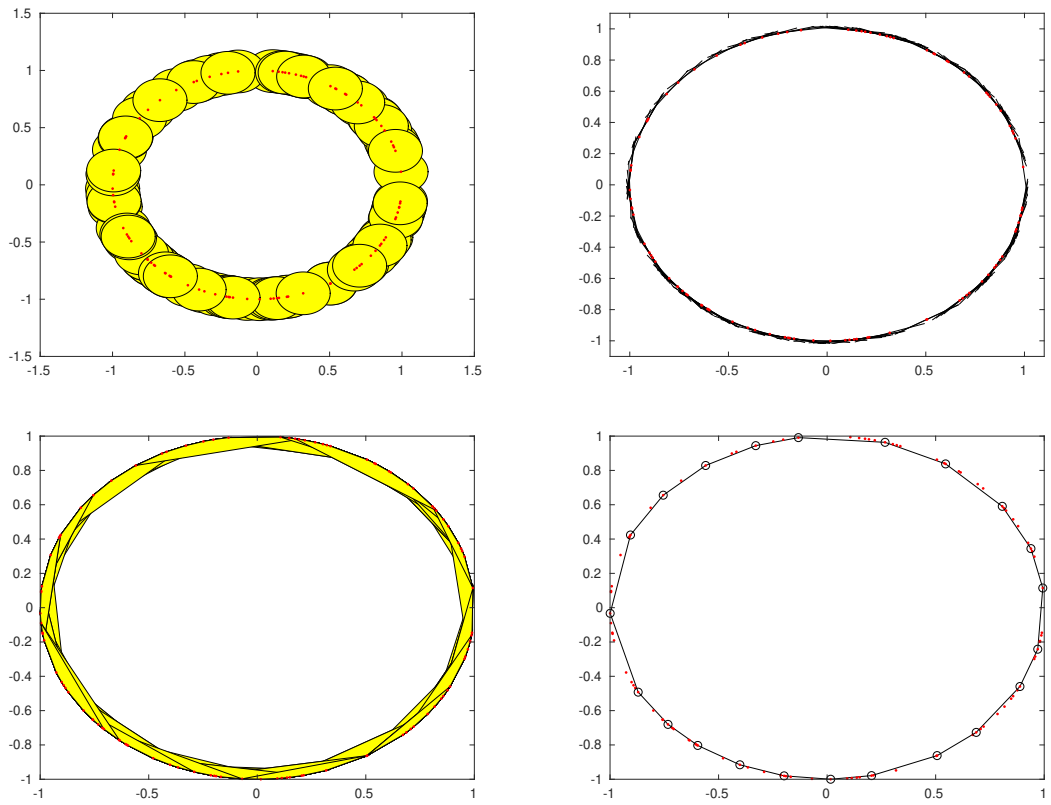

Figure 1.3: 4 support estimators computed on the same sample of size 100 drawn on $\mathcal{S}(0,1)$. Top and Left: DW. Top and right the Patch estimator. Bottom and Left LcH_r. Bottom and right: the generalization of the r shape (that needs a sparsification of the sample).

non-empty boundary (see [P1]);

4. estimation of $|S|_d$ (see [J5]).

In a third part of this chapter we will focus on the "noisy lower dimensional" case where the support of the distribution is included in a tubular neighborhood of the manifold and we aim to get informations on the manifold (see Figure 1.4).

If the "amount of noise" (i.e. the radius of the tubular neighborhood) is vanishing (that is depends on the sample size and converges toward 0) most of the aforementioned methods in the "true lower dimensional" case are still consistent (with a possibly deteriorate rate). We are mostly interested in the case of a "fixed" amount of noise (see [Genovese *et al.* 2012b] for the minimax rate and [Genovese *et al.* 2012a] for the case of 1-dimensional manifolds). In [J5] we propose two estimators for the amount of noise and a "denoising method" (i.e. a method that have, as output, data in a vanishing tubular neighborhood of the manifold) . In [J2] another "denoising method", based on the medial axis estimation (as in [Genovese *et al.* 2012a]) is proposed.

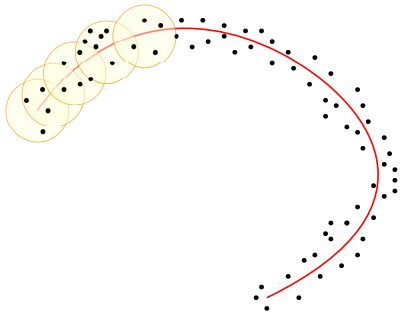

Figure 1.4: In the noisy lower dimensional context, the data (here the black dots) are at a small distance to an unknown manifold (red line) that we aim to estimate.

A final chapter (Chapter 4) concerns the possible applications to data analysis and statistical learning which are various (see [Tenenbaum *et al.* 2000, Srivastava *et al.* 2008, Singh *et al.* 2007a, Singh *et al.* 2007b, Kohonen 2004, Chaudhuri & Dasgupta 2010]). First we detail the link between geometric inference and dimension reduction, giving a list of applications of the tools developed in Chapters 2 and 3. Then we present results of [J4] where convergence rates for geodesic distance estimation (used as input of many non linear dimension reduction method) are derived (and applied to estimation of Frechet moments on a unknown manifold). We then provide a discussion on how aforementioned tools can apply to Clustering and Classification. If clustering application is well known, the classification is presented as a perspective. We also present some results of [J3], which is less linked with geometric and topological problems.

Chapter 2

Full dimensional context (set estimation and related topics)

2.1 Introduction and main hypotheses

Let $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ be an iid sample of some random variable with (unknown) distribution $\mathbb{P}_X$ supported by $S = \overline{\{x \in \mathbb{R}^D \text{ such that } f_X(x) > 0\}}$ which is the smallest, for the inclusion, closed set such that $\mathbb{P}_X(S) = 1$. Assume that $\mathbb{P}_X$ is absolutely continuous with respect to the Lebesgue Measure and denote by f_X its density. Under such an assumption it comes that we have that $\overline{\mathring{S}} = S$ also classically known as **regularity of the support** which is the **full dimensional** context.

In this chapter we are interested in estimating S, ∂S and some of their functionals such as $|S|_D$ its volume and $|S|_{D-1}$ its surface area. We also aim at estimating f_X and the level sets $L_\lambda = \overline{\{x \in \mathbb{R}^D \text{ such that } f_X(x) > \lambda\}}$.

With regard to set estimation, performance of an estimator $\hat{S}_n$ is usually evaluated through the Hausdorff distance or through the measure of the symmetric difference. This two ways are not equivalent and have both advantages and drawbacks.

1. Suppose the distribution is the uniform one on a compact set S and that $\hat{S}_n = \mathbb{X}_n$ we have $d_H(S, \hat{S}_n) \to 0$ while $d_\mu(S, \hat{S}_n) = \mu(S)$

2. Suppose that the distribution is the standard normal on $\mathbb{R}^2$ and that $\hat{S}_n = B(O, r_n)$ with $r_n \to +\infty$ we have $d_H(S, \hat{S}_n) = +\infty$ while $d_\mu(S, \hat{S}_n) \to 0$

Because d_H and d_μ are not fully satisfying we also propose to take into account the boundary of the estimator (see [Cuevas & Rodríguez-Casal 2004] and [Rodríguez Casal 2007]). Indeed having $d_H(S, S_n) \leq \varepsilon_n$ and $d_H(\partial S, \partial S_n) \leq \varepsilon_n$ we have that $S \Delta S_n \subset \partial S \oplus d_H(\partial S, \partial S_n)$ which is a more precise information on the support estimator asymptotic behavior. We also expect that the proposed estimator catch the topology of the unknown support.

Considering the assumptions on S, the support is usually supposed compact. The case where the support is compact, convex and estimated by the convex hull of the sample has been extensively studied (see, for instance the list being far from exhaustive [Efron 1965], [Schneider 1988], [Dümbgen & Walther 1996], [Bárány 1992], [Brunel 2013], [Brunel 2020], [Beermann & Reitzner 2015]). It thus seems interesting to have a statistical test allowing to decide whether the support is convex of not that was the aim of [J6] which is summurized in section 2.2.1.2.

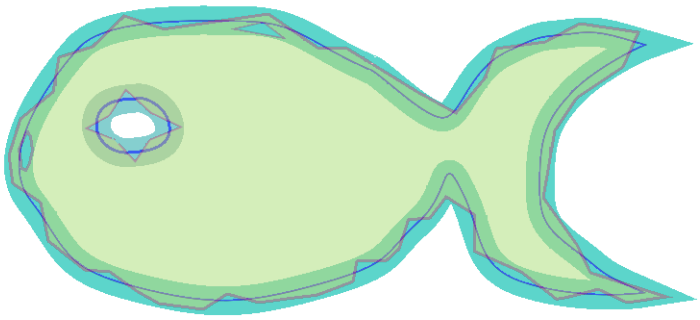

Figure 2.1: ∂F The boundary of a "fish-shape" set (blue line), $\partial F \oplus \varepsilon B$, and an estimator satisfying $d_H(\hat{F}, F)$ is "small" and $\partial \hat{F} \subset \partial F \oplus \varepsilon B$.

Under convexity assumption, we will mainly refer to [Dümbgen & Walther 1996] where it is proved that, when the density is uniform (which easily extends to the case of bounded bellow by a positive constant) we have $d_H(S, \hat{S}_n) = O((\ln n/n)^{1/D})$ and, additionally assuming that the boundary is $\mathcal{C}_1^1$ we have that $d_H(S, \hat{S}_n) = O((\ln n/n)^{2/(D+1)})$. We aim at obtaining support estimators having similar rates with less restrictive shape hypothesis than convexity.

First notice that, in the full dimensional context convexity implies that there exists $\delta > 0$ such that $|B(x, r) \cap S|_D \geq \delta \omega_D r^D$ for small enough radius r. And we will soften the convexity hypothesis by this less restrictive hypothesis, also known as standardness and first introduced in [Cuevas 1990].

Definition 1 (standardness)**.** *A regular set S ($\overline{\overset{\circ}{S}} = S$) is said to be δ-standard ($\delta > 0$) if there exists $r_0 > 0$ such that, for all $r \leq r_0$ and all $x \in S$ $|B(x, r) \cap S|_D \geq \delta \omega_D r^D$*

It is sometime useful to use a lightly stronger but morally similar condition. See Figure 2.2.

Definition 2 (Ball standardness)**.** *A regular set S ($\overline{\overset{\circ}{S}} = S$) is said inside (resp. outside) (ε_0, δ)-ball standard if for all $x \in S$ (resp. S^c) and all $\varepsilon \leq \varepsilon_0$ there exists y_{in} (resp. y_{out}) such that $x \in B(y_{in}, \varepsilon)$ and $B(y_{in}, \delta\varepsilon) \subset S$ (resp. $x \in B(y_{out}, \varepsilon)$ and $B(y_{out}, \delta\varepsilon) \subset \overline{S^c}$).*

Inside and outside (ε_0, δ)-ball standardness is sumurized in (ε_0, δ)-ball standardness.

To generalize the results on the $\mathcal{C}_1^1$ convex hull we will see that the regularity of the boundary is the only important condition there. Due to [Walther 1999] this condition can also been expressed as a "rolling ball" condition (see Figure 2.2) which allows to obtain results applying classical Euclidean geometry and not having any differential or Riemannian calculus.

Definition 3 (Inside and outside rolling ball condition)**.** *A regular set S ($\overline{\overset{\circ}{S}} = S$) is said to satisfy the r_0 inside and outside rolling ball conditions if, for all $x \in \partial S$ there exists O_{in} and O_{out} such that*

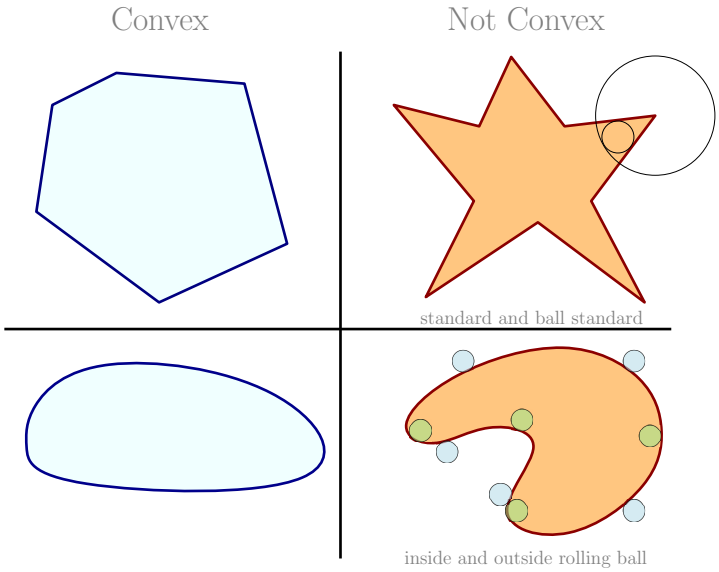

Figure 2.2: Ball standarness generalizes the convex polytope, while the rolling ball condition generalizes the smooth convex case.

1. $||x - O_{in}|| = r_0$ *and* $B(O_{in}, r_0) \subset \overline{S}$
2. $||x - O_{out}|| = r_0$ *and* $B(O_{out}, r_0) \subset \overline{S^c}$

Together with the shape hypothesis we will require some distribution hypothesis. Morally we need that the density decreases quickly enough when going closer to the boundary. The strongest and most currently used hypothesis being that f is bounded bellow by a positive constant on S also denoted as f is "almost uniform" that is a particular case of α-quickly decreasing densities defined as follows.

Definition 4 (α-quickly decreasing)**.** *The density f_X is said to be α-quickly decreasing if, there exists $f_0 > 0$ such that, for all $x \in S$, $f(x) \geq f_0 d(x, \partial S)^\alpha$.*

If f_X is 0-quickly decreasing, f_X is said to be ***almost uniform****.*

Originally the standardness, firstly introduced in [Cuevas 1990], was mixing geometric and distribution hypothesis

Definition 5 (Standard distribution)**.** *A distribution $\mathbb{P}$ supported by S is δ-standard with regard to a measure μ if there exists $r_0 > 0$ such that for all $r \leq r_0$, and all $x \in S$* $\mathbb{P}(B(x, r)) \geq \delta \mu(B(x, r))$

Notice that we clearly have that either standardness or ball standardness for the support together with almost uniformity for the distribution imply standardness of the distribution.

Section 2.2.2 is dedicated to present some support estimators and there properties and convergence rates under some of the aforementioned hypotheses.

Under our most common hypothesis of compact support and almost uniform density, kernel density estimation is biased for point at the boundary. This may perturb the Level set estimation. Existing bias correction methods (see [Funke & Kawka 2015a], [Charpentier & Gallic 2015], [Funke & Kawka 2015b], [Jones *et al.* 1995], [Karunamuni & Zhang 2008], [Leblanc 2010], [Marron & Ruppert 1994] or [Ruppert & Cline 1994]), are based on the knowledge of the support but what can we do in case of unknown support? Obviously plug-in of support estimates seems a reasonable proposal for a theoretic point of view, unfortunately that does not seem really computationally feasible. In section 2.3.1 we introduce a new density estimator allowing to correct the bias, when unknown support and with easy computation. This is a "work in progress" but the proof is really short and it opens a wide field of development wich is presented presented as a perspective.

Section 2.3 at the end of this chapter is dedicated to estimation of the volume and surface area. If there exists yet (see [Arias-Castro *et al.* 2019]) minimax volume estimator under uniformity assumption one can expect to generalize it to non uniform distribution (which is far from obvious). The surface area estimation in any dimension is far from minimax but, up to our knowledge, [P2] is the first work about it when observing only points in S, it is summarize in section 2.3.2.

2.2 Support (and boundary) estimation

2.2.1 Convexity hypothesis

2.2.1.1 Inference under convexity hypothesis

As previously mentioned, the convex hull of a sample drawn on a convex set has been extensively studied. We present here a small time-line of a selection of some results on the convex hull of random points.

1965 [Efron 1965] gives integral expression for expected values of interest (such as the number of vertex and the probability contents) for dimension $D = 2$ or $D = 3$.

1988-1994 As mentioned in the introduction for uniform samples in [Schneider 1988] convergence rates for the convex hull are given, the asymptotic behavior being improved in [Bárány 1992] then [Schutt 1993].

1998 In [Bräker & Hsing 1998] estimation of the perimeter and area of a convex sate is studied.

2003 In [Reitzner 2003] Reitzner derive strong law of large number for the volume and of the number of vertices of the convex hull of the sample

2015 In [Baldin & Reiß 2016] an unbiased estimator of the volume of the convex hull (with a correction with the number of vertices) is proposed in the Poisson point process setting.

2017 In [Brunel 2020] probabilistic bounds for the probabilistic contents of convex hull of sets is given in a very general setting (improving [Brunel 2013] results derived under uniformity hypothesis).

Also mention that when the support is supposed to be a convex polytope it can be estimate with a parametric rate as proved in [Brunel 2016a].

2.2.1.2 Convexity test

With regard to all the existing results about the convex hull as a support estimator when the support is convex it appears convenient to have a statistical test allowing to decide whether the support is convex or not. A first test was proposed in [Delicado *et al.* 2014] considering the following test statistics

$$T = \max_{i,j} \left\| \frac{X_i + X_j}{2}, \mathbb{X}_n \right\|$$

that provides a consistent decision rules with a Monte-Carlo calibration of the threshold of the decision rule.

With Alejandro Cholaquidis and Ricardo Fraiman we proposed another way of testing the convexity of the support based on the maximal spacing.

The maximal spacing and Janson's results For ease of reading we are going to present here a lightly modified version of the maximal spacing introduced by Janson [Janson 1987] that take into account our extension but that is limited to the size of balls (instead of a general convex) that do not contain any observations.

Definition 6. *Let $\mathbb{X}_n = \{X_1, \dots, X_n\}$ be an iid random sample of points in $\mathbb{R}^D$, drawn according to a density f_X with bounded support S. We define*

$$\Delta(\mathbb{X}_n) = \sup \left\{ r : \ \exists x \text{ such that } B\left(x, \frac{r}{(f_X(x)\omega_D)^{1/D}}\right) \subset S \setminus \mathbb{X}_n \right\},$$

$$V(\mathbb{X}_n) = \Delta^D(\mathbb{X}_n),$$

and

$$U(\mathbb{X}_n) = nV(\mathbb{X}_n) - \log(n) - (D-1)\log\big(\log(n)\big) - \log(\alpha_D).$$

where

$$\alpha_D = \frac{1}{D!}\left(\frac{\sqrt{\pi}\Gamma\left(\frac{D}{2}+1\right)}{\Gamma\left(\frac{D+1}{2}\right)}\right)^{D-1}.$$

In this section, U is a random variable such that $\mathbb{P}(U \le t) = \exp\big(-\exp(-t)\big)$

The following result can be found in [Janson 1987].

Theorem 1 (Jansen 87). *Let $S \subset \mathbb{R}^D$ be a bounded set with $|\partial S|_D = 0$ Let $\mathbb{X}_n = \{X_1, \dots, X_n\}$ be iid random vectors **uniformly** distributed on S. Then,*

i)

$$U(\mathbb{X}_n) \xrightarrow{\mathcal{L}} U \quad \textit{when } n \to \infty,$$

ii)

$$\liminf_{n\to+\infty} \frac{nV(\mathbb{X}_n) - \log(n)}{\log(\log(n))} = D - 1 \ \textit{a.s.},$$

iii)

$$\limsup_{n\to+\infty} \frac{nV(\mathbb{X}_n) - \log(n)}{\log(\log(n))} = D + 1 \ \textit{a.s.}$$

From constant to Holder continuous densities Our extension (see [J6]) consists in having a similar result for Holder continuous densities (instead of constant ones).

Theorem 2 (A., Cholaquidis and Fraiman 17)**.** *Let $S \subset \mathbb{R}^D$ be a bounded set that has a boundary which is regular enough. Assume that $Dim_{MinK}(\partial S) < D$. Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be iid random vectors distributed on S with an almost uniform density f_X whose restriction on S is **Holder continuous**. Then,*

i)

$$U(\mathbb{X}_n) \xrightarrow{\mathcal{L}} U \quad \textit{when } n \to \infty,$$

ii)

$$\liminf_{n\to+\infty} \frac{nV(\mathbb{X}_n) - \log(n)}{\log(\log(n))} = D - 1 \ \textit{a.s.},$$

iii)

$$\limsup_{n\to+\infty} \frac{nV(\mathbb{X}_n) - \log(n)}{\log(\log(n))} = D + 1 \ \textit{a.s.}$$

The proof of this theorem is only technical with approximation of a Holder continuous density with piece-wise constant densities, application of Jansen's theorem on any constant piece and argument that allows to neglect boundary effects.

This theorem has the following Corollary that is more realistically useful because it allows to use density estimators instead of the real density (commonly unknown).

Corollary 1 (A., Cholaquidis and Fraiman 17)**.** *Let $S \subset \mathbb{R}^D$ be a bounded set that has a boundary which is regular enough also assume that $Dim_{MinK}(\partial S) < D$. Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be iid random vectors distributed on S with an almost uniform density f_X whose restriction on S is Holder continuous. Suppose that we have $\hat{f}_n$ an estimation of f that satisfies $(f(x)/\hat{f}_n(x)) \geq 1 - \varepsilon_n$ for all $x \in S$ define with $(\ln n)\varepsilon_n \to 0$*

$$\hat{\Delta}(\mathbb{X}_n) = \sup\left\{r : \ \exists x \textit{ such that } B\left(x, \frac{r}{(\hat{f}_n(x)\omega_D)^{1/D}}\right) \subset S \setminus \mathbb{X}_n\right\},$$

$$\hat{V}(\mathbb{X}_n) = \hat{\Delta}^D(\mathbb{X}_n),$$

and

$$\hat{U}(\mathbb{X}_n) = n\hat{\Delta}^D(\mathbb{X}_n) - \log(n) - (D-1)\log\big(\log(n)\big) - \log(\alpha_D).$$

Then $\mathbb{P}(\hat{U}(\mathbb{X}_n) \geq t) \leq 1 - \exp(-\exp(-t)) + o(1)$

The proof first consists in noticing that we have that, since,

$$\frac{r}{(\hat{f}_n(x)\omega_D)^{1/D}} = \frac{r}{(f(x)\omega_D)^{1/D}} \left(\frac{f(x)}{\hat{f}_n(x)}\right)^{1/D} \geq \frac{r(1-\varepsilon_n)^{1/D}}{(f(x)\omega_D)^{1/D}}$$

then:

$$B\left(x, \frac{r(1-\varepsilon_n)^{1/D}}{(f(x)\omega_D)^{1/D}}\right) \subset B\left(x, \frac{r}{(\hat{f}_n(x)\omega_D)^{1/D}}\right).$$

Thus, if $B\left(x, \frac{r}{(\hat{f}_n(x)\omega_D)^{1/D}}\right) \subset S \setminus \mathbb{X}_n$ then $B\left(x, \frac{r(1-\varepsilon_n)^{1/D}}{(f(x)\omega_D)^{1/D}}\right) \subset S \setminus \mathbb{X}_n$, leading to :

$$\frac{\Delta_n(\mathbb{X})}{(1-\varepsilon_n)^{1/D}} \geq \hat{\Delta}(\mathbb{X}_n).$$

Thus $\hat{U}(\mathbb{X}_n) \leq U(\mathbb{X}_n) + \frac{\varepsilon_n}{1-\varepsilon_n} nV(\mathbb{X}_n)$. To conclude the proof, by application of points *ii)* and *iii)* in Theorem 2 $\frac{\varepsilon_n}{1-\varepsilon_n} nV(\mathbb{X}_n) \xrightarrow{a.s.} 0$.

Possible improvement: Under the hypothesis on f_X the common density estimators may underestimate the density near the boundary, that is taken into account in this corollary and is not a problem for the forthcoming convexity test. If we apply density estimators that uniformly converges on the all support we may obtain a better result with $\hat{U}(\mathbb{X}_n) \xrightarrow{\mathcal{L}} U$.

The associated convexity test The maximal spacing based convexity test lies on a tricky density estimator. Namely, let $\tilde{f}_n$ be a classical kernel density estimator with a kernel and a windows size satisfying the hypothesis of [Giné & Guillou 2002] so that we can apply there uniform convergence result and then previous corollary. In particular we require that the kernel belongs to the set $\mathcal{K}$ defined as follows according to [Giné & Guillou 2002].

Definition 7. *Let $\mathcal{K}$ be the set of kernel functions $K(u) = \phi(p(u))$, where p is a polynomial and ϕ is a bounded real function of bounded variation, such that $c_K = \int \|u\| K(u)du < \infty$, $K \geq 0$ and there exists r_K and $c'_K > 0$ such that $K(x) \geq c'_K$ for all $x \in \mathcal{B}(0, r_K)$.*

Note that, for example, the Gaussian and the uniform kernel are in $\mathcal{K}$.

Based on $\tilde{f}_n = \frac{1}{nh_n^D} \sum_{i=1}^n K\left(\frac{\|x-X_i\|}{h_n}\right)$ with $K \in \mathcal{K}$, we propose to estimate the density with $\hat{f}_n$ defined as follows:

$$\hat{f}_n(x) = \max_{i, x \in \mathrm{Vor}(X_i)} \tilde{f}_n(X_i) \mathbb{I}_{\mathcal{H}(\mathbb{X}_n)}(x) \tag{2.1}$$

We proposed to use the test statistics $\hat{\hat{U}}(\mathbb{X}_n)$ defined as follows.

$$\hat{\hat{\Delta}}(\mathbb{X}_n) = \sup\left\{r : \ \exists x \text{ such that } B\left(x, \frac{r}{(\hat{f}_n(x)\omega_D)^{1/D}}\right) \subset \mathcal{H}(\mathbb{X}_n) \setminus \mathbb{X}_n\right\},$$

$$\hat{\hat{V}}(\mathbb{X}_n) = \hat{\hat{\Delta}}^D(\mathbb{X}_n),$$

and

$$\hat{\hat{U}}(\mathbb{X}_n) = n\hat{\hat{V}}(\mathbb{X}_n) - \log(n) - (D-1)\log\big(\log(n)\big) - \log(\alpha_D).$$

Such a method will allow to test the convexity with an upper bound on the level and a power 1 for n large enough. More precisely

Theorem 3 (A., Cholaquidis and Fraiman 17)**.** *Let $\hat{f}_n$ be as defined in (2.1) based on $\tilde{f}_n$ a kernel density estimator with a kernel $K \in \mathcal{K}$. Assume that $h_n = \mathcal{O}(n^{-\beta})$ for some $0 < \beta < 1/D$. Assume also that the unknown density is almost uniform and is Lipschitz continuous on S.*

For the following decision problem,

$$\begin{cases} H_0: & S \text{ is convex} \\ H_1: & S \text{ is not convex}, \end{cases} \tag{2.2}$$

a) the test based on the statistic $\hat{\hat{V}}_n$ with critical region $RC = \{\hat{\hat{V}}_n \geq c_{n,\gamma}\}$, where

$$c_{n,\gamma} = \frac{1}{n}\Big(-\log(-\log(1-\gamma)) + \log(n) + (D-1)\log(\log(n)) + \log(\alpha_D)\Big),$$

has an asymptotic level less than γ.

b) Moreover, if S not convex but is ball standard, the power is 1 for sufficiently large n.

Roughly speaking.

1. If S is convex then
   (a) Because of $\mathcal{H}(\mathbb{X}_n) \subset S$, the continuity of f, the compactness of S and the almost uniform hypothesis of f we have $\ln n \max_{x\in\mathcal{H}(\mathbb{X}_n)} |\hat{f}_n(x) - \tilde{f}_n(x)| \to 0$.
   (b) Due to application of [Giné & Guillou 2002] we have that $\max_S |\tilde{f}_n(x) - \mathbb{E}(\tilde{f}_n(x))| \leq \varepsilon_n$ with $(\ln n)\varepsilon_n' \to 0$ Thus we have, for all $x \in \mathcal{H}(\mathbb{X}_n) \subset S$, $\frac{f(x)}{2} - \varepsilon_n' \leq \tilde{f}(x) \leq f(x) + \varepsilon_n'$ and using the almost uniform hypothesis we then can obtain $(f(x)/\hat{f}_n(x)) \geq 1 - \varepsilon_n$ with $\ln n\varepsilon_n \to 0$.
   (c) Apply 1 to obtain the upper (asymptotic) bound on $\hat{\Delta}$ (using $\tilde{f}$ as an estimator) and remark that $\hat{\hat{\Delta}} \leq \hat{\Delta}$ to bound the level.
2. If S is not convex
   (a) Topological arguments (because S is closed) allows to have the existence of a $B(x_0, r_0) \subset \mathcal{H}(S) \setminus S$.
   (b) Due to standardness of S and because we ask the kernel to be bounded above there exist a λ_0 such that, for all i we have $\tilde{f}(X_i) \geq \lambda_0$
   (c) Thus we have $\hat{\hat{V}}(\mathbb{X}_n) \geq \omega_D \lambda_0 r_0^D + o(1)$ that is sufficient to get a power one for n large enough.

2.2.2 Non convex support estimation

Now we have a statistical rule allowing to decide whether S is convex or not. One can aim in finding convenient estimators S (and its boundary) when non convexity is inferred.

2.2.2.1 Support estimation

Minimax rates. Concerning the minimax rates we can naturally expect that the minimax rates are the same convergence rates than the one obtained when estimating a convex support with the convex hull of the observation. This has been proved in [Härdle *et al.* 1995] when $D = 2$

Theorem 4 (Hardle, Park and Tsybakov. 1995)**.** *When $D = 2$, if the boundary of the support is $\mathcal{C}^r$ and the density is α-quickly decreasing then the minimax convergence rate for support estimation is $n^{-\frac{r}{1+(\alpha+1)r}}$*

In higher dimension the minimax rate in support estimation can easily be conjectured to be at least the rate obtained for the convex hull of the sample when the support is convex. This rate is $n^{-2/(D+1)}$ when the density is almost uniform and ∂S is $\mathcal{C}^1_1$ (see [Dümbgen & Walther 1996]). Surprisingly, up to our knowledge there is no explicit proof that it is indeed the minimax rate before [P1] (see Theorems 3.11, 3.12, 3.14 and 3.15 for $d = D$).

Theorem 5 (Aamari, A. and Levrard 2021)**.** *If ∂S is $\mathcal{C}^2$ and the density is almost uniform then, up to a logarithm factor, the minimax rate for manifold and boundary estimation is $(\ln n/n)^{\frac{2}{D+1}}$*

Let us now give a (non exhaustive) list of support estimator and (some of) there properties.

Devroye-Wise Estimator The first support estimator see [Devroye & Wise 1980] (or [Chevalier 1976]) which may be the most intuitive one consists in a union of balls centered at the observations

$$DW_r(\mathbb{X}_n) = \bigcup_i B(X_i, r)$$

It has first been proved that it provides a universally consistent estimator with regard to the measure of the symmetric difference (note that in the following theorem the compactness of the support is not required).

Theorem 6 (Devroye and Wise (1980))**.** *when $r_n \to 0$ and $nr_n^D \to +\infty$ then*

$$d_{\mathbb{P}_X}(DW_{r_n}\Delta S) \xrightarrow{a.s.} 0$$

Considering the Hausdorff distance and the boundary estimation (see [Cuevas & Rodríguez-Casal 2004]), the Devroye-Wise estimator provides an universally consistent estimator.

Theorem 7 (Cuevas and Rodriguez-Casal (2004)). *If S is compact then $d_H(\mathbb{X}_n, S) \xrightarrow{a.s.} 0$ and for any sequence r_n such that $r_n \to 0$ and $r_n \geq d_H(\mathbb{X}_n, S)$ (e.a.s) we have*

$$d_H(DW_{r_n}, S) \xrightarrow{a.s.} 0 \text{ and } d_H(\partial DW_{r_n}, \partial S) \xrightarrow{a.s.} 0.$$

Moreover if the distribution is standard and if there exit positive constants C and ε_0 such that for all $\varepsilon \leq \varepsilon_0$ we have $d_H(\partial S, \partial(S \oplus \varepsilon B)) \leq C\varepsilon$, choosing then $r_n = a(\ln n/n)^{1/d}$ for $a \geq a_0$ (an explicit constant depending on the regularity of the model) we have

$$d_H(DW_{r_n}, S) = O\left(\frac{\ln n}{n}\right)^{1/D} \text{ e.a.s. and } d_H(\partial DW_{r_n}, \partial S) = O\left(\frac{\ln n}{n}\right)^{1/D} \text{ e.a.s.}$$

Roughly speaking the Devroye-Wise estimator generalizes the "convex hull with corner" results under a similar hypothesis.

Many other results have been obtained on the Devroye Wise support estimator, see for instance [Biau *et al.* 2008] and [Biau *et al.* 2009] for more precise results on convergence rates and central limit theorem or [Baíllo *et al.* 2000] for heuristics about tuning the r_n parameter.

Note that, if it is very easy (computationally) to decide whether a point belong to $DW_r(\mathbb{X}_n)$ the computation of $\partial DW_r(\mathbb{X}_n)$ is hard and the convergence rates are far from optimal when additional regularity hypothesis are assumed. Concerning the topological preservation, we are deeply convinced that it is realized e.a.s. for well chosen sequence of radius (as in previous Theorem), but this result has not yet been proved.

To obtain better convergence rates we have to get inspired on the convex case and generalize it. A first generalization of the convex hull is the r-convex hull.

Definition 8. *Let E be a set, its r-convex hull is*

$$C_r(E) = \left(\bigcup_{x, \mathring{B}(x,r) \cap E = \emptyset} \mathring{B}(x, r)\right)^c$$

E is said to be r-convex if $C_r(E) = E$.

When the support S satisfies the r_0-inside and outside rolling ball condition then S is r-convex for any $r \leq r_0$. We then can estimate it with the r-convex hull of the observations see [Rodríguez Casal 2007].

Theorem 8 (Rodriguez Casal 2007). *If the support S satisfies the r_0-inside and outside rolling ball condition and the density is almost uniform we have that, for all $r \leq r_0$ $d_H(C_r, S) = O\left(\frac{\ln n}{n}\right)^{2/(D+1)}$ e.a.s. and $d_H(\partial C_r, \partial S) = O\left(\frac{\ln n}{n}\right)^{2/(D+1)}$ e.a.s..*

About the parameter selection r, a fully data driven procedure based on the maximal spacing, under uniformity assumption, has been proposed in [Rodríguez-Casal, A. & Saavedra-Nieves, P. 2016]. The method had been generalized to Holder continuous density using our maximal spacing generalization presented in previous section in [Rodríguez-Casal & Saavedra-Nieves. 2019a].

Concerning the topological preservation we obtained a new result has been set the boundary of $\partial C_r(\mathbb{X}_n)$.

Theorem 9 (A., Cholaquidis and Fraiman 21). *If the support S satisfies the r_0-inside and outside rolling ball condition, $\mathcal{C}^2$ boundary and the density is almost uniform we have that, for all $r < r_0$ we have that $\partial C_r(\mathbb{X}_n) \approx \partial S$ e.a.s.*

Sketch of proof. First according to $\partial C_r(\mathbb{X})$ is a finite union of portion of spheres of radius r and $\hat{\eta}_x$ then the unit normal (to $\partial C_r(\mathbb{X})$) outward (to $C_r(\mathbb{X})$) vector at x is, almost everywhere, well define for $x \in \partial C_r(\mathbb{X})$ and $\mathring{B}(x + r\hat{\eta}_x, r) \subset C_r(\mathbb{X})^c$ we then have:

$$\angle \hat{\eta}_x, \eta_{\pi_{\partial S}(x)} = O(\sqrt{d_H(C_r(\mathbb{X}), S)}).$$

Indeed, introduce $O_{\text{out}} = x + r\hat{\eta}_x$, $O_{\text{in}} = \pi_{\partial S}(x) - r_0\eta_{\pi_{\partial S}(x)}$, $y_1 \in \partial B(O_{\text{in}}, r_0) \cap [O_{\text{in}}, O_{\text{out}}]$, $y_2 \in \partial B(O_{\text{out}}, r) \cap [O_{\text{in}}, O_{\text{out}}]$ (see Figure 2.3), we have $y_1 \in C_r^c$ and $d(y_1, \partial S) \geq ||y_1 - y_2||$ thus $d(S, C_r) \geq ||y_1 - y_2||$. Small basic euclidean geometry calculus (taking into account that $||x - \pi_{\partial S}(x)|| \leq d_H(C_r, S)$) give the announced inequality in angle.

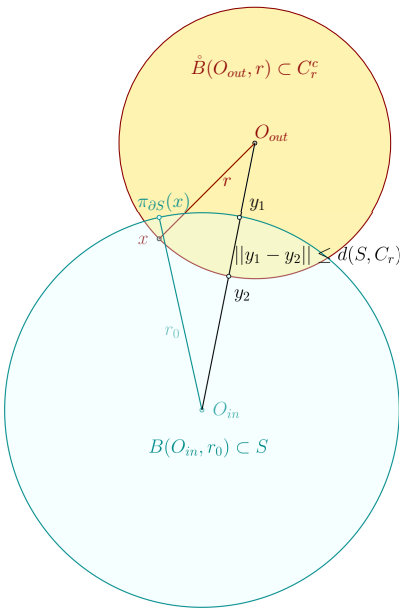

Figure 2.3: O_{out}, O_{oin}, y_1 and y_2

Injectivity of $\pi_{\partial S} : \partial C_r(\mathbb{X}) \to \partial S$ comes from the angle inequality. Continuity of $\pi_{\partial S}$ being a consequence of the positive reach of ∂S and surjectivity coming from positive reach and results in [Rodríguez Casal 2007]. We obtain that $\pi_{\partial S}$ is one to one. Continuity of $\pi_{\partial S}^{-1}$ comes from positive reach allowing to obtain the attempted homeomorphism. □

A similar result setting that $C_r(\mathbb{X}_n) \approx S$ e.a.s. should be easy to deduce following the same idea of proof than in [J7] but it is not written yet.

Note that other generalizations of the convex hull exists as [Cholaquidis *et al.* 2014] and [Cholaquidis & Cuevas 2020].

2.2.2.2 Local Convex Hull

In [J7] we studied another kind of convex hull generalization. The proposed support estimator initially introduced in [Getz & Wilmers 2004] in the applied context of the home range estimation is:

$$LCH_r(\mathbb{X}_n) = \bigcup_i \mathcal{H}(B(X_i, r) \cap \mathbb{X}_n).$$

We obtain the following result

Theorem 10 (A. and Bodart 16 (universality))**.** *If S is compact and $Dim_{Mink}(\partial S) < D$ then $d_H(\mathbb{X}_n, S) \overset{a.s.}{\to} 0$ and for any sequence r_n satisfying $r_n \overset{a.s.}{\to} 0$ and $r_n \geq 4d_H(S, \mathbb{X}_n)$ we have that: $|LCH_{r_n} \Delta S|_D \overset{a.s.}{\longrightarrow} 0$, $d_H(LCH_{r_n}(\mathbb{X}_n), S) \overset{a.s.}{\longrightarrow} 0$ and $d_H(\partial LCH_{r_n}(\mathbb{X}_n), \partial S) \overset{a.s.}{\longrightarrow} 0$*

When additionally assuming that the boundary is smooth enough (given by the rolling ball condition) and that the density is α-decreasing we obtain much better convergence rates, namely:

Theorem 11 (A. and Bodart 16 (optimality and Topology preservation))**.** *If S is compact and satisfies the r_0 inside and outside rolling ball condition and if f_X is α-decreasing. then for sequences $r_n = \lambda(\ln n/n)^{\frac{1}{D+1+2\alpha}}$ we have:*

$$|LCH_{r_n} \Delta S| = O\left(\frac{\ln n}{n}\right)^{\frac{2}{D+1+2\alpha}} \quad e.a.s$$

$$d_H(LCH_{r_n}(\mathbb{X}_n), S) = O\left(\frac{\ln n}{n}\right)^{\frac{2}{D+1+2\alpha}} \quad e.a.s$$

$$d_H(\partial LCH_{r_n}(\mathbb{X}_n), \partial S) = O\left(\frac{\ln n}{n}\right)^{\frac{2}{D+1+2\alpha}} \quad e.a.s$$

$$LCH_{r_n}(\mathbb{X}_n) \approx S \ e.a.s \text{ and } \partial LCH_{r_n}(\mathbb{X}_n) \approx \partial S \ e.a.sS$$

Remark: It thus has a minimax (up to a ln) rate when $D = 2$ or when $\alpha = 0$.

2.2.2.3 Perspectives

Computational limit We initially decided to study the Local Convex Hull support estimator for two main reasons:

1. We wanted to have theoretical results on a very commonly used (in ecology) tool ([Getz & Wilmers 2004] has more than 450 citations).

2. It has minimax rates and does not degenerate in smaller dimension. If we are not in the full dimensional case we may have $C_r(\mathbb{X}_n) = \mathbb{X}_n$ (this result will be exploited later) while we really believe that $LCH_{r_n}(\mathbb{X}_n)$ is still minimax in the lower dimensional case (see [Divol 2020], that proved such a result for different but very close estimator, when the support is a sub-manifold of $\mathbb{R}^D$ without boundary).

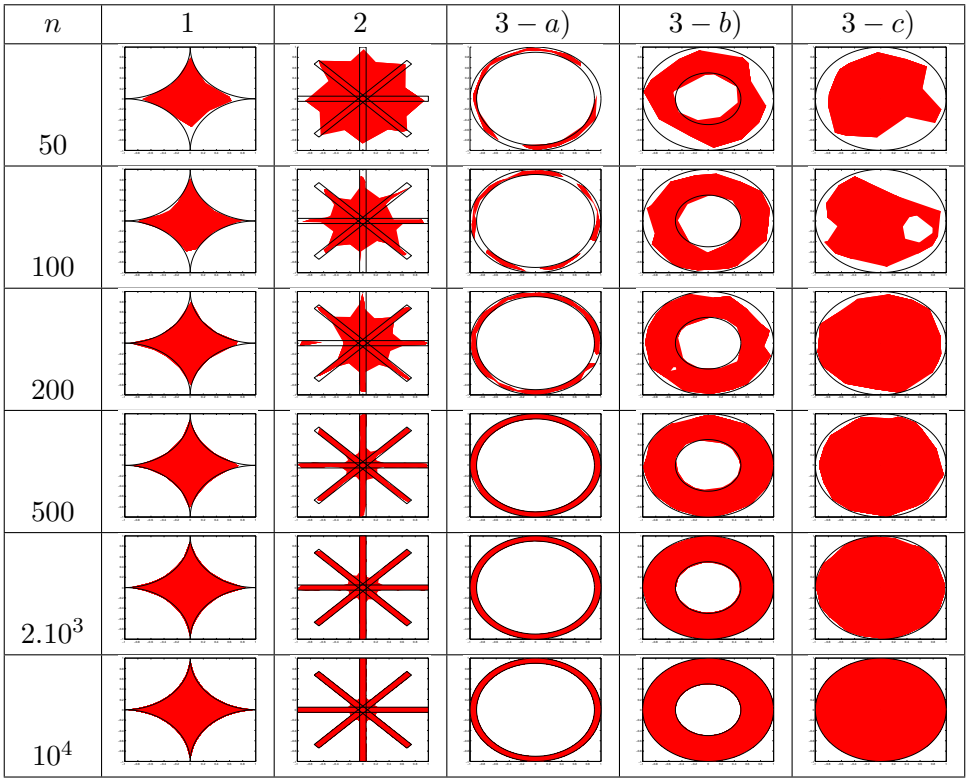

Figure 2.4: LCH estimator uniform drawn on a (only) compact set (1), on a standard set (2) and on sets with the r_0-inside and outside ball property $(3-a)$, $3-b)$ and $3-c)$ with different values of r_0

3. We thought it was computationally easier than the local convex hull.

If the two first reasons are actually good ones and if, in dimension 2 it is very easy to plot estimated support according to the local convex hull and possible to have an explicit description of the boundary, it appears that the computation of the boundary of the local convex hull in higher dimension is very costly.

In fact we believe now that, when aiming at considering the boundary, a really good tool, on theoretical and practical aspects is the r-shape. To define the r-shape we must first define the Delaunay simplices.

Definition 9 (Delaunay simplex)**.** *$\sigma = \mathcal{H}(X_{i_1}, \ldots, X_{i_{D+1}})$ is a Delaunay simplex of $\mathbb{X}_n$ if $\mathring{B}(O_\sigma, r_\sigma) \cap \mathbb{X}_n = \emptyset$ where 0_σ and r_σ are respectively the center and radius of the circumscribed sphere of $X_{i_1}, \ldots, X_{i_{D+1}}$. We will denote by Σ the set of the Delaunay simplices of $\mathbb{X}_n$*

The r-Shape of $\mathbb{X}_n$, denoted by $Sh_r(\mathbb{X}_n)$ is then:

$$Sh_r(\mathbb{X}_n) = \bigcup_{\sigma \in \Sigma, r_\sigma \leq r} \sigma$$

It has some very good computational advantages :

1. For a given sample, there is only a finite number of possible r-shape which is a great advantage when aiming at tuning the parameter r
2. Due to the fact that we have a description which is a triangulation of $\hat{S}_r$ it is easy to compute its boundary (that are is the set of faces that are only present once).
3. Also, triangulation aspects allows to compute topological invariants.

It is easy to deduce from [Rodríguez Casal 2007] that we have the following property.

Proposition 1. *When S satisfies the r_0-rolling ball property, for $r < r_0$ we have the existence of sequences $r_n = O(\ln n/n)^{2/(D+1)}$ such that*

$$C_r(\mathbb{X}_n) \subset Sh_r(\mathbb{X}_n) \subset S \oplus r_n B \text{ e.a.s.}$$

From that we can infer that r-shape has minimax rates under rolling ball condition and almost uniform densities.

The tricky point consists in proving the topological preservation properties (that is easily obtained when $d = 2$ but not in higher dimensions).

KNN approach and Robustness problems If we forgot about the computational cost of the Local Convex hull one can point out another advantage of such a method is that we can easily obtain a "local" estimator turning the fix radius parameter r into a function of i and considering

$$\bigcup_i \mathcal{H}(B(X_i, r_i) \cap \mathbb{X}_n)$$

which can be useful when the density has great variations. A classical way to choose r_i is to consider $r_i^{(k)}$ the distance from X_i to its k-the nearest neighbor (in $\mathbb{X}_n$) and define

$$LCH_k^{KNN} = \bigcup_i \mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)$$

Classically we can expect that since $k_n \to +\infty$ with $k_n/n \to 0$, it will provide consistant estimators and minimax rates for well chosen value of the k_n sequence (See Figure 2.5 first line).

Unfortunately such an idea provides drastically not robust estimators (See Figure 2.5 second line). To be convinced, suppose that you have a unique outlier X_{n+1} and a sample points $\mathbb{X}_n$ drawn with a almost uniform distribution on a smooth enough support S. Suppose that X_{n+1} is far enough from the support ($d(X_{n+1}, S) \geq r_n$) consider $\mathbb{Y} = \mathbb{X}_n \cup \{X_{n+1}\}$ we have $LCH_{r_n}(\mathbb{Y}) = LCH_{r_n}(\mathbb{X}) \cup \{X_{n+1}\}$ and then we have, at least $\mathring{LCH}_{r_n}(\mathbb{Y}) = \mathring{LCH}_{r_n}(\mathbb{X}_n)$ but $LCH_k^{KNN}(\mathbb{Y})$ far from $LCH_k^{KNN}(\mathbb{X}_n)$.

A way to overcome this problem, keeping the KNN advantages should be to get inspired by [Brunel 2016b] and to consider the (κ)-convex hull defined as follows and illustrated in Figure 2.6.

Definition 10 ((κ)-convex hull)**.** *let Y be a finite set of point $\mathcal{H}_\kappa(Y)$ is the intersection of all the half spaces that contains $\#Y - \kappa$ points of Y*

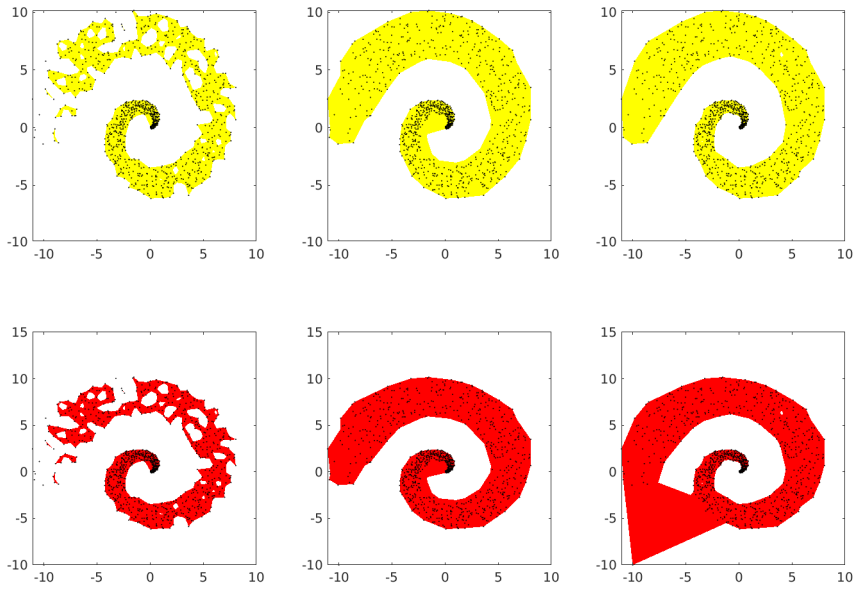

Figure 2.5: Upper line: 1000 points drawn on a "galaxy" shape spiral, lower line : the same points and an outlier located at $(-10,-10)$ then, from left to right: $LCH_{0.8}$ (fits quite well the support for points close to the origin but is too small for "queue" points); LCH_2 (fits quite well for "queue" points but is too large near the origin) ; LCH_{15}^{KNN} (provide good results when the data is not corrupted not when the outlier is added).

and use

$$LCH_{k,\kappa}^{KNN} = \bigcup_i \mathcal{H}_\kappa(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)$$

Which might be quite robust but may have an heavy computational cost (at least for $d \geq 2$).

We also may be inspired by the really recent [Brunel *et al.* 2021] that also deals with convex hull in a noisy environment.

Such ideas propose an alternative from [Brécheteau & Levrard 2020] way of dealing with robustness in geometric inference.

2.3 Perspectives in density and level set estimation

2.3.1 Density estimation

The density Level Set (at the level t) usually defined as $L_t = \overline{\{z \in \mathbb{R}^D, f(z) > t\}}$ or its "quantile" version $Q_p = L_{t^{(p)}}$ with $t^{(p)} = \sup\{t \geq 0, \mathbb{P}(L_t) \geq p\}$ (see [Cadre *et al.* 2013]) has been extensively studied due to its wide field of practical applications. See for instance [Baíllo *et al.* 2001], [Baíllo 2003],

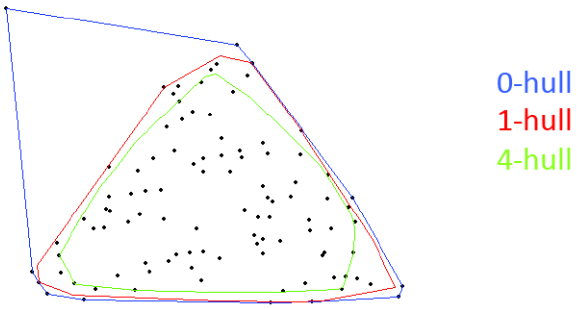

Figure 2.6: Different κ-convex hulls, picture taken from V.E. Brunel's presentation CIRM 2021

[Biau *et al.* 2007], [Butucea *et al.* 2007], [Cadre 2005], [Chen *et al.* 2015], [Polonik 1995], [Rigollet & Vert 2009], [Rodríguez-Casal & Saavedra-Nieves. 2019b] or [Tsybakov 1997].

There is two main way of estimating levels sets, one of theme being a plug-in of a density estimator and propose to use $L_t = \overline{\{z \in \mathbb{R}^D, \hat{f}_n(z) > t\}}$.

Such method will fail if the support is compact and the density almost uniform. Indeed near to the boundary of ∂S the usual density estimators such as kernel or KNN ones are biased (when $x \in \partial S$ and ∂S is a $\mathcal{C}^2$ manifold $\mathbb{E}(\hat{f}_n(x)) \to f_X(x)/2$.

We aim at correcting this bias with an easy to implement algorithm when the support is unknown.

As for support estimation we can get inspired of [Getz & Wilmers 2004] that proposed the following algorithm to plot the core area (which, in a mathematical sense is the level set of the density).

For some given k

1. Compute $V_i = |\mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)|$.
2. Sort the i so that $|\mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)|$ is decreasing.
3. Plot $\mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)$ with an increasing level of gray.

that can mathematically be formulated as:

$$\hat{L}_\lambda = \bigcup_{i, V_i \leq \frac{k}{n\lambda}} \mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)$$

and

$$\hat{f}(x) = \max_{i, x \in \mathcal{H}(B(X_i, r_i^{(k)}) \cap \mathbb{X}_n)} \frac{k}{nV_i}$$

We definitively think that taking into account the volume of local convex hull is a very good idea since estimating f_X with $\hat{f}_X(x) = \frac{k}{n|\widehat{B(x, r_x^{(k)}) \cap S}|}$ seems to be an intuitive way to

obtain an asymptotically unbiased density estimator. If the bias is expected to converge towards 0 the density is "twice overestimated". Indeed $|H(B(x,r_x^{(k)})\cap \mathbb{X}_n|$ underestimates $|B(x,r_x^{(k)})\cap S|$ and this overestimation is emphasized by the max.

We propose to get inspired by [Baldin & Reiß 2016] to decrease the bias.

We are here going to present some results very recently obtained (a work currently in progress). Even if a bit premature the proof is (objectively) so brief and (subjectively) so nice that I cannot help presenting it here.

We propose to estimate the density with:

$$\hat{f}_{r_n,A}(x) = \frac{N^o_{x,r_n}}{(n-N^\partial_{x,r_n})V_{x,r_n}}\mathbb{I}_{V_{x,r_n}\geq A\omega_D r^D}\mathbb{I}_{N^\partial_{x,r_n}\leq n/2}$$

Where $N_{x,r_n} = \#\{\mathbb{X}_n \cap B(x,r_n)\}$, $H_{x,r_n} = \mathcal{H}(B(x,r_n),\cap\mathbb{X}_n)$, $V_{x,r_n} = |H_{x,r_n}|$, $N^\partial_{x,r_n} = \#\{\mathbb{X}_n \cap \partial H_{x,r_n}\}$ and $N^o_{x,r_n} = N_{x,r_n} - N^\partial_{x,r_n}$.

Theorem 12 (A. and Fraiman (21))**.** *If S is compact, satisfying the inside and outside rolling ball condition, if f is almost uniform and $f|_S$ is of class $\mathcal{C}^2$ then, there exists $A_{\min}$ such that, if $r_n = cn^{-1/(D+4)}$, for all $A \leq A_{\min}$:*

$$\max_{x\in S}\mathbb{E}(\hat{f}_{r_n,A}(x) - f(x))^2 = O(n^{-2/(D+4)})$$

If $d \leq 7$,

$$\max_{x\in S\ominus r_n B}\mathbb{E}(\hat{f}_{r_n,A}(x) - f(x))^2 \leq O(n^{-4/(D+4)})$$

Sketch of proof. First condition on S and the almost uniformity of the density warranty that, since $nr_n^D \to +\infty$ we have that $\mathbb{P}(V_{x,r_n} \leq A\omega_D r^D) \leq c_1\exp(-c_1'(nr_n^D)^{1/3})$ (easily obtained using classical calculus in that topic).

Second, $N^\partial_{x,r_n} \leq N_{x,r_n}$ and then, applying Hoeffding's inequality and $r_n \to 0$ it comes that $\mathbb{P}(N^\partial_{x,r_n}) \geq n/2) \leq \exp(-n/8)$ for n large enough.

Introduce $\tilde{\Gamma}_{x,r_n} = \int_{H_{x,r_n}} f(z)dz$, we have the following decomposition of $\hat{f}_{r_n,A}$

$$\hat{f}_{r_n,A}(x) - f(x) = \frac{1}{|H_{x,r_n}|(n-N^\partial_{x,r_n})}\left(N^o_{x,r_n} - \tilde{\Gamma}_{x,r_n}(n-N^\partial_{x,r_n})\right) + \left(\frac{\tilde{\Gamma}_{x,r_n}}{|H_{x,r_n}|} - f(x)\right)$$

Let $\varepsilon_1 = \frac{1}{|H_{x,r_n}|(n-N^\partial_{x,r_n})}\left(N^o_{x,r_n} - \tilde{\Gamma}_{x,r_n}(n-N^\partial_{x,r_n})\right)$

$N^o_{x,r_n}|H_{x,r_n} \sim Binom(n-N^\partial_{x,r_n},\tilde{\Gamma}_{x,r_n})$ (see [Baldin & Reiß 2016] for the possibility of conditioning by H_{x,r_n})

Thus $\mathbb{E}(\varepsilon_1^2|C_{x,r_n}) \leq \frac{\Gamma_{x,r_n}}{|C_{x,r_n}|^2(n-N^\partial_{x,r_n})}$ and then

$$\mathbb{E}(\varepsilon_1^2|C_{x,r_n} \geq A\omega_d r_d, N^\partial_{x,r_n} \leq n/2) \leq \frac{2\max_S f}{A^2\omega_d n r_n^d}$$

Let now $\varepsilon_2 = \frac{\int_{H_{x,r_n}} f(z)dz - \int_{H_{x,r_n}} f(x)dz}{|H_{x,r_n}|}$ for all x : $\varepsilon_2 = O(r)$

If $B(x,r_n) \subset S$ then $\varepsilon_2 = \frac{\int_{B(x,r_n)}(f(z)-f(x))dz - \int_{B(x,r_n)\setminus H_{x,r_n}}(f(z)-f(x))dz}{|H_{x,r_n}|}$

A 2 order Taylor expansion (possible since f is $\mathcal{C}^2$ on S gives $\int_{B(x,r_n)} |f(z)-f(x)|dz = O(r_n^2)$ then a direct corollary of results in [Brunel 2020] is that

$$\mathbb{E}\left(\int_{B(x,r_n)\setminus H_{x,r_n}} \frac{|f(z)-f(x)|dz}{|H_{x,r_n}|}\right)^2 = O(r_n^2(nr_n^D)^{-4/d(d+1)})$$

Finally notice that $r_n(nr_n^D)^{-2/D(D+1)} \leq O(r_n^2)$ when $D \leq 7$ □

We also have, with very similar proof use of Bennet inequality to bound the probability that $\mathbb{P}(|\varepsilon_1| \geq x)$ and [Brunel 2020] for $\mathbb{P}(|\varepsilon_2| \geq x)$

Theorem 13 (A. and Fraiman (21)). *If S is compact, satisfying the inside and outside rolling ball condition, if f is almost uniform and $f|_S$ is of class $\mathcal{C}^2$ then, there exists $A_{\min}$ such that, if $r_n = cn^{-1/(D+4)}$, for all $A \leq A_{\min}$:*

for all $x \in S$ we have that for n large enough

$$\mathbb{P}\left(|\hat{f}_{r_n,A}(x) - f(x)| \geq c_1 \ln n n^{-1/(D+4)}\right) \leq 3n^{-2.4}$$

When $d \leq 7$ for all $x \in S \ominus rB$ we have that for n is large enough

$$\mathbb{P}\left(|\hat{f}_{r_n,A}(x) - f(x)| \geq c_2 \ln n n^{-2/(D+4)}\right) \leq 3n^{-2.4}$$

From which we obtain, by sum and Borrel Cantelli lemma, the following Corollary.

Corollary 2 (A. and Fraiman (21)). *If S is compact, satisfying the inside and outside rolling ball condition, if f is almost uniform and $f|_S$ is of class $\mathcal{C}^2$ then, there exists $A_{\min}$ such that, if $r_n = cn^{-1/(D+4)}$, for all $A \leq A_{\min}$ we have*

$$\max_i |\hat{f}_{r_n,A}(X_i) - f(X_i)| \leq c_1 \ln n.n^{-1/(D+4)} \ \ e.a.s.$$

And, when $d \leq 7$

$$\max_{i, X_i \in S\ominus rB} |\hat{f}_{r_n,A}(X_i) - f(X_i)| \leq c_1 \ln n.n^{-2/(D+4)} \ \ e.a.s.$$

This last corollary being the starting points of new results on level set estimation that is convenient with almost uniform densities on compact support.

2.3.1.1 Intensity estimation

We can also aim to estimate level sets to draw "risk" maps when data are obtained in a quite different way than the presented one. Most of the time the location are deterministic, the territory being subdivided into K-deterministic subsets C_k containing a population of N_k individuals. In each cell C_k we then observe $n_k \sim \mathcal{P}(p_k N_k)$ "victims" and we aim to draw a risk map that is mostly a level set on p.

2.4 Estimation of the volume and surface

For some practical application it also sound convenient to estimate the volume and the surface area of the support. For instance, in medicine the ratio surface area/volume of a tumor is correlated with its dangerosity (see [Andea *et al.* 2004] for instance).

2.4.1 Perspective in Volume estimation

Concerning the volume, under inside and outside rolling ball hypothesis, $|C_r(\mathbb{X}_n)|_D$ underestimates $|S|_D$. Under uniformity hypothesis, this problem has been nicely overcame in [Arias-Castro *et al.* 2019] obtaining a minimax convergence rate of $n^{-\frac{D+3}{2(D+2)}}$ with use a decomposition of $\mathbb{X}_n$ in two subsets, $\mathbb{X}_m^{(1)}$ being used to compute $C_r(\mathbb{X}_m^{(1)})$ an $X_{n-m}^{(2)}$ is used to obtain a Monte Carlo estimation of the missing volume.

Two main developments may be envisaged:

1. Can we skip the two subsets decomposition with the use of a correction based on the number of observations on the boundary (as in [Baldin & Reiß 2016]) ?

2. Can we extend the result to non uniform samples (in this special case the generalization to almost uniform densities is far from trivial)

2.4.2 Surface estimation

Let now be interested in the surface area ($|\partial S|_{D-1}$) estimation. Minimax convergence rates are conjectured to be the same than for volume estimation but they are far from being achieved. Notice than, even if the much easier problem of surface area estimation under uniform hypothesis and a "double" sample information (a sample uniformly drawn on S and another one uniformly drawn "outside" S) convergence rates are far from optimality:

1. The proposals given in [Cuevas *et al.* 2007], [Pateiro-López & Rodríguez-Casal 2008] and [Cuevas *et al.* 2013] aim to estimate the Minkowski content of ∂S. In [Cuevas *et al.* 2013] a very general convergence result is obtained, while in [Cuevas *et al.* 2007] a convergence rate of order $n^{-1/2D}$ is obtained under some mild hypotheses, and later on, in [Pateiro-López & Rodríguez-Casal 2008] a convergence rate of order $n^{-1/(D+1)}$ is achieved, under stronger assumptions.

2. In [Jimìenez & Yukich 2011] a very nice fully data driven method, based on the Delaunay triangulation is proposed under an homogeneous point process sampling scheme. The asymptotic rate of convergence of the variance is given, but there is no global convergence rate because no result is obtained for the bias.

3. Lastly, in [Thäle & Yukich 2016] a parameter-free procedure, based on the Voronoi triangulation is proposed, and a rate of convergence of order $\lambda^{-1/D}$ is obtained, under a Poisson Point Process (PPP) sampling scheme (where λ is the intensity of the PPP).

When we only have a sample from the inside (i.e. only observation drawn according to a distribution supported by S) and when the support is not supposed to be convex, up to our knowledge, only the $D = 2$ dimensional case has been studied in [Arias-Castro & Rodríguez-Casal 2017].

In [P2], with Alejandro Cholaquidis and Ricardo Fraiman we propose to study two surface area estimators that make use of the Crofton formula.

To introduce the general Crofton's formula in $\mathbb{R}^D$ for a compact $(D-1)$-dimensional manifold M, let us define first the constant

$$\beta(D) = \Gamma(D/2)\Gamma((D+1)/2)^{-1}\pi^{-1/2},$$

where Γ stands for the well known Gamma function.

Given a vector $\theta \in (\mathcal{S}^+)^{D-1}$ and a point y, $r_{\theta,y}$ denotes the line $\{y + \lambda\theta, \lambda \in \mathbb{R}\}$.

Let $\theta \in (\mathcal{S}^+)^{D-1}$, θ determine a $(D-1)$-dimensional linear space $\theta^\perp = \{v : \langle v, \theta\rangle = 0\}$. Given $y \in \theta^\perp$, let us write $n_M(\theta, y) = \#(r_{\theta,y} \cap M)$, where $\#$ is the cardinality of the set.

It is proved in [Federer 1969] (see Theorem 3.2.26) that if M is an $(D-1)$-dimensional rectifiable set, then the integral-geometric measure of M (which will be denote by $I_{D-1}(M)$, and is defined by the right-hand side of 2.3) equals its $(D-1)$-dimensional Hausdorff measure, i.e.,

$$|M|_{D-1} = I_{D-1}(M) = \frac{1}{\beta(D)} \int_{\theta\in(\mathcal{S}^+)^{D-1}} \int_{y\in\theta^\perp} n_M(\theta, y) d\mu_{D-1}(y) d\theta. \tag{2.3}$$

The measure $d\theta$ is the uniform measure on $(\mathcal{S}^+)^{D-1}$ (with total mass 1).

Along this section we assume that ∂S is the boundary of a compact set $S \subset \mathbb{R}^D$ such that $S = \overline{\mathring{S}}$. We also assume that S fulfills the outside and inside α-rolling condition, and then ∂S is rectifiable (see Theorem 1 in [Walther 1999]). From this it follows that $I_{D-1}(\partial S) = |\partial S|_{D-1} < \infty$, which implies (by (2.3)) that, except for a set of measure zero with respect to $d\mu_{D-1}(y)d\theta$, any line $r_{\theta,y}$ meets ∂S a finite number of times: $n_{\partial S}(\theta, y) < \infty$. From Theorem 1 in [Walther 1999], it also follows that ∂S is a $\mathcal{C}^1$ manifold, which allows us to consider for all $x \in \partial S$, η_x, the unit outward normal vector.

For the Devroye–Wise based surface area estimator we will assume that ∂S satisfies a technical hypothesis named (C, ε_0)-regularity.

Definition 11. *Let us define $E_\theta(\partial S) = \{x \in \partial S, \langle \eta_x, \theta\rangle = 0\}$ and F_θ its orthogonal projection onto $\theta^\perp$. Let us define, for $\varepsilon > 0$,*

$$\varphi_\theta(\varepsilon) = \left|\theta^\perp \cap B(F_\theta, \varepsilon)\right|_{D-1}.$$

We will say that ∂S is (C, ε_0)-regular if for all θ and all $\varepsilon \in (0, \varepsilon_0)$, $\varphi'_\theta(\varepsilon)$ exists and $\varphi'_\theta(\varepsilon) \le C$.

Once the rolling balls condition is imposed, we are deeply convinced that the (C, ε_0)-regularity of the boundary is not a too restrictive hypothesis. Roughly speacking we have that, if ∂S is a $\mathcal{C}^2$ manifold with a positive reach α, F_θ is an union of $(D-2)$-dimensional manifold that and there norm of second fundamental form is uniformly upper bounded by α. When the number of manifolds in the union is a finite number N_θ and taking the

minimum reach of all the manifolds τ_θ, due to polynomial volume approximation it comes that we have the (C, ε_0)-regularity when $\sup_\theta N_\theta < +\infty$ and $\inf_\theta \tau_\theta > 0$ that is, we think, not that restrictive. Moreover it is conjectured in [Alesker 2018] as even more general.

For the Devroye-Wise type estimator we will also show that the convergence rate can be quadratically improved if we additionally assume that the number of intersections between any line and ∂S is bounded from above (that exclude the case of a linear part in ∂S).

Definition 12. *Given $S \subset \mathbb{R}^D$, we say that ∂S has a bounded number of linear intersections if there exists N_S such that , for all $\theta \in (\mathcal{S}^+)^{D-1}$ and $y \in \theta^\perp$, $n_{\partial S}(\theta, y) \leq N_S$.*

2.4.2.1 Devroye–Wise based approach

The Devroye based surface area approach is not only a plug-in the Crofton formula to the Devroye Wise estimator. The idea is to estimate the number of intersection of a line with the boundary of the support with the following procedure based on two offsets of the data recall that $DW_r(\mathbb{X}_n) = \cup B(X_i, r)$.

Definition 13. *Let ε_n a sequence of positive real numbers, such that $\varepsilon_n \to 0$ and $\mathbb{X}_n \subset S$ a set not necessarily finite. Consider a line $r_{\theta,y}$. If $DW_{\varepsilon_n}(\mathbb{X}_n) \cap r_{\theta,y} = \emptyset$, define $\hat{n}_{\varepsilon_n}(\theta, y) = 0$, otherwise:*

- *denote by $I_1, \ldots, I_m$ the connected components of $DW_{\varepsilon_n}(\mathbb{X}_n) \cap r_{\theta,y}$. Order this sequence in such a way that $I_i = (a_i, b_i)$, with $a_1 < b_1 < a_2 < b_2 < \cdots < a_m < b_m$.*
- *Define new intervals $A_j = (a_{i(j)}, b_{i(j)+\ell(j)})$ that are the intervals such that $A_j \subset DW_{4\varepsilon_n}$ and when exists we also have $(b_{i(j)-1}, a_{i(j)}) \nsubseteq SW_{4\varepsilon_n}$ and $(b_{i(j)+\ell(j)}, a_{i(j)+\ell(j)+1}) \nsubseteq DW_{4\varepsilon_n}$.*
- *Let m' be the number of disjoint open intervals A_j that this process ended with. Then define $\hat{n}_{\varepsilon_n}(\theta, y) = 2m'$.*

Our first proposed estimator is

$$\hat{I}_{D-1}(\partial S) = \frac{1}{\beta(D)} \int_{\theta \in (\mathcal{S}^+)^{D-1}} \int_{y \in \theta^\perp} \hat{n}_{\varepsilon_n}(\theta, y) d\mu_{d-1}(y) d\theta.$$

Under the assumption that ∂S has a bounded number N_S of linear intersections (see Definition 12) we will consider, for a given $N_0 \geq N_S$,

$$\hat{I}^{N_0}_{D-1}(\partial S) = \frac{1}{\beta(D)} \int_{\theta \in (\mathcal{S}^+)^{D-1}} \int_{y \in \theta^\perp} \min(\hat{n}_{\varepsilon_n}(\theta, y), N_0) d\mu_{d-1}(y) d\theta.$$

Theorem 14 (A. Cholaquidis and Fraiman 21)**.** *Let $S \subset \mathbb{R}^D$ be a compact set fulfilling the outside and inside α-rolling conditions. Assume also that S is (C, ε_0)-regular for some positive constants C and ε_0. Let $\mathbb{X}_n = \{X_1, \ldots, X_n\} \subset S$. Let $\varepsilon_n \to 0$ such that $d_H(\mathbb{X}_n, S) \leq \varepsilon_n$. Then*

$$\hat{I}_{D-1}(\partial S) = |\partial S|_{D-1} + \mathcal{O}(\sqrt{\varepsilon_n}). \tag{2.4}$$

Moreover, for n large enough,

$$|\mathcal{O}(\sqrt{\varepsilon_n})| \leq \frac{4C\,diam(S)}{3\beta(D)\sqrt{\alpha}}\sqrt{\varepsilon_n},$$

C being the constant of the (C,ε_0)-regularity of S.

When it is additionlly supposed that the maximum number of interesection between a line and ∂S is finite, the convergence rate is quadratically improved using $\hat{I}_{D-1}^{N_0}(\partial S)$ instead of $\hat{I}_{D-1}$.

Theorem 15 (A. Cholaquidis and Fraiman 21)**.** *Let $S \subset \mathbb{R}^D$ be a compact set fulfilling the outside and inside α-rolling conditions. Assume also that S is (C,ε_0)-regular for some positive constants C and ε_0 and that ∂S has a number of linear intersection bounded by N_S. Let $\mathbb{X}_n = \{X_1,\dots,X_n\} \subset S$. Let $\varepsilon_n \to 0$ such that $d_H(\mathbb{X}_n, S) \leq \varepsilon_n$ and $N_0 \geq N_S$. Then*

$$\hat{I}_{D-1}^{N_0}(\partial S) = |\partial S|_{D-1} + \mathcal{O}(\varepsilon_n).$$

Moreover, for n large enough,

$$|\mathcal{O}(\varepsilon_n)| \leq \frac{5}{\beta(D)} C N_0 \varepsilon_n,$$

C being the constant of the (C,ε_0)-regularity of S.

From this deterministic results we can obtain various convergence rate that only depends on $d_H(\mathbb{X}_n, S)$. For instance, when $\mathbb{X}_n$ is an iid sample with the distribution is almost uniform then the rates are respectively of order $(\ln n/n)^{1/2D}$ or $(\ln n/n)^{1/D}$ depending on the hypotheses (bounded number of linear intersection of not). Even if these rates are associated to slow convergence method it has to be noticed that, some of the proposed estimators in the "inside/oustide" sampling model are not better than that.

2.4.2.2 The r-convex hull based approach

Theorem 16 (A., Cholaquidis and Fraiman 21)**.** *Let $S \subset \mathbb{R}^D$ be a compact set fulfilling the inside and outside r_0-rolling conditions. Let $r < r_0$ be a positive constant and let $\mathbb{X}_n \subset S$ be a finite set such that $d_H(\partial C_r(\mathbb{X}_n), \partial S) \leq \varepsilon_n$ with $\varepsilon_n \leq \min\left(\frac{r_0 r}{16(r+r_0)}, \frac{1}{(D-1)r_0}\right)$. Then*

1. *$\pi_{\partial S} : \partial C_r(\mathbb{X}_n) \to \partial S$ (where $\pi_{\partial S}(x)$ denotes the projection onto ∂S) is one to one*
2. *$||\partial S|_{D-1} - |\partial C_r(\mathbb{X}_n)|_{D-1}| = \left(3r_0 + 64\frac{r+r_0}{r r_0}\right)\varepsilon_n$,*

Sketch of proof. The proof is based on the above mentioned results setting that $\angle\eta_{\pi_{\partial S}(x)}, \hat{\eta}_x \leq \varepsilon_n$ almost everywhere on $\partial C_r(\mathbb{X}_n)$ where η_x is the outward (to S) unit normal (to ∂S) vector at $x \in \partial S$ and $\hat{\eta}_x$ is the outward (to $C_r(\mathbb{X}_n)$) unit normal (to $\partial C_r(\mathbb{X}_n)$) vector at $x \in \partial C_r(\mathbb{X}_n)$. This, combined with $d_H(\partial C_r(\mathbb{X}_n), \partial S) \leq \varepsilon_n^2$ (in [Rodríguez Casal 2007]) allows with a little of differential geometry and a Taylor expansion to have the announced result. □

2.4.2.3 Perspectives and Conjecture with the r-Shape

There is two main axis to improve this work.

1. Computationally the Monte Carlo step is not fully satisfactory. It is shown in [Arias-Castro & Rodríguez-Casal 2017] that the use of $Sh_r(\mathbb{X}_n)$ allows to estimate the perimeter. This is due to the fact that, in dimension 2, ∂C_r is an union of arcs whose length is "close" to length of segments. The problem is that, in higher dimension it may not be the case. We yet have an explicit and computationally feasible correction but we can prove it works only once we can prove that the surface of the r-shape is homeomorphic to the surface of the r-convex hull that is an aforementioned difficult perspective.

2. Can we find an estimator with minimax rates ?

Chapter 3

Lower dimensional context (Manifold learning)

3.1 Introduction

The lower dimensional context is the case where the support of the distribution is either a d-dimensional ($d < D$) sub-manifold of $\mathbb{R}^D$ (true lower dimensional case) or "close" to a d-dimensional ($d < D$) sub-manifold of $\mathbb{R}^D$ (noisy lower dimensional case)

The true lower dimensional case In [Edelsbrunner & Shah 1994] an algorithm for general manifold reconstruction based on Delaunay triangulation is proposed and is proved to have topological guarantees when a "reasonable" condition is satisfied. It has been proved that this condition is really reasonable when $D \leq 3$, but it is no more the case for higher dimensions, a counter example being build in [Oudot 2008] when $D = 4$ and $d = 2$. When assuming the manifold is without boundary a topologically preserving estimator ($D = 3$, $d = 2$) is presented in [Amenta *et al.* 2002]). Minimax rates are derived in [Aamari & Levrard 2019] where an estimator is proposed, another minimax estimator is proposed in [Divol 2020] and, based on [Boissonnat & Ghosh 2014], an estimator that is both minimax and topology preserving is proposed in [Aamari & Levrard 2018] (the price to pay being the algorithmic complexity).

Our main contribution consists in studying methods that allows boundary existence. In a first section of this chapter, we give a geometric setting that allows boundary. We then present tests that allow to decide if we are in the lower dimensional case or not (section 3.2.1) and if the boundary is empty or not (section 3.2.2). We propose solutions for manifold and its boundary estimation (section 3.2.3) and for volume, based on Minkowsky contents, estimation (section 3.2.4) Finally, as a perspective, we also propose estimators for volume and surface area estimation and for density estimation.

The noisy dimensional case is much more challenging. We will consider the case where $S \subset M \oplus \varepsilon B$ where M is a d-dimensional manifold. When $S \subset M \oplus \varepsilon_n B$ with $\varepsilon_n \to 0$ quickly enough, results in [Aamari & Levrard 2018] an [Aamari & Levrard 2019] are still valid. We strongly believe that it is also true for the one obtain in [P1] (but calculus were difficult and long enough in the noiseless model). We nevertheless think that such an hypothesis of vanishing noise is a bit too restrictive and that we should be able to have data supported by $S \subset M \oplus \varepsilon B$ where ε might be small enough but constant. In [Genovese *et al.* 2012b] it is proved that the minimax rate is $n^{-2/(d+2)}$ (when M is a

d-dimensional manifold with positive reach and without boundary) but no computable estimator with such a rate have been yet found.

In section 3.3 we propose an estimator for the "amount of noise" ε and two denoising methods that allow to obtain, from data in $S \subset M \oplus \varepsilon B$, data in $S \subset M \oplus \varepsilon_n B$ that can then be taken as input of previously proposed methods.

3.2 Geometric Setting

3.2.1 Sub-manifolds with (possible) Boundary

By definition, the d-dimensional sub manifolds $M \subset \mathbb{R}^D$ with boundary are the subsets of $\mathbb{R}^D$ that can locally be parametrized either by the Euclidean space $\mathbb{R}^d$, or the half-space $\mathbb{R}^{d-1} \times \mathbb{R}_+$ [Lee 2011, Chapter 2].

We will only be interest in $\mathcal{C}^k$ manifold with $\mathcal{C}^k$ (or empty) boundary that can be defined as follows.

Definition 14 (Sub manifold with Boundary, Boundary, Interior)**.** *A closed subset $M \subset \mathbb{R}^D$ is a d-dimensional $\mathcal{C}^k$-*sub-manifold with boundary *of $\mathbb{R}^D$, if for all $p \in M$ and all small enough open neighborhood V_p of p in $\mathbb{R}^D$, there exists an open neighborhood U_0 of 0 in $\mathbb{R}^D$ and a $\mathcal{C}^k$-diffeomorphism $\Psi_p : U_0 \to V_p$ with $\Psi_p(0) = p$, such that either:*

1. $\Psi_p\left(U_0 \cap \left(\mathbb{R}^d \times \{0\}^{D-d}\right)\right) = M \cap V_p$.
   Such a $p \in M$ is called an interior *point of M, the set of which is denoted by* $\operatorname{Int} M$.

2. $\Psi_p\left(U_0 \cap \left(\mathbb{R}^{d-1} \times \mathbb{R}_+ \times \{0\}^{D-d}\right)\right) = M \cap V_p$.
   Such a $p \in M$ is called a boundary *point of M, the set of which is denoted by ∂M.*

Remark 1 (Boundaries)**.** *The* geometric *(or* differential*) boundary ∂M is not to be confused with the ambient* topological *boundary defined as $\bar{\partial} S := \bar{S} \setminus \mathring{S}$ for $S \subset \mathbb{R}^D$, where the closure and interior are taken with respect to the ambient topology of $\mathbb{R}^D$. Indeed, one easily checks that if $d < D$, then $\bar{\partial} M = M$. On the other hand, the two sets $\bar{\partial} M$ and ∂M coincide when $d = D$.*

Then, sub manifolds *without* boundary are those M that fulfill $\partial M = \emptyset$, i.e. that are everywhere locally parametrized by $\mathbb{R}^d$, and nowhere by $\mathbb{R}^{d-1} \times \mathbb{R}_+$. From this perspective — as confusing as this standard terminology can be —, sub manifolds without boundary are special cases of sub manifolds with boundary. Note that key instances of manifolds without boundary are given by boundaries of manifolds, as expressed by the following result.

Proposition 2 ([Lee 2011, Example 2.17])**.** *If $M \subset \mathbb{R}^D$ is a d-dimensional $\mathcal{C}^2$-sub-manifold with nonempty boundary ∂M, then ∂M is a $(d-1)$-dimensional $\mathcal{C}^2$-sub-manifold without boundary.*

3.2.2 Tangent and Normal Structures

In the present $\mathcal{C}^2$-smoothness framework, the difference between boundary and interior points sharply translates in terms of local first order approximation properties of M either by its so-called tangent cones or tangent spaces, which we now define.

Definition 15 (Tangent and Normal Cones and Spaces)**.** *Let $p \in M$, and Ψ_p its local parametrization from Definition 14.*

- *The* tangent cone $Tan(p, M)$ *of M at p is defined as*

$$Tan(p, M) := \begin{cases} d_0\Psi_p(\mathbb{R}^d \times \{0\}^{D-d}) & \text{if } p \in \operatorname{Int} M, \\ d_0\Psi_p\left(\mathbb{R}^{d-1} \times \mathbb{R}_+ \times \{0\}^{D-d}\right) & \text{if } p \in \partial M, \end{cases}$$

  where $d_0\Psi_p$ denotes the differential of Ψ_p at 0.
  The tangent space T_pM *is then defined as the linear span $T_pM := \operatorname{span}(Tan(p, M))$.*

- *The* normal cone $Nor(p, M)$ *of M at p is the dual cone of $Tan(p, M)$:*

$$Nor(p, M) := \{v \in \mathbb{R}^d \mid \forall u \in Tan(p, M), \langle u, v\rangle \leq 0\}.$$

  The normal space *of M at p is defined accordingly by $N_p(M) := \operatorname{span}(Nor(p, M))$.*

Whenever $p \in \operatorname{Int} M$, it falls under the intuition that $Tan(p, M) = T_pM$ and $Nor(p, M) = N_pM$, while when $p \in \partial M$, N_pM and T_pM share one direction which is orthogonal to $T_p\partial M$. These properties are summarized in the following proposition.

Proposition 3 (Outward-Pointing Vector)**.** *Let M be a $\mathcal{C}^2$-sub-manifold with boundary.*

- *If $p \in \operatorname{Int} M$, then $Tan(p, M) = T_pM$ and $Nor(p, M) = N_pM$ are orthogonal linear spaces spanning $\mathbb{R}^D$.*

- *If $p \in \partial M$, then $Tan(p, M)$ and $Nor(p, M)$ are complementary half-spaces. In particular, $T_pM \cap N_pM$ is one-dimensional. The unique unit vector η_p in $Nor(p, M) \cap T_pM$ is called the* outward-pointing *vector. It satisfies*

$$Tan(p, M) = T_pM \cap \{\langle \eta_p, .\rangle \leq 0\}, \quad Nor(p, M) = N_pM \cap \{\langle \eta_p, .\rangle \geq 0\},$$

  and

$$T_p\partial M \overset{\perp}{\oplus} \operatorname{span}(\eta_p) = T_pM,$$

  where $\overset{\perp}{\oplus}$ denotes the orthogonal direct sum relation.

3.2.3 Geometric Assumptions

Any $\mathcal{C}^2$-sub-manifold M of $\mathbb{R}^D$ admits a tubular neighborhood in which any point has a unique nearest neighbor on M [Bredon 1993, p.93]. However, the width of this tubular neighborhood might be arbitrarily small. This scenario occurs when M exhibits high

curvature or nearly self-intersecting areas [Aamari *et al.* 2019]. In this case, the estimation of M gets more difficult, since such locations require denser sample to be reconstructed accurately. The width of such a tubular neighborhood is given by the so-called *reach* ([Federer 1959, Definition 4.1]), whose formal definition goes as follows.

Given a closed set $S \subset \mathbb{R}^D$, the *medial axis* $\mathcal{M}(S)$ of S is the set of ambient points that do not have a unique nearest neighbor on S. More precisely, if

$$\mathrm{d}(z, S) := \min_{x \in S} \|z - x\|$$

stands for the *distance function* to S, then

$$\mathcal{M}(S) := \left\{ z \in \mathbb{R}^D | \exists x \neq y \in S, \|z - x\| = \|z - y\| = \mathrm{d}(z, S) \right\}. \tag{3.1}$$

The reach of S is then defined as the minimal distance from S to $\mathcal{M}(S)$.

Definition 16 (Reach)**.** *The* reach *of a closed set $S \subset \mathbb{R}^D$ is*

$$\tau_S := \min_{x \in S} \mathrm{d}\left(x, \mathcal{M}(S)\right) = \inf_{z \in \mathcal{M}(S)} \mathrm{d}\left(z, S\right).$$

Remark 2.

- *By construction of the medial axis Equation (3.1), the projection on S*

$$\pi_S(z) := \arg\min_{x \in S} \|x - z\|$$

is well defined (exactly) on $\mathbb{R}^D \setminus \mathcal{M}(S)$. In particular, π_S is well defined on any r-neighborhood of S of radius $r < \tau_S$.

- *One easily checks that S is convex if and only if $\tau_S = \infty$ [Federer 1959, Remark 4.2]. In particular, for the empty set $S = \emptyset$, we have $\tau_\emptyset = \infty$.*

Requiring a lower bound on the reach of a manifold amounts to bound its curvature [Federer 1959, Proposition 6.1], and prevents quasi self-intersection at scales smaller than the reach [Aamari *et al.* 2019, Theorem 3.4]. Moreover, it allows to assess the quality of the linear approximation of the manifold by its tangent cones. In fact, [Federer 1959, Theorem 4.18] shows that for all closed set $S \subset \mathbb{R}^D$ with reach $\tau_S > 0$, its tangent cone $Tan(x, S)$ is well defined at all $x \in S$, and $\mathrm{d}(y - x, Tan(x, S)) \leq \|y - x\|^2 / (2\tau_S)$ for all $y \in S$. This motivates the introduction of our geometric model below.

Definition 17 (Smooth Geometric Model)**.** *Given integers $1 \leq d \leq D$ and positive numbers $\tau_{\min}, \tau_{\partial,\min}$, we let $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ denote the set of compact connected d-dimensional $\mathcal{C}^2$-sub manifolds $M \subset \mathbb{R}^D$ with boundary, such that*

$$\tau_M \geq \tau_{\min} \quad \text{and} \quad \tau_{\partial M} \geq \tau_{\partial,\min}.$$

Remark 3.

- *Let us emphasize that the model $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ includes both sub manifolds with empty and non-empty boundary ∂M, the main requirement being that $\tau_{\partial M} \geq \tau_{\partial,\min}$. If $\partial M = \emptyset$, this requirement is always fulfilled since $\tau_\emptyset = \infty$. Note also that Definition 17 does not exclude the case $d = D$, in which case M consists of a*

domain of $\mathbb{R}^D$ with non-empty interior. Furthermore, since the boundary ∂M of a sub-manifold M is either empty or itself a sub-manifold without boundary, a non-empty ∂M cannot be convex. As a result, $\mathcal{M}^{d,D}_{\tau_{\min},\infty}$ is exactly the set of sub manifolds $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that have empty boundary. In particular, Definition 17 encompasses the model of [Genovese et al. 2012b, Kim & Zhou 2015, Aamari & Levrard 2019].

- *Similarly, since $\tau_M = \infty$ if and only if M is convex, $\mathcal{M}^{d,D}_{\infty,\tau_{\partial,\min}}$ is exactly the set of sub manifolds $M \in \mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ that are convex (and hence have non-empty boundary). In particular, Definition 17 encompasses the model of [Dümbgen & Walther 1996].*
- *In full generality, the two lower bounds on the respective reaches of M and ∂M are* not *redundant with one another. As shown in Figure 3, τ_M and $\tau_{\partial M}$ are not related when $d < D$. However, for $d = D$, ∂M is the topological boundary of M (see remark 1). In this case, [Federer 1959, Remark 4.2] and an elementary connectedness argument show that $\tau_M \geq \tau_{\partial M}$. Said otherwise, this means that the reach regularity of a full-dimensional domain is no worse than that of its boundary. Hence, $\mathcal{M}^{D,D}_{\tau_{\min},\tau_{\partial,\min}} = \mathcal{M}^{D,D}_{\tau_{\partial,\min},\tau_{\partial,\min}}$ for all $\tau_{\min} \leq \tau_{\partial,\min}$, so that for $d = D$, one may set $\tau_{\min} = \tau_{\partial,\min}$ without loss of generality.*

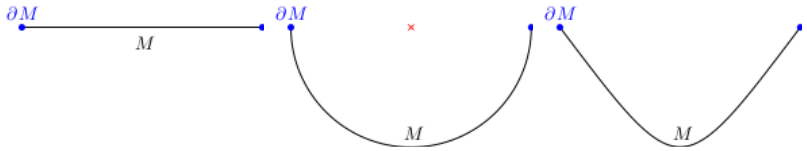

Figure 3.1: For $d < D$, the reach of a sub-manifold M and that of its boundary ∂M are not related. First example: $\tau_{\partial M} < \tau_M = \infty$, second: $\tau_{\partial M} = \tau_M$, and third: $\tau_{\partial M} > \tau_M$.

Remark: The model with $\mathcal{C}^2$ manifold with positive reach is an extension of the full dimensional case with the rolling ball condition. Due to [Federer 1959] it allows to deal with manifolds with few use of differential geometry making the proof easy to read by statisticians (and being fully honest, easy to do by statisticians also).

We sometime have results that also apply in less restrictive condition on the manifold and somehow authorize the boundary to have corner.

Definition 18 (Geometric Model with corner)**.** *Given integers $1 \leq d \leq D$ and positive number $\tau_{\min}$, we let $\mathcal{N}^{d,D}_{\tau_{\min},\delta,r_0}$ denote the set of compact connected d-dimensional $\mathcal{C}^0$-sub manifolds $M \subset \mathbb{R}^D$ with boundary, such that, $\tau_M \geq \tau_{\min}$ and for all $x \in M$, and $r \leq r_0$, $|M \cap B(x,r)|_d \geq \delta\omega_d r^d$*

Remark: The technical lemmas in [P1] implies that any manifold in $\mathcal{M}^{d,D}_{\tau_{\min},\tau_{\partial,\min}}$ is in $\mathcal{N}^{d,D}_{\tau_{\min},\delta,r_0}$ for any $\delta < 1/2$ and any r_0 small enough.

3.3 True lower dimensional context

The true lower dimensional context correspond to the class of distribution P supported by M a compact d-dimensional sub-manifold of $\mathbb{R}^D$ that have a density f with respect

to the volume measure $\text{vol}_M(A) = |A \cap M|_d$ on M. When f satisfies that for all $x \in M$ $0 < f_0 \leq f(x) \leq f_1 < +\infty$ we will say that the distribution is almost uniform on M.

Practically, the reach of M can be estimated (and thus we can suppose it is known) due to [Aamari *et al.* 2019] or [Berenfeld *et al.* 2021] but it is not yet the case for $\tau_{\partial M}$.

3.3.1 Testing the true lower dimensional case

In [J5] we provide a test helping to decide whether the support of the observations is "full dimensional" or "lower dimensional". It is less general than the test proposed in [Fefferman *et al.* 2013] since we do not allow to have a little amount of noise but it is much more easy to implement. The propose test takes into consideration the "size" of the Voronoi cells of the observations $\rho_i = \sup_{z \in \text{Vor}_{\mathbb{X}_n}(X_i)} ||z - X_i||$.

Roughly speaking, one one hand, when the support M is full dimensional ($\overline{\mathring{M}} = M$) its interior is not empty and then there exists some $B(x, r_0) \subset M$ and, if the distribution put enough weight around x (that is the case in our classical hypothesis of density "almost uniform", the Voronoi cells of observations close to x becomes small when n goes to infinity). One the other hand, when the support is a d-dimensional manifold with $d < D$ with positive reach τ_M then, for any observation $X_i \in M$ and any $r < \tau_M$ there exists $x \in \mathbb{R}^D$ such that $||x - X_i|| \geq r$ and thus the Voronoi of X_i has a large radius.

This is more mathematically expressed in the following theorem which is very easy to prove.

Theorem 17 (A., Cholaquidis and Cuevas 17). *1. If M is a d-dimensional manifold with positive reach τ_M then, for all i we have $\min_i \rho_i \geq \tau_M$*

2. If there exists x_0 and ρ_0 r and δ such that

(a) $B(x_0, \rho_0) \subset M$

(b) for all $x \in B(x_0, \rho_0)$ and all $r' < r$ we have $\mathbb{P}(B(x, r)) \geq \delta \omega_D r^D$

we have that $\min_i \rho_i \leq \left(\frac{\ln n}{\delta \omega_D n}\right)^{1/D}$ e.a.s.

It is thus possible to test the "lower dimensional" versus the "full dimensional" case with the test statistic $\min_i \rho_i$ with a decision rule "decide H_0 (M is lower dimensional)" if $\min_i \rho_i \geq r_n$. Due to [Penrose 1999] we can obtain a fully data driven method with a choice of r_n that depend on $\max_i \min_j ||X_i - X_j||$ which allows to decide correctly with probability one when n large enough (under the almost uniform assumption).

Note that, even if it appears a very easy and naive idea that is practically subject to the curse of dimensionality this initial idea of considering ρ_i is the source of most of the tools exposed in this chapter.

3.3.2 Testing the boundary existence

As mentioned in the introduction, most of the proposed methods in manifold estimation are adapted to manifold without boundaries. For instance in [Aamari & Levrard 2019] it is proved that, when M is a $\mathcal{C}^k$ manifold **without boundary** the minimax rate for manifold

estimation is of order $n^{-k/d}$, when aiming at estimating the tangent spaces the minimax rate is of order $n^{-(k-1)/d}$ and when aiming at estimating the second differential form the rate is of order $n^{-(k-2)/d}$.

When applying manifold estimation method proposed in [Aamari & Levrard 2019] to $\mathcal{C}^2$ manifold with boundary we obtain the deteriorated rate of $n^{-1/d}$ (surprisingly this loss do not apply to the tangent space estimation). More generally it has been empirically observed that methods for manifold estimation for manifold without boundary, fail (near to the boundary) when applied to manifold with boundary. This motivated us to develop a test of boundary existence and manifold estimators adapted to the non-empty boundary case.

The test statistics proposed in In [J1] is the following.

Definition 19. *Given an i.i.d. sample $X_1, \dots, X_n$ of a random row vector X with support $M \subset \mathbb{R}^D$, where M is a d-dimensional manifold with $d \leq D$, we will denote by $X_{j(i)}$ the j-nearest neighbor of X_i. For a given sequence of positive integers k_n, let us define, for $i = 1, \dots, n$,*

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

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

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

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

Theorem 18. *Assume that $X_1, \dots, X_n$ is an i.i.d. sample drawn according to an unknown distribution $\mathbb{P}_X$ supported by M, a $\mathcal{C}^2$ d-dimensional (d is assumed to be known) manifold with positive reach and that its boundary ∂M is either empty of a $\mathcal{C}^2$ manifold with positive reach. Also suppose that the density is almost uniform and Lipschitz continuous on S. For sequences (k_n) fulfilling condition that $k_n/n^{1/(d+1)} \to 0$, $k_n/(\ln(n))^4 \to \infty$ when $d > 1$ and $k_n/\sqrt{n \ln n} \to +\infty$ when $d = 1$.*

Then the test

$$\begin{cases} H_0 : & \partial M = \emptyset \\ H_1 : & \partial M \neq \emptyset \end{cases} \tag{3.2}$$

with the rejection zone

$$W_n = \left\{ \Delta_{n,k_n} \geq F_d^{-1}(9\alpha/(2e^3 n)) \right\}, \tag{3.3}$$

satisfies $\mathbb{P}_{H_0}(W_n) \leq \alpha + o(1)$ and has power 1 for n large enough.

Sketch of proof. The first step consist in proving that we have $r_n := \max_i r_{i,k_n} \xrightarrow{a.s.} 0$ (using that the density is bounded from below and the condition $k_n/n^{1/(d+1)} \to 0$).

Consider an observation X_{i_0} such that $d(X_{i_0}, \partial M) \geq r_{i_0,k_n}$. The regularity of the manifold and the continuity of the density given by condition will imply 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" to an uniform sample on the d-dimensional unit ball, 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 satisfy $d(X_i, \partial M) \geq r_{i,k_n}$. Even though the $\{\delta_{i,k_n}\}_i$ are not independent, we will obtain an asymptotic result for Δ_{n,k_n} that involves the $\chi^2(d)$ distribution due to [Pinelis 1994] (from where the constant $9/(2e^3)$ appears). That gives the bound for the level.

If $\partial M \neq \emptyset$, the almost uniformity of the density and the regularity of the boundary ensure a.s. the existence of an observation X_{i_0} with $d(X_{i_0}, \partial M) = O(\ln n/n)$, and then condition K ($k_n/(\ln n)^4 \to +\infty$) ensures that $d(X_{i_0}, \partial M) \ll r_{i_0,k_n}$. Note that this condition is stronger than the usual $k_n \to +\infty$. The sample $\{r_{i_0,k_n}^{-1} X^*_{1(i_0)}, \ldots, r_{i_0,k_n}^{-1} X^*_{k_n(i_0)}\}$ thus "looks like" an uniform sample on a half d-dimensional unit ball and $\|\overline{X}_{k_n,i_0}\| r_{i_0,k_n}^{-1} \xrightarrow{a.s.} \alpha_{d'} > 0$. This consideration allows to obtain the power 1 for n large enough. □

A practical way for the choice of the parameter k_n is given making in [J1] making the test quite easy to practically implement.

3.3.3 Boundary estimation and estimation with boundary

3.3.3.1 Detecting Boundary Observations

Intuition In the full-dimensional case ($d = D$), data points close to the boundary may be identified by how (macroscopically) large their Voronoi cells tend to be [Rodríguez Casal 2007]. That is, if $\rho > 0$ is a detection radius, the *boundary observations* may be defined by

$$\mathcal{Y}_\rho = \{X_i \in \mathbb{X}_n \mid \exists O \in \mathbb{R}^D, \|O - X_i\| \geq \rho \text{ and } \mathring{\mathrm{B}}(O, \|O - X_i\|) \cap \mathbb{X}_n = \emptyset\}.$$

If X_i belongs to $\mathcal{Y}_\rho$ with associated $O \in \mathbb{R}^D$, then $\hat{\eta}_i := \frac{O - X_i}{\|O - X_i\|}$ appears to provide an consistent estimator of the unit outer normal vector of ∂M at $\pi_{\partial M}(X_i)$ [P2]. The present work leverages the above intuition and extends it to the case where M is a d-dimensional manifold with $d < D$. In fact, the manifold M not being full-dimensional raises the following additional subtleties:

- Even if X_i is far from ∂M, its Voronoi cell is large in the directions of $T_{X_i} M^\perp$, as it actually contains at least $X_i + \mathrm{B}_{T_{X_i} M^\perp}(0, \tau_{\min})$. To detect points close to the boundary *only*, we shall hence avoid these normal non-informative directions and solely focus on the tangential components of the Voronoi cells. For instance, by first projecting points onto (an estimate of) $T_{X_i} M$.
- If X_i is close to ∂M but M is folded over X_i, then the Voronoi cell of X_i in the Voronoi diagram of the projected sample might be small. To detect enough points close to the boundary, not all the sample should thus be projected, but rather just a neighborhood $\mathbb{X}_n \cap \mathrm{B}(X_i, R_0)$ of X_i, for some localization radius $R_0 > 0$ to be tuned.

These two remarks lead to the following first detection procedure: for a collection of estimated tangent spaces $\hat{T}_i$'s, one *may* label X_i as being a *boundary observation* if it has a large Voronoi cell within its R_0-neighborhood, when projected onto $X_i + \hat{T}_i$. That is, if there exists $O \in \hat{T}_i$ such that $\|O\| \geq \rho$ and $\mathring{\mathrm{B}}(O, \|O\|) \cap \pi_{\hat{T}_i}(\mathbb{X}_n \cap \mathrm{B}(X_i, R_0) - X_i) = \emptyset$. Unfortunately, when $1 < d < D$, this intuitive detection method is not sufficient to detect enough observations close to the boundary. This issue can be overcome by investigating *all* the Voronoi cells of $\pi_{\hat{T}_j}(X_i)$ for $X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n$, where r is a small scale parameter.

As it is now clear how critical the knowledge of tangent spaces is to build a Voronoi-based boundary detection scheme, let us first briefly detail how we estimate them.

3.3.3.2 Tangent Space Estimation

Following the ideas of [Aamari & Levrard 2019], we will estimate tangent spaces using local principal component analysis.

Definition 20 (Tangent Space Estimator)**.** *For $i \in \{1, \dots, n\}$ and $h > 0$, we introduce the local covariance matrix*

$$\hat{\Sigma}_i(h) := \frac{1}{n-1} \sum_{j \neq i} (X_j - X_i)(X_j - X_i)^t \mathbb{I}_{\mathrm{B}(X_i, h)}(X_j),$$

and define $\hat{T}_i$ as the linear span of the first d eigenvectors of $\hat{\Sigma}_i(h)$.

Note that $\hat{T}_i$ is a local estimator, in the sense that it is $\left((X_j - X_i)\mathbb{I}_{X_j \in \mathrm{B}(X_i,h)}\right)_{1 \leq j \leq n}$-measurable. For a suitable choice of h, the following proposition provides guarantees on the principal angle between $T_{X_i}M$ and $\hat{T}_i$. In what follows, given two linear subspaces $T, T' \subset \mathbb{R}^D$, the *principal angle* between them is

$$\angle(T, T') := \|\pi_T - \pi_{T'}\|_{\mathrm{op}},$$

where $\|A\|_{\mathrm{op}} := \sup_{\|x\| \leq 1} \|Ax\|$ stands for the operator norm of $A \in \mathbb{R}^{n \times n}$.

Proposition 4 (Tangent Space Estimation)**.** *Let $h = \left(C_d \frac{f_{\max}^4}{f_{\min}^5} \frac{\log n}{n-1}\right)^{\frac{1}{d}}$, for a large enough constant C_d. For n large enough so that $h \leq \frac{\tau_{\min}}{32} \wedge \frac{\tau_{\partial,min}}{3} \wedge \frac{\tau_{\min}}{\sqrt{d}}$, with probability larger than $1 - 2\left(\frac{1}{n}\right)^{\frac{2}{d}}$, we have*

$$\max_{1 \leq i \leq n} \angle(T_{X_i}M, \hat{T}_i) \leq C_d \frac{f_{\max}}{f_{\min}} \frac{h}{\tau_{\min}}.$$

Detection Method and Normal Vector Estimation Now, for a local (but macroscopic) scale $R_0 > 0$, a detection radius $\rho > 0$ and a local bandwidth $r > 0$, we compute the d-dimensional Voronoi diagrams of $(\pi_{\hat{T}_i}(\mathrm{B}(X_i, R_0) \cap \mathbb{X}_n - X_i))_{1 \leq i \leq n}$ and define our boundary observations detection procedure as follows.

Definition 21 (Boundary Observations). *For $i \in \{1, \dots, n\}$, we let $J_{R_0,r,\rho}(X_i)$ be the set of r-neighbors X_j of X_i for which X_i has a ρ-large Voronoi cell in the projected Voronoi diagram at X_j. That is, writing*

$$\mathrm{Vor}^{(j)}_{R_0,\rho}(X_i) := \left\{ O \in \hat{T}_j \,\middle|\, \mathring{\mathrm{B}}\big(O, \|O - \pi_{\hat{T}_j}(X_i - X_j)\|\big) \cap \pi_{\hat{T}_j}(\mathrm{B}(X_j, R_0) \cap \mathbb{X}_n - X_j) = \emptyset \right\},$$

we define

$$J_{R_0,r,\rho}(X_i) := \left\{ X_j \in \mathrm{B}(X_i, r) \cap \mathbb{X}_n \,\middle|\, \mathrm{Vor}^{(j)}_{R_0,\rho}(X_i) \cap \mathring{\mathrm{B}}_{\hat{T}_j}(\pi_{\hat{T}_j}(X_i - X_j), \rho)^c \neq \emptyset \right\}.$$

The set of boundary observations $\mathcal{Y}_{R_0,r,\rho} \subset \mathbb{X}_n$ *is then defined as the set of data points that have at least one such large Voronoi cell:*

$$\mathcal{Y}_{R_0,r,\rho} := \{X_i \in \mathbb{X}_n \mid J_{R_0,r,\rho}(X_i) \neq \emptyset\}. \tag{3.4}$$

Remark 4. *Detecting boundary observations requires to compute n Voronoi diagrams in dimension d. Note that this step does not depend on the ambient dimension D, and can run in parallel.*

This strategy also provides a natural way to estimate unit normal outward-pointing vectors. For this, given a boundary observation $X_i \in \mathcal{Y}_{R_0,r,\rho}$, we simply consider directions in which $\mathrm{Vor}^{(j)}_{R_0,\rho}(X_i)$ is ρ-wide (see Figure 3.2). A formal definition goes as follows.

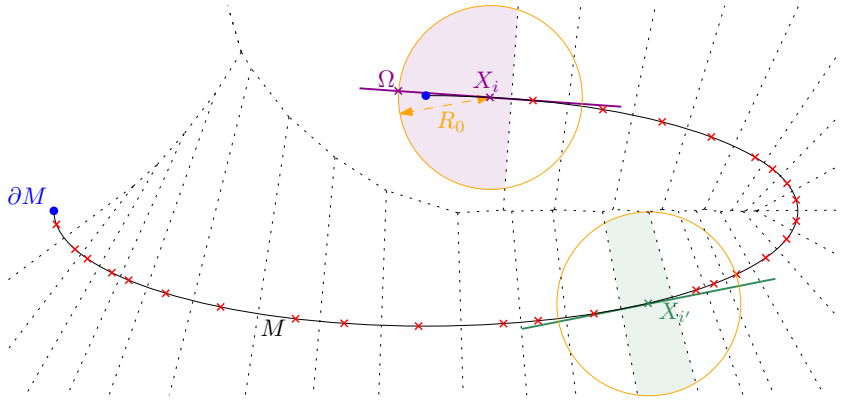

Figure 3.2: An ambient Voronoi diagram built on top of observations $\mathbb{X}_n$ lying on an open plane curve ($d = 1$, $D = 2$). The denser $\mathbb{X}_n$ in M, the narrower the Voronoi cell of the X_i's in the tangent directions $T_{X_i}M$. Observations close to ∂M yield cells that extend in the outward pointing direction. Localization radius $R_0 > 0$ prevents global foldings of M that would mix different ambient neighborhoods of M when projecting onto $T_{X_i}M$.

Definition 22 (Normal Vector Estimator). *For $X_i \in \mathcal{Y}_{R_0,r,\rho}$ and $X_j \in J_{R_0,r,\rho}(X_i)$, let*

$$\Omega^{(j)}_{R_0,r,\rho} \in \arg\min \left\{ \left\| \Omega - \pi_{\hat{T}_j}(X_i - X_j) \right\| \,\middle|\, \Omega \in \mathrm{Vor}^{(j)}_{R_0,\rho}(X_i) \cap \mathring{\mathrm{B}}_{\hat{T}_j}(\pi_{\hat{T}_j}(X_i - X_j), \rho)^c \right\}.$$

The estimator of the unit normal outward-pointing vector in $\hat{T}_j$ is defined by

$$\tilde{\eta}_i^{(j)} := \frac{\Omega^{(j)}_{R_0,r,\rho} - \pi_{\hat{T}_j}(X_i - X_j)}{\left\|\Omega^{(j)}_{R_0,r,\rho} - \pi_{\hat{T}_j}(X_i - X_j)\right\|}.$$

The final estimator of the unit outward-pointing normal vector at X_i is then defined by

$$\tilde{\eta}_i := \frac{1}{\#J_{R_0,r,\rho}(X_i)} \sum_{j \in J_{R_0,r,\rho}(X_i)} \tilde{\eta}_i^{(j)}. \tag{3.5}$$

As expected, when localization radii are chosen properly, Theorem below provides quantitative bounds for boundary detection and normal estimation.

Theorem 19 (Guarantees for Boundary Detection and Normals)**.** *ake $R_0 \leq \frac{\tau_{\min} \wedge \tau_{\partial,\min}}{40}$. Define*

$$r_- := \sqrt{(\tau_{\min} \wedge \tau_{\partial,\min})R_0}\left(c_d \frac{f_{\max}^5 \log n}{f_{\min}^6 n R_0^d}\right)^{\frac{1}{d+1}}, r_+ := \frac{R_0}{12}, \text{ and } \rho_- := \frac{R_0}{4} =: \frac{\rho_+}{2}$$

Then, for n large enough, with probability at least $1 - 4n^{-\frac{2}{d}}$, we have that for all $\rho \in [\rho_-, \rho_+]$ and $r \in [r_-, r_+]$:

1. *If $\partial M = \emptyset$, then $\mathcal{Y}_{R_0,r,\rho} = \emptyset$;*
2. *If $\partial M \neq \emptyset$ then:*
   (a) *For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,*
   $$\mathrm{d}(X_i, \partial M) \leq \frac{2r^2}{\tau_{\min} \wedge \tau_{\partial,\min}};$$
   (b) *For all $x \in \partial M$,*
   $$\mathrm{d}(x, \mathcal{Y}_{R_0,r,\rho}) \leq 3r;$$
   (c) *For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$,*
   $$\|\eta_{\pi_{\partial M}(X_i)} - \tilde{\eta}_i\| \leq \frac{20r}{\sqrt{R_0(\tau_{\min} \wedge \tau_{\partial,\min})}}.$$

A key quantity is the scale R_0, that needs to be carefully tuned in practice. Whenever prior information on the reaches $\tau_{\min}$ and $\tau_{\partial,min}$ is at hand, then choosing R_0 as large as $\frac{\tau_{\min}}{40} \wedge \frac{\tau_{\partial,\min}}{4}$ leads to near-optimal bounds. If no information on the reaches is at our disposal, choosing $R_0 = (\log n)^{-1}$ would meet the requirements for n large enough, while only adding extra $\log n$ terms in the bounds. The choice of r is less crucial: choosing $r = n^{-\alpha}$, with $\alpha \in \left(\frac{1}{d+1}, \frac{1}{d}\right)$ ensures that the requirements are fulfilled for n large enough, with no impact on the final bounds.

In a nutshell, Point 1. guarantees that no false positive occur if $\partial M = \emptyset$. On the other hand, if $\partial M \neq \emptyset$, for $\varepsilon \asymp (\log n/n)^{1/(d+1)}$, point 2.($b$) ensure that $\mathcal{Y}_{R_0,r,\rho}$ is an ε-covering

of ∂M that consists of points (ε^2/R_0)-close to ∂M. In the convex case $\tau_{\min} = \infty$, taking the convex hull of $\mathcal{Y}_{R_0,r,\rho}$ — similarly to [Dümbgen & Walther 1996] — would result in an (ε^2/R_0)-approximation of M, and the boundary of this convex hull in an (ε^2/R_0)-approximation of ∂M.

Still based on $\mathcal{Y}_{R_0,r,\rho}$, we extend this "hull" construction to the non-convex case by leveraging the additional tangential and normal estimates, to provide estimators of M and ∂M.

3.3.3.3 Boundary Estimation

Assume that $\partial M \neq \emptyset$. Then ∂M is a $(d-1)$-dimensional $\mathcal{C}^2$-sub-manifold without boundary. Therefore, using manifold estimators of [Aamari & Levrard 2018, Aamari & Levrard 2019, Maggioni *et al.* 2016] designed for the empty boundary case with input points $\mathcal{Y}_{R_0,r,\rho}$ seems relevant. We choose to focus on the manifold estimator proposed in [Aamari & Levrard 2018], based on the Tangential Delaunay Complex [Boissonnat & Ghosh 2014], as it also provides a topologically consistent estimation. This procedure, as well as the aforementioned two others, takes as input boundary points but also estimates of the tangent spaces (of the boundary). Thus, a preliminary step is to provide estimators for the boundary tangent spaces at points of $\mathcal{Y}_{R_0,r,\rho}$.

Definition 23 (Boundary's Tangent Space Estimator)**.** *For all $X_i \in \mathcal{Y}_{R_0,r,\rho}$, $\hat{T}_{\partial,i}$ is defined as the orthogonal complement of $\pi_{\hat{T}_i}(\tilde{\eta}_i)$ in $\hat{T}_i$. That is,*

$$\hat{T}_{\partial,i} := (\pi_{\hat{T}_i}(\tilde{\eta}_i))^{\perp} \cap \hat{T}_i.$$

A straightforward consequence of Property on tangent space estimation and Guarantee for Boundary and normal Theorem is that we easily can estimate tangent to the boundary spaces

Corollary 3 (Boundary's Tangent Space Estimation)**.** *Under the assumptions of previous property and theorem we have, for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$,*

$$\max_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \angle(T_{\pi_{\partial M}(X_i)}\partial M, \hat{T}_{\partial,i}) \leq \left(C_d \frac{f_{max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} R_0^d (n-1)} \right)^{\frac{1}{d+1}}.$$

We are now in position to provide an estimator for ∂M. Following [Aamari & Levrard 2018], we let $\varepsilon = R_0 \left(C_d \frac{f_{max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} R_0^d (n-1)} \right)^{\frac{1}{d+1}}$, and let $\mathbb{Y}_\partial$ denote an ε-sparsification of $\mathcal{Y}_{R_0,r,\rho}$, i.e. a subset of $\mathcal{Y}_{R_0,r,\rho}$ that forms an ε-covering of $\mathcal{Y}_{R_0,r,\rho}$ with points are ε-separated.

We also denote by $\mathbb{T}_\partial$ the collection of $\hat{T}_{\partial,i}$'s, for $X_i \in \mathbb{Y}_\partial$, and define our estimator of ∂M as the (weighted) Tangential Delaunay Complex [Boissonnat & Ghosh 2014] based on $(\mathbb{Y}_\partial, \mathbb{T}_\partial)$:

$$\widehat{\partial M} := \mathrm{Del}^{\omega_*}(\mathbb{Y}_\partial, \mathbb{T}_\partial).$$

Since ∂M has no boundary, [Aamari & Levrard 2018, Theorem 4.4] applies, and leads to the following reconstruction result.

Theorem 20 (Boundary Estimation – Upper Bound). *Provided that $\partial M \neq \emptyset$ and under the assumptions of previous results, we have for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$,*

1. $\mathrm{d_H}(\partial M, \widehat{\partial M}) \leq R_0 \left(C_d \frac{f_{max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} R_0^d (n-1)} \right)^{\frac{2}{d+1}}$,
2. *∂M and $\widehat{\partial M}$ are ambient isotopic.*

As a consequence, for n large enough, we have

$$\mathbb{E}\left[\mathrm{d_H}(\partial M, \widehat{\partial M})\right] \leq R_0 \left(C_d \frac{f_{max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} R_0^d (n-1)} \right)^{\frac{2}{d+1}}.$$

And the proposed estimator is minimax (up to a log) due to following result.

Theorem 21 (Boundary Estimation – Lower Bound). *Assume that $f_{\min} \leq c_d / \tau_{\partial,\min}^d$ and $c_d' / \tau_{\partial,\min}^d \leq f_{\max}$ for some small enough $c_d, (c_d')^{-1} > 0$. Then for all $n \geq 1$,*

$$\inf_{\hat{B}} \sup_{P \in \mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(f_{\min}, f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(\partial M, \hat{B}\big)\right] \geq C_d \tau_{\partial,\min} \left\{ 1 \wedge \left(\frac{1}{f_{\min} \tau_{\partial,\min}^d n} \right)^{\frac{2}{d+1}} \right\}.$$

3.3.3.4 Boundary-Adaptive Manifold Estimation

If $\partial M = \emptyset$, it is known that M can be estimated optimally by local linear patches [Aamari & Levrard 2019]. That is, choosing $\varepsilon_{\mathring{M}} = \left(C_d \frac{f_{\max}^4 \log n}{f_{\min}^5 n} \right)^{1/d}$, and estimating M via the union of tangential balls $\hat{M} = \bigcup_{i=1}^n X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ leads to $\mathrm{d_H}(M, \hat{M}) \leq C_d f_{\max} \varepsilon_{\mathring{M}}^2 / (f_{\min} \tau_{\min})$ [Aamari & Levrard 2019, Theorem 6], recovering the minimax rate $O\big((\log n / n)^{2/d}\big)$ over the class of $\mathcal{C}^2$ manifolds without boundary [Kim & Zhou 2015].

If $\partial M \neq \emptyset$ and X_i is close to ∂M, a tangential ball $X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ may go past ∂M along the normal direction $\eta_{\pi_{\partial M}(X_i)}$, leading to a poor approximation of M in terms of Hausdorff distance. In this case, replacing $X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}})$ by a tangential half-ball oriented at the opposite of the outward-pointing normal vector $\eta_{\pi_{\partial M}(X_i)}$ seems more appropriate. We formalize this intuition as follows.

Let $\mathcal{Y}_{R_0,r,\rho}$ denote the detected boundary observations of Definition 21. These points will generate half-balls, with radius $\varepsilon_{\partial M}$, that will roughly approximate the inward slab $M \cap \mathrm{B}(\partial M, \varepsilon_{\partial M})$ of radius $\varepsilon_{\partial M}$. To approximate the remaining part of M, we further define the *$\varepsilon_{\partial M}$-inner points* as

$$\mathring{\mathcal{Y}}_{\varepsilon_{\partial M}} := \{ X_i \in \mathbb{X}_n \mid \mathrm{d}(X_i, \mathcal{Y}_{R_0,r,\rho}) \geq \varepsilon_{\partial M}/2 \}. \tag{3.6}$$

Then, the manifold M may be reconstructed as follows.

Definition 24 (Boundary-Adaptive Manifold Estimator). *Given some inner and boundary radii parameters $\varepsilon_{\mathring{M}}$ and $\varepsilon_{\partial M}$, the manifold estimator $\hat{M}$ is defined by*

$$\hat{M} := \hat{M}_{\mathrm{Int}} \cup \hat{M}_{\partial},$$

where

$$\hat{M}_{\text{Int}} := \bigcup_{X_i \in \mathring{\mathcal{Y}}_{\varepsilon_{\partial M}}} X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\mathring{M}}),$$
$$\hat{M}_{\partial} := \bigcup_{X_i \in \mathcal{Y}_{R_0,r,\rho}} \Big(X_i + \mathrm{B}_{\hat{T}_i}(0, \varepsilon_{\partial M})\Big) \cap \{z, \langle z - X_i, \tilde{\eta}_i\rangle \leq 0\},$$

with

- *the $\hat{T}_i$'s being the estimated tangent spaces,*
- *the $\tilde{\eta}_i$'s being the estimated of the outward-pointing normals.*

Note that $\hat{M}$ is adaptive in the sense that it does not require information about emptiness of ∂M. If $\partial M = \emptyset$, then $\mathcal{Y}_{R_0,r,\rho} = \emptyset$ with high probability. In this case $\hat{M}$ coincides (with high probability) with the estimator from [Aamari & Levrard 2019], which is minimax over the class of boundariless $\mathcal{C}^2$-manifolds.

Theorem 22 (Estimation with Boundary – Upper Bound)**.** *Choose (R_0, r, ρ) as in first Theorem set*

$$\varepsilon_{\mathring{M}} = \left(C_d \frac{\log n}{f_{\min} n}\right)^{\frac{1}{d}} \text{ and } \varepsilon_{\partial M} = R_0 \left(C_d \frac{f_{max}^5}{f_{\min}^5} \frac{\log n}{f_{\min} R_0^d (n-1)}\right)^{\frac{1}{d+1}}.$$

Then for n large enough, with probability larger than $1 - 4n^{-\frac{2}{d}}$, we have

$$\mathrm{d_H}(M, \hat{M}) \leq C_d \begin{cases} (f_{\max}/f_{\min})^{\frac{4}{d}+1} \varepsilon_{\mathring{M}}^2 / \tau_{\min} & \text{if } \partial M = \emptyset, \\ \varepsilon_{\partial M}^2 / R_0 & \text{if } \partial M \neq \emptyset. \end{cases}$$

As a consequence, for n large enough,

$$\mathbb{E}\Big[\mathrm{d_H}(M, \hat{M})\Big] \leq C_d \begin{cases} \tau_{\min} \left(\dfrac{f_{\max}^{2+d/2}}{f_{\min}^{2+d/2}} \dfrac{\log n}{f_{\min} \tau_{\min}^d n}\right)^{\frac{2}{d}} & \text{if } \partial M = \emptyset, \\ R_0 \left(\dfrac{f_{\max}^5}{f_{\min}^5} \dfrac{\log n}{f_{\min} R_0^d n}\right)^{\frac{2}{d+1}} & \text{if } \partial M \neq \emptyset. \end{cases}$$

This method being minimax (up to a log factor).

Theorem 23 (Manifold Estimation – Lower Bounds)**.**

1. *(Boundaryless) Assume that $f_{\min} \leq c_d / \tau_{\min}^d$ and $c'_d / \tau_{\min}^d \leq f_{\max}$, for some small enough $c_d, (c'_d)^{-1} > 0$. If $d \leq D - 1$, then for all $n \geq 1$,*

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}^{d,D}_{\tau_{\min},\infty}(f_{\min}, f_{\max})} \mathbb{E}_{P^n}\Big[\mathrm{d_H}\big(M, \hat{M}\big)\Big] \geq C_d \tau_{\min} \left\{1 \wedge \left(\frac{1}{f_{\min} \tau_{\min}^d n}\right)^{\frac{2}{d}}\right\}.$$

2. *(Convex) Assume that* $f_{\min} \leq c_d/\tau_{\partial,\min}^d$ *and* $c'_d/\tau_{\partial,\min}^d \leq f_{\max}$*, for some small enough* $c_d, (c'_d)^{-1} > 0$*. Then for all* $n \geq 1$*,*

$$\inf_{\hat{M}} \sup_{P \in \mathcal{P}_{\infty,\tau_{\partial,\min}}^{d,D}(f_{\min},f_{\max})} \mathbb{E}_{P^n}\left[\mathrm{d_H}\big(M,\hat{M}\big)\right] \geq C_d \tau_{\partial,\min} \left\{1 \wedge \left(\frac{1}{f_{\min}\tau_{\partial,\min}^d n}\right)^{\frac{2}{d+1}}\right\}.$$

3.3.4 Volume Estimation

In this section the target is to estimate the d-dimensional Minkowski content of M, as given by

$$\lim_{\varepsilon \to 0} \frac{\mu_D(M \oplus \varepsilon B)}{\omega_{D-d}\varepsilon^{D-d}} = L_0(M) < \infty. \tag{3.7}$$

This is just (alongside with Hausdorff measure, among others) one of the possible ways to measure lower-dimensional sets; see [Mattila 1995] for background.

Regarding the estimation of lower-dimensional measures, with $d < D$, the available literature mostly concerns the problem of estimating $L_0(M)$, M being the boundary of some compact support S. The sample model is also a bit different, as it is assumed that we have sample points *inside and outside* S. Here, typically, $d = D - 1$; see, [Armendáriz *et al.* 2009], [Cuevas *et al.* 2007], [Cuevas *et al.* 2013], [Jimìenez & Yukich 2011].

Again, in the case $M = \partial S$ with $D = 2$, under the extra assumption of r-convexity for S, the consistency of the plug-in estimator $L_0(\partial C_r(\mathcal{X}_n))$ of $L_0(\partial S)$ is proved in [Cuevas *et al.* 2012] under the usual *inside* model (points taken on S). Finally, in [Berrendero *et al.* 2014], assuming an *outside* model (points drawn in $B(S,R) \setminus S$), estimators of $\mu_D(S)$ and $L_0(\partial S)$ are proposed, under the condition of *polynomial volume* for S

From the perspective of the above references, our contribution here (Th. 24 below) could be seen as a sort of lower-dimensional extension of the mentioned results of type $\mu_D(M_n) \to \mu_D(M)$ regarding volume estimation. But, obviously, in this case the Lebesgue measure μ_D must be replaced with a lower-dimensional counterpart, such as the Minkowski content (3.7).

Theorem 24. *Let* $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ *be an iid sample drawn according to an almost uniform distribution* P_X *supported by* $M \subset \mathbb{R}^D$ *a* d*-dimensional sub-manifold in the geometric model with corner (see 18). Let us take* r_n *such that* $r_n \to 0$ *and* $(\log(n)/n)^{1/d} = o(r_n)$*, then*

$$\frac{\mu\big(B(\mathbb{X}_n, r_n)\big)}{\omega_{D-d} r_n^{D-d}} - L_0(M) = \mathcal{O}\Big(\frac{\beta_n}{r_n} + r_n\Big),$$

where $\beta_n := d_H(\mathbb{X}_n, M) = \mathcal{O}\big(\log(n)/n\big)^{1/d}$.

Sketch of proof. First with classical technique we have $\beta_n := d_H(\mathbb{X}_n, M) = \mathcal{O}\big(\log(n)/n\big)^{1/d}$ e.a.s. Then the results directly comes from the double inclusion:

$$M \oplus (r_n - \beta_n)B \subset \mathbb{X}_n \oplus r_n B \subset M \oplus r_n B,$$

and the fact that the positive reach of M implies that when $x \leq \tau_M$, $|M \oplus xB|_d$ is a polynomial that can be written $|M \oplus xB|_d = \omega_{D-d} L_O(M) x^{D-d} + \ldots + q_D x^D$. □

Figure 3.3: The linear patch support estimator proposed in [P1] looks like pangolin or t shirt for children. Patches overlap too much to plug measure estimator, and there is no topological preservation.

3.3.5 Perspectives

3.3.5.1 Support estimator with topological guaranty in case of boundary

If the "linear" patch support estimator proposed in [P1] is adaptive, minimax with regard to the Hausdorff distance and has the "good"dimension, it is far from being satisfactory for a topological point of view (it looks like pangolin scales an is quite irregular, see Figure 3.3). Finding a support estimator that is minimax (for the Hausdorff distance) **and** is (eventually almost surely) homeomorphic to the support even when the boundary is not empty is (one of the) Graal(s) of the manifold estimation community.

Before achieving (or not) this ambitious goal one may first concentrate on a more realistic goal having manifold estimators that can be plugged to estimate $|M|_d$ and $|\partial M|_{d-1}$.

See Figure 3.4 to be convinced that, the sum of the volume of the patches widely over-estimates the volume.

3.3.5.2 Volume and surface estimation

When we aim to estimate $|M|_d$ the solution proposed in [J1] seems to have very poor convergence rates and, up to our knowledge there is no yet estimators for $|\partial M|_{d-1}$. The idea is to refine the linear patch estimator proposed in [P1] to reduce overlapping. The idea being to consider the following linear patches for X_i

$$P_i = (\hat{T}_i \cap \{x, \langle x, \hat{\eta}_{\pi_{\partial M}(X_i)}\rangle \leq 0\}) \cap (\mathrm{Vor}_{\pi_{\hat{T}_i}(\mathbb{X}_n \cap B(X_i, R_0) - X_i)}(0)) + X_i$$

and the following support and boundary estimators

$$\hat{M} = \bigcup_i P_i \text{ and } \widehat{\partial M} = \bigcup_i (P_i \cap \{x, \langle x, \hat{\eta}_{\pi_{\partial M}(X_i)}\rangle = 0\})$$

Preliminary calculus indicates that such an estimator has good behavior (minimax rates for manifold and boundary estimation but also we can plug it for volume and perimeter estimation) if the data is preliminarily sparsified (see Figure 3.4 to help representing $\hat{M}$

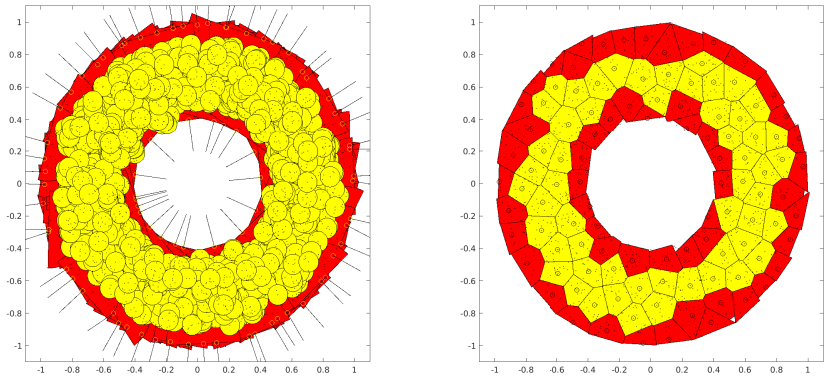

Figure 3.4: The original set is $B(O,1) \setminus B(O,0.5)$. The linear patches of [P1] (left) versus a Voronoi refinement on a scarification (indicated by o points) of the data (right). In red the "boundary" patches and in yellow the "interior" patches. Be convinced that the sum of the volumes of the linear patches (left) drastically overestimates the volume of the set, which is no more the case when patches are Voronoi-refined.

3.3.5.3 Density estimation on unknown manifold with possible boundary

Since [Hendriks 1990], estimation of densities supported by manifold have arisen a lot of attention. See for instance [Pelletier 2005], [Kim & Park 2013], [Berry & Sauer 2017] [Divol 2021] or [Berenfeld & Hoffmann 2021]. Most of the time (except for [Berry & Sauer 2017]) the manifolds are assumed to be without boundary. Knowledge of the manifold might be required (as in [Pelletier 2005]), sometime only the knowledge of the dimension [Kim & Park 2013]) is used and in, [Berenfeld & Hoffmann 2021] no preliminary knowledge on the manifold is necessary. We aim to propose and study a density estimator, that is convenient with the boundary case and only requires prior knowledge on the dimension of the manifold. Namely we propose a "local PCA projection" version of the density estimator proposed in 2.3.1 We propose to study the following density estimator:

$$\hat{f}_{r_n,A,d}(x) = \frac{N^o_{x,r_n}}{(n - N^\partial_{x,r_n})V_{x,r_n,d}} \mathbb{I}_{V_{x,r_n} \geq A\omega_D r^d} \mathbb{I}_{N^\partial_{x,r_n} \leq n/2}$$

Where $\pi_{\widehat{T_xM}}$ is an estimator of T_xM, $N_{x,r_n} = \#\{\mathbb{X}_n \cap B(x,r_n)\}$, $H_{x,r_n,d} = \mathcal{H}(\pi_{\widehat{T_xM}}(B(x,r_n) \cap \mathbb{X}_n))$, $V_{x,r_n} = |H_{x,r_n}|_d$, $N^\partial_{x,r_n} = \#\{\pi_{\widehat{T_xM}}(B(x,r_n) \cap \mathbb{X}_n) \cap \partial H_{x,r_n}\}$ and $N^o_{x,r_n} = N_{x,r_n} - N^\partial_{x,r_n}$.

3.4 Noisy lower dimensional context

Recall that, in this section we are interest in the following problem. Let $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ be iid observation drawn with a distribution $\mathbb{P}_X$ supported by $M \oplus rB$ where M is a d-dimensional manifold and we aim to get information on M. First, in [Niyogi *et al.* 2008]

it is proved that, under regularity condition (almost uniform distribution, positive reach τ_M and support included in $M \oplus rB$ with $r < \tau_M$) we can compute the homology of M from a Devroye-Wise estimator. We aim at obtaining more precise information on M such as a set $\mathbb{Y}_N$ included in $M \oplus \varepsilon_n B$ with $\varepsilon_n \to 0$ (so that we can apply the true lower dimensional methods -that mostly also works under such condition- with a possibly deteriorated rate). In [Genovese *et al.* 2012b] it has been proved that, under classical assumptions (positive reach, absence of boundary and almost uniform density) we have a minimax rate of order $n^{-\frac{2}{d+2}}$. In the same paper authors proposed a far from optimal, but computable, manifold estimator. The problem seems to have been very recently solved in [Aizenbud & Sober 2021] when data are uniformly drawn in the tubular neighborhood of a Manifold without boundary.

3.4.1 Estimation of the amount of noise

In [J5] we proposed two estimators for the the amount of noise. The first one, wich is based on the "boundary balls" of the Devroye-Wise estimator is the simplest. The second one which is based on the r-convex hull estimator has better rates.

Let $\rho_i = \sup_{z \in \mathrm{Vor}_{\mathbb{X}_n}(X_i)} ||z - X_i||$ and $\mathbb{B}_\rho = \{X_i, \rho_i \geq \rho\}$, define then $\hat{R}_\rho(\mathbb{X}_n) = \max_i d(X_i, \mathbb{B}_\rho)$

Theorem 25 (A., Cholaquidis and Cuevas 17)**.** *Suppose that M is a d-dimensional manifold with positive reach τ_M and that the distribution is almost uniform on $S = M \oplus RB$, where $R < \tau_M$ let $\varepsilon_n = (c \ln n/n)^{1/D}$ with $c \geq 6/(f_0 \omega_D)$ then, with probability one for n large enough $|R - \hat{R}_{\varepsilon_n}(\mathbb{X}_n)| \leq 2\varepsilon_n$*

Sketch of proof. The proof is based on the fact that we will have $d_H(\mathbb{X}_n, S) \leq \varepsilon_n$ and $d_H(\mathbb{B}_n, \partial S) \leq \varepsilon_n$. □

This theorem can be easily extend to the following one, suppose that $d_H(S, M) = R < \tau_M$, that S has the r_0-rolling ball condition for some positive r_0 and that the distribution is almost uniform then $\hat{R}_{\rho_n}(\mathbb{X}_n)$ is also a consistent estimator for R.

When such an extension is not required there exists a slightly better estimator of the amount of noise $\tilde{R}_\rho(\mathbb{X}_n) = \max_i d(X_i, C_\rho(\mathbb{X}_n))$ and we have

Theorem 26 (A., Cholaquidis and Cuevas 17)**.** *Suppose that M is a d-dimensional manifold with positive reach τ_M and that the distribution is almost uniform on $S = M \oplus RB$, where $R < \tau_M$ let $\rho \leq \tau_M - R$ then, with probability one for n large enough $|R - \tilde{R}_\rho(\mathbb{X}_n)| \leq O((\ln n/n)^{\min(\frac{1}{D-d}, \frac{2}{d+1})})$*

Sketch of proof. $|M \oplus a(\ln n/n)^{\frac{1}{D-d}}|_D = O(\ln n/n)$ that guarantees that, with probability one for n large enough there exists an observation (let say X_1) in $|M \oplus a(\ln n/n)^{\frac{1}{D-d}}|_D = O(\ln n/n)$ then using the property of the r-convex hull it comes that

$$d(X_1, \partial S) - O((\ln n/n)^{2/(d+1)} \leq d(X_1, \partial C_r(\mathbb{X}_n)) \leq d(X_1, \partial S)$$

giving the inequality $\hat{R} \geq R - O((\ln n/n)^{2/(d+1)} - O(a(\ln n/n)^{\frac{1}{D-d}})$. The reverse inequality being purely geometric. □

3.4.2 Denoising with use of reflexion on the boundary

Suppose that M is a d-dimensional manifold with positive reach τ_M and that the distribution is almost uniform on $S = M \oplus RB$, where $R < \tau_M$ and that we have also obtained:

1. $\hat{R}_n$ an estimator of R such that $|\hat{R}_n - R| \leq \varepsilon_n$
2. $\widehat{\partial S}_n$ a boundary estimator such that $d_H(\widehat{\partial S}_n, \partial S) \leq a_n$
3. $\mathbb{Y}_\lambda = \mathbb{X}_n \cap \{x, d(x, \widehat{\partial S}_n) \geq \lambda \hat{R}\}$

define $\phi : \mathbb{R}^D \to \mathbb{R}^D$ such that $\phi(y) = \pi_{\widehat{\partial S}_n}(y) + \hat{R}\frac{y - \pi_{\widehat{\partial S}_n}(y)}{\left\|y - \pi_{\widehat{\partial S}_n}(y)\right\|}$

Theorem 27 (A., Cholaquidis and Cuevas 17)**.** *Suppose that M is a d-dimensional manifold with positive reach τ_M and that the distribution is almost uniform on $S = M \oplus RB$, then for any given $\lambda \in]0, 1[$*

$$d_H(M, \phi(\mathbb{Y}_\lambda)) = O(\max(a_n^{1/3}, \varepsilon_n, d_H(\mathbb{X}_n, S)) \ \ e.a.s$$

sketch of proof. The fact that we apply ϕ on $\mathbb{Y}$ which is far enough from ∂S ensures that $\frac{y - \pi_{\widehat{\partial S}_n}(y)}{\left\|y - \pi_{\widehat{\partial S}_n(y)}\right\|}$ is close to $-\eta_{\pi_{\partial S}(y)}$. Hypothesis ensures that $\pi_{\widehat{\partial S}_n(y)}$ is close to $\pi_{\partial S}(y)$. Thus because $\pi_{\partial S}(y) - R\eta_{\pi_{\partial S}(y)} \in M$ (due to positive reach), by continuity we have $\phi(y)$ close to M. Also see Figure 3.5.

Every points $x \in M$ have an observation X_i at distance $O(\ln n/n)^{1/D}$ (by almost uniform assumption), for n large enough X_i is in $\mathbb{Y}_\lambda$ then $\phi(X_i)$ is close to x (as in the first part of the proof).

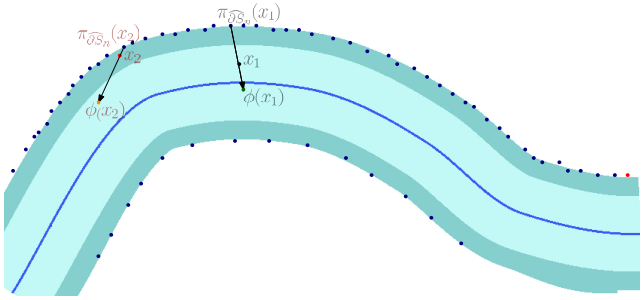

Figure 3.5: blue points are $\widehat{\partial S}_n$, $\phi(x_1)$ with $d(x_1, \widehat{\partial S}_n) \geq \lambda \hat{R}$ and $\phi(x_2)$ with $d(x_2, \widehat{\partial S}_n) \leq \lambda \hat{R}$

□

3.4.3 Denoising with use of the medial axis

We are going to generalize the results in [Genovese *et al.* 2012a] which make use of the medial axis for filament estimation ($D = 2$ and $d = 1$).

3.4.3.1 The medial axis

Let $S \subset \mathbb{R}^D$ be a compact set, its medial axis, introduced in [Blum 1967] as the set of points in $\mathbb{R}^D$ that has at least two different projections on ∂S (see Figure 3.6) has been initially proposed as a tool for biological shape recognition. Note that the medial axis can be decompose into two parts: its inner part, that is $\mathcal{M}(S) \cap S$ and its outer part that is $\mathcal{M}(S) \cap S^c$ (see Figure 3.6).

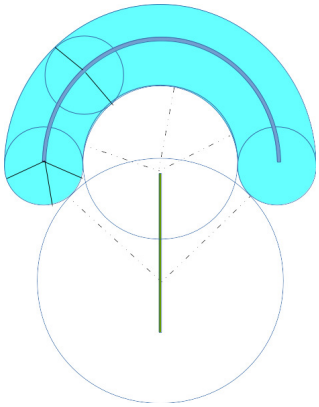

Figure 3.6: A set, its inner medial axis (blue) and its outer medial axis (green). Some points of the medial axis are presented together with some of there projection onto the boundary.

When dealing with compact sets we can only focus on $\mathcal{M}(S)$ the inner part of the medial axis.

Definition 25.

$$\mathcal{M}(S) = \{x \in S, diam(\Gamma(x)) > 0\}$$

where $\Gamma(x) = \{y \in \partial S, ||y - x|| = d(x, \partial S)\}$ *and* $diam(A) = \max\{||x - y||, (x, y) \in A^2\}$.

3.4.3.2 Link with manifold estimation

Indeed, suppose that Y is a random variable drawn on M a compact manifold with positive reach τ_M. also suppose that the distribution of Y is almost uniform. Now, we do not observe Y but X, that satisfies $X|Y \sim \mathcal{U}(B(O, r_Y))$ and denote by S the support of X. If $r_Y = r \leq \tau_M$ then $M = \mathcal{M}(S)$ that, we think can extend to, if $y \mapsto r_y$ is smooth enough and upper bounded by a $r < \tau_M$.

3.4.3.3 λ-medial axis

Unfortunately, the medial axis is difficult to estimate because it is not continuous with respect to the Hausdorff distance d_H. This is detailed in [Nagel] (see pages $217 - 238$) and illustrated in Figure 3.7 part a)). This implies that estimating the medial axis using a finite sample of points $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ can not be solved using classical plug-in methods (see

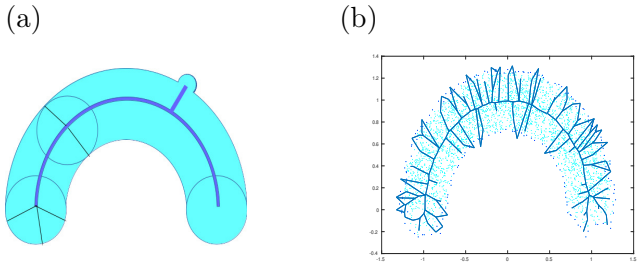

Figure 3.7: Two sets close to S the one of Figure 3.7 and their medial axis. (a) $S \cup \overline{B}(x, r_0)$ with $x \in \partial S$: a parasite branch appear whatever is the value of r_0 that illustrates the non continuity of the medial axis with regard to the Hausdorff distance. (b) plug-in estimator of the medial axis computed on a sample points there exist a lot of parasite branches.

Figure 3.7 part b)) and so provides a challenging problem that has been investigated in various papers (see [Attali *et al.*] for a state-of-the-art report).

Mainly two different approaches have been investigated. The first one consists in pruning the medial axis of an estimation of S (see, [Brandt & Algazi 1992] or [Attali & Montanvert 1996]); the second one consists in estimating the λ-medial axis defined as $\mathcal{M}_\lambda(S) = \{x \in \mathcal{M}(S), \Gamma(x) \subset \mathcal{B}(a, r) \Rightarrow r \geq \lambda\}$ instead of the medial axis. The λ-media axis has been introduce and studied in [Chazal & Lieutier 2005] where it has been proved to be stable with respect to the Hausdorff distance. More precisely the Authors prove that, if $d_H(S'^c, S^c) = O(\varepsilon)$ then $d_H(\mathcal{M}_\lambda(S), \mathcal{M}_\lambda(S')) = O(\sqrt{\varepsilon})$, then they propose an algorithm to estimate the λ-medial axis given sample points located near the boundary and prove that it converges.

Later on, given a sample point $\mathbb{X}_n$ drawn on S (instead of "near ∂S"), it is proved in [Cuevas *et al.* 2014], under no more shape hypothesis than regularity, that given a support estimator $\hat{S}_n$ such that $d_H(\hat{S}_n, S) \to 0$ a.s. and $d_H(\partial\hat{S}_n, \partial S) \to 0$ a.s. then $d_H(\mathcal{M}_\lambda(\hat{S}_n), \mathcal{M}_\lambda(S)) \to 0$ a.s.

3.4.3.4 The medial axis estimator and main result

We propose to study the following medial axis estimator, given $\hat{S}_n$ a support estimator and $\mathbb{Y}$ a subset of $\mathbb{X}_n$ define

$$\hat{\mathcal{M}}_\lambda(S_n, \mathbb{Y}) = \left\{ x \in \mathrm{Vor}_{\mathbb{Y}}(y) \cap \mathrm{Vor}_{\mathbb{Y}}(z) \cap \hat{S}_n, (y, z) \in \mathbb{Y}^2, ||y - z|| > \lambda \right\} \tag{3.8}$$

All the geometric assumptions made on S are listed in following Definition.

Definition 26. *Let $r_0 > 0$ and $K < 1$ be two numbers, S be a compact set in $\mathbb{R}^D$. We say S is (K, r_0)-regular if:*

1. *balls of radius r_0 roll freely inside and outside S;*
2. *$\mathcal{M}(S)$ is closed;*

3. for all $(x, y) \in \mathcal{M}(S)^2$, $|d(x, \partial S) - d(y, \partial S)|/||x - y|| \leq K$

Theorem 28. *Let* $\mathbb{X}_n = \{X_1 \dots X_n\} \subset \mathbb{R}^D$ *be an iid sample of points, drawn on* S *a* (K, r_0)*-regular compact. Assume that the distribution of* X *is almost uniform. For all* $r < r_0$ *denote by* $\hat{C}_r(\mathbb{X}_n)$ *the* r*-convex hull of* $\mathbb{X}_n$ *and put* $\mathbb{Y} = \partial \hat{C}_r(\mathbb{X}_n) \cap \mathbb{X}_n$.

There exists λ_0 *such that, for all* $0 < \lambda < \lambda_0$ *there exists* B_ρ *such that*

$$d_H(\mathcal{M}(S), \hat{\mathcal{M}}_\lambda(\hat{C}_r(\mathbb{X}_n), \mathbb{Y})) \leq B_\rho \left(\frac{\ln n}{n}\right)^{\frac{2}{D+1}} \quad e.a.s.$$

Remark: in the manifold estimation context, when $d = D - 1$ we have the minimax rate cf [Genovese *et al.* 2012b].

In Figure 3.8 we present some medial axis reconstruction with the proposed algorithm.

Note that in [J2] we propose a posteriori indicators to tune the different parameters of the medial axis estimator and more precisely the λ (the r parameter being possibly fully data driven as proposed in [Rodríguez-Casal & Saavedra-Nieves. 2019a]).

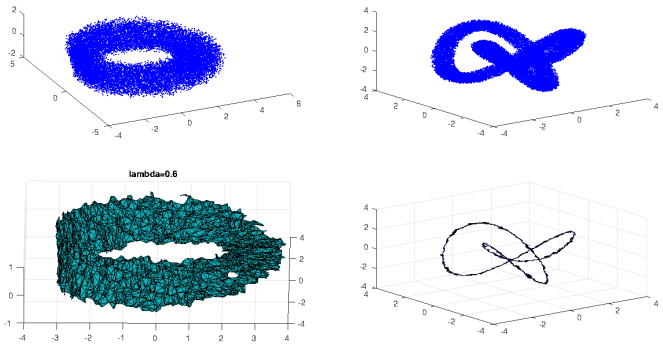

Figure 3.8: Manifold estimation with use of the medial axis. Left $D = 3$ and $d = 2$ data on a "noise" Moebus ring (top) and its medial axis reconstruction (down), right $D = 3$ and $d = 1$ data on a "noise" trefoil knot (top) and its medial axis reconstruction (down)

3.4.4 Perspectives

As mentionned in the introduction, finding a procedure that with a convergence rates depending on d and not D has been achieved in [Aizenbud & Sober 2021] but is not minimax and needs data are uniformly drawn in $S = M \oplus RB$ the tubular neighborhood of a manifold without boundary. Note that the proposed algorithm requires the amount of noise and the reach as inputs. The proposed amount of noise estimators of section 3.4.1 together with [Rodríguez-Casal, A. & Saavedra-Nieves, P. 2016] where an estimator of $\gamma = \sup_r C_r(S) = S$ allows to obtain a reach estimator since the geometric assumtions implies that $\tau_M = R + \gamma_0$.

Can we extend there estimator and find minimax denoising procedures adapted to manifolds without boundaries and non uniform samples (as in the reflexion on the boundary method) and possibly samples drawn in a subset of $S = M \oplus RB$ (as in the medial axis method)?.

3.5 General perspective: sub-manifold of a known manifold

A general perspective of works of set and manifold estimation consists in extending our results to the following purpose. Suppose that we know that data live in a known d'-dimensional sub-manifold $\mathcal{M}$ of $\mathbb{R}^D$ and the support M is now a d-dimensional sub-manifold of $\mathcal{M}$. We can either have $d' = d$ and then try to extend the set estimations methods and results or $d < d'$ and then try to extend the manifold estimations methods and results. This project requires to join the set/manifold estimation and the shape analysis communities. We expect that

1. Concerning the practical aspects the most trivial idea consists in replacing the Euclidean geometry by the Reamannian (on $\mathcal{M}$) one. We can generalize the Devroye Wise estimator by union of geodesic ball. Concerning other classical estimators such as r-convex hull or the r-shape, generalization is not that obvious when $d = 2$ (see [Boissonnat *et al.* 2018]) nevertheless we can try to find geometric assumptions allowing the generalization or explore the local convex hull idea.

2. Concerning the convergence rates we can expect to replace the dependence in D of the convergence rates by a dependence in d'. Morally we expect that the knowledge of $\mathcal{M}$ will “only"change the constant in the convergence rate. This is much more important that it looks since the constants are usually power of D (of number greater than one).

Chapter 4

Application to statistical learning

The link between Geometry, topology and data analysis is well known and is currently an attractive field of research (see [Carlsson 2009]). It started in the early 80′ in different application fields such as Neural Network [Kohonen 2004] for instance), Physical science [Grassberger & Procaccia 1983] or Informatics [Edelsbrunner *et al.* 1983] and arose from 2005 with, for instance the development of Persistent Homology [Singh *et al.* 2007b], [Zomorodian & Carlsson 2004] or [Carlsson *et al.* 2005] and the study of its asymptotic [Chazal *et al.* 2014], the mapper algorithm [Singh *et al.* 2007a], the Morse Theory [Bubenik *et al.* 2009].

In this chapter we won't aim at making an exhaustive state of the art (that should need another, much longer report) but only focus on the links between aforementioned set and manifold inference tools and Dimension reduction, Clustering or Classification. Even if most of them are yet well known or only perspective, it seems important to emphasize that the proposed tests and estimators are not only nice toys but may have many practical applications.

Notice first, that when data lies in a manifold, usual standardization, may hide the lower dimensional structure and that an appropriate normalization should be preliminary applied see [J8] where a local and iterative standardization process based on the idea of applying normalization on a neighborhoud data (i.e. on the $\mathbb{Y}_N = \{X_i - X_j,$ when i is a "neighbor" of $j\}$) (after each steps the neighborhoud may change which is the reason for iterations). If we were unable to have theoretical result on such a normalization process it has been empirically observed that it has nice performances.

4.1 Dimension reduction and some statistics in lower dimension

It is easy to see the link between dimension reduction and manifold estimation.

Many of the previously mentioned method may be used as preliminary tools before applying dimension reduction tools.

1. If the support (or a level set more robustness) is inferred to be convex then a simple well known PCA will do the job.

2. The estimation of the amount of noise given in section 3.3.1 quantify the distance between the data and a lower dimensional manifold and then is an a priori indicator of goodness of fit to a lower dimensional structure. Namely, having a "small" value of $\hat{R}/\text{diam}(\mathbb{X}_n)$ indicates the amount of noise from the sample to a lower dimensional

manifold is small with regard to the diameter of the sample and so that applying a dimension reduction method, is, a priori, a good idea.

3. After a denoising such as the medial axis based denoising we obtain points close to the wanted manifold, a dimension estimation on the "denoised data" (such as [Grassberger & Procaccia 1983] for one of the oldest reference of [Brito *et al.* 2013] or [Erba *et al.* 2019] for one of the newest ones) helps for tunning the "dimension parameter" of dimension reduction methods.

4. If the homology group (that can be determined by use of [Niyogi *et al.* 2008]) is not trivial then "classical" dimension reduction method (i.e. the one that send S into $A \subset \mathbb{R}^d$ with A which is d-dimensional won't work. It will be necessary, or to make projection in higher (than d dimension) or to apply reduction dimension method that allows to "cut" the data (such as [Lee *et al.* 2004])

They also can be used as a posteriori indicators

1. In [Delicado 2001] the authors proposed to apply there convexity test as an a posteriori way to tune the number of neighbors in the graph matrix used for geodesic distance computation in isomap algorithm (they namely propose to chose the smaller radius that provides a convex output).

2. By use of the estimation of the amount of noise given in section 3.3.1 we have an indication on the quality of the nonlinear dimension reduction method and particularly we can detect over-fitting.

4.1.1 Convergences rates for the geodesic distances

Since [Tenenbaum *et al.* 2000], a lot of dimension reduction methods [Demartines & Herault 1997], [Srivastava *et al.* 2008], [Lee *et al.* 2004] or [Lennon *et al.* 2002] are based on the use of the geodesic distance (instead of the euclidean one) between pairs of observations, that was the motivation to study, in [J4], derive the convergence rate of geodesic distance estimation.

When data are drawn according to a distribution that is supported by a path connected manifold M, the geodesic distance $\gamma(X_i, X_j)$ is the length of the shortest continuous path that links X_i to X_j, it is usually estimated by

$$\hat{\gamma}_{r_n}(X_i, X_j) = \min \left\{ \sum_{k=1}^{K} \|X_{i_{k+1}} - X_{i_k}\|, i_1 = i, i_K = j, \forall k \text{ s.t. } \|X_{i_{k+1}} - X_{i_k}\| \leq r_n \right\},$$

That can be computed using the Dijkstra's algorithm.

And because the method that uses the geodesic distances requires the computations of all the geodesic distances between every pairs of observations we are interest on the evaluation of

$$\max_{i,j} |\hat{\gamma}_{r_n}(X_i, X_j) - \gamma(X_i, X_j)|$$

In that purpose we will impose a regularity condition on M which is the following:

Definition 27. *Let $M \subset \mathbb{R}^D$ be a compact set, M is said to be K_M-geodesically smooth (later denoted as GS) for some positive number K_M if:*

1. *for all $(x, y) \in M^2$ there exists a geodesic path $\gamma_{x\to y}$ of class $\mathcal{C}^1$ that links x to y;*

2. *there exists a real function β such that $\lim_{t\to 0} \beta(t) = 0$ and $\forall (x, y) \in M^2$, $|\gamma_{x\to y}| \leq \beta(\|x - y\|)$;*

3. *let $\Gamma_{x\to y} : [0, |\gamma_{x\to y}|] \to \mathbb{R}^D$ be the parametrization of $\gamma_{x\to y}$ such that $\Gamma_{x\to y}(s)$ is the point of $\gamma_{x\to y}$ that is at a (curvilinear) distance s from x (along the geodesic curve). For all $(x, y) \in M^2$, the gradient of $\Gamma_{x\to y}$, denoted $\dot{\Gamma}_{x\to y}$, is K_M-Lipschitz continuous.*

Notice that a compact manifold of class $\mathcal{C}^2$ with a $\mathcal{C}^2$ boundary respects this hypothesis but many more sets (that even may not be manifold) can satisfies it.

Our main results consist in the following deterministic theorem.

Theorem 29 (A., Bodart 18)**.** *Assume that $d(\mathbb{X}_n, S) \longrightarrow 0$ and let (r_n) be a sequence such that $r_n > 2d(\mathbb{X}_n, S)$ and $d(\mathbb{X}_n, S)/r_n \longrightarrow 0$. Then,*

$$\max_{i,j} \left| |\hat{\gamma}_{r_n}(X_i, X_j)| - |\gamma_{X_i\to X_j}| \right| = O\left(\max\left(r_n, \frac{d(\mathbb{X}_n, S)^2}{r_n^2}\right)\right)$$

Which is optimal when choosing $r_n = O(d(\mathbb{X}_n, S))^{2/3})$ associated to a rate of order $d(\mathbb{X}_n, S))^{2/3}$.

When S is a compact d-dimensional manifold of $\mathbb{R}^D$ and the drawn is almost uniform we then obtain a rate of convergence of order $(\ln n/n)^{\frac{2}{3d}}$.

4.1.2 Perspectives and open questions

One can naturally wonder what is the minimax rate for geodesic distance estimation and if we can improve the proposed estimator.

We present here some possible way of improvement (that obviously can be combined).

1. "smoothing the curve" : ca we improve the rate with a smoothing (for instance with spines) of the piecewise linear estimation of the geodesic ?

2. Limit behavior when $r_n = O(d_H(S, \mathbb{X}_n))$: we conjecture than, when $r_n = cd_H(\S_n, \mathbb{X})$ the geodesic distance estimator is biased ($\mathbb{E}|\hat{\gamma}_{r_n}(x, y)| \to \alpha_{c,d}\gamma_{x\to y}$, $\alpha_{c,d}$ depending on the mean angle between tangent to $\hat{\gamma}_{r_n}(x, y)$ at t and tangent to $\gamma_{x\to y}$ "near t" as in [Jimìenez & Yukich 2011] or [Thäle & Yukich 2016]. Is this conjecture true, can this improve the rate ?

4.1.3 Application to Frechet mean estimation

In this section we assume the set M to be a compact d-manifold of class $\mathcal{C}^2$. Following the ideas of X. Pennec (see [Pennec 2006]), we consider the Fréchet expectations of the random variable X (which distribution is supported on M):

$$\mathbb{E}_k^{\mathrm{Fr}}(X) = \arg\min_{x\in M} \mathbb{E}(|\gamma_{x\to X}|^k), \quad k \in \mathbb{N}^*, \tag{4.1}$$

which are generalizations of the expected value for $k = 2$ and of the median (or depth) for $k = 1$. Note that Estimation of the Fréchet expectations has been widely studied when M is known (For instance, concerning CLT there is a first version in [Bhattacharya & Patrangenaru 2005] and many generalizations that can be found in [Bhattacharya & Bhattacharya. 2008, Bhattacharya & Lin 2016, Ellingson *et al.* 2013, Patrangenaru & Ellingson 2015, Eltzner & Huckemann 2018, Bhattacharya & Patrangenaru 2014]) and we are interested here to the case of unknown M.

As it is pointed out in [Pennec 2006], these expectations are not necessarily unique. For example, if M is a sphere and $\mathbb{P}_X$ the uniform distribution, then obviously all the points of M realize the minimum in (4.1) (for any $k \geq 1$).

To avoid dealing with such situations, we are going to make the following assumption, considering that k is fixed:

$$\begin{cases} \Phi(x) = \mathbb{E}(|\gamma_{x\to X}|^k) \text{ admits a unique minimum } x^* \in M, \\ \Phi \text{ is of class } \mathcal{C}^2 \text{ in a neighborhood of } x^*, \\ H_\Phi(x^*) \text{ is positive definite,} \end{cases} \tag{4.2}$$

where H_Φ denotes the Hessian matrix of Φ (i.e. $(H_\Phi)_{i,j} = \frac{\partial^2 \Phi}{\partial x_i, \partial x_j}$).

Remark: It must be noted that Φ is a continuous function on M. Indeed the triangle and Minkowski inequalities give $|\Phi(x)^{1/k} - \Phi(y)^{1/k}| \leq |\gamma_{x\to y}|$, for any $(x, y) \in M^2$. The extra (local) regularity in the conditions (4.2) is required for the sake of simplicity , allowing to apply basic differential calculus results at the optimal point x^*.

The first part of this assumption is very strong, but the second part is not. For example, when $d = 1$ and M is homeomorphic to a segment, explicit computations show that (4.2) holds for $k = 1$ if $f_X(x^*) \neq 0$. For $k = 2$, when M is a bounded closed convex set of dimension d, the geodesic distance on M coincides with the euclidean distance, the expectation $\mathbb{E}(X)$ lies in M, it minimizes the function $\Phi(x)$ and the condition (4.2) is satisfied (with $H_\Phi \equiv 2I_d$). This leads to think that, for $k = 2$, this condition is general enough and may hold for a wide class of regular sub manifolds of $\mathbb{R}^d$.

In this section we aim at studying the behavior of the natural estimator of $\mathbb{E}_k^{\mathrm{Fr}}(X)$:

$$\hat{\mathbb{E}}_{k,r_n}^{\mathrm{Fr}}(\mathbb{X}_n) = \arg\min_{X_i\in M} \frac{1}{n} \sum_j |\hat{\gamma}_{r_n}(X_i, X_j)|^k. \tag{4.3}$$

Theorem 30. *Assume that $M \subset \mathbb{R}^D$, $d \geq 2$ is a $d-$dimensional manifold, $d < D$ of class $\mathcal{C}^2$ with no boundary and that $\mathbb{P}_X$ is a probability distribution on M with continuous and*

bounded from below probability density f_X. *Moreover, suppose that assumption* (4.2) *holds. Then, choosing* $r_n = c(\max_i(\min_j ||X_i - X_j||))^{2/3}$ *in the definition of* $\hat{\gamma}_{r_n}$, *we have*

$$|\mathbb{E}_k^{Fr}(X) - \hat{\mathbb{E}}_{k,r_n}^{Fr}(\mathbb{X}_n)| = O\left(\left(\frac{\ln n}{n}\right)^{\min(1/4,1/3d)}\right) \quad e.a.s.$$

4.2 Clustering

The link between clustering and estimation of the level sets has been proposed and studied for long see [Biau *et al.* 2007] . Namely the method, which is really intuitive is to fix a level λ and to classify observation with respect to the connected component of $\hat{L}_\lambda$ they belong to. This method has the great advantage that the number of clusters is not an input. But... we need to tune λ, observations that do not belong to $\hat{L}_\lambda$ are not affected to any cluster, there might be "incompatible clusters". To solve this problem it has been proposed to study the level-set tree of the density (see [Rinaldo & Wasserman 2010], [Chaudhuri & Dasgupta 2010] for convergence rates or [Balakrishnan *et al.* 2013] for convergence rates when the density is supported by a manifold).

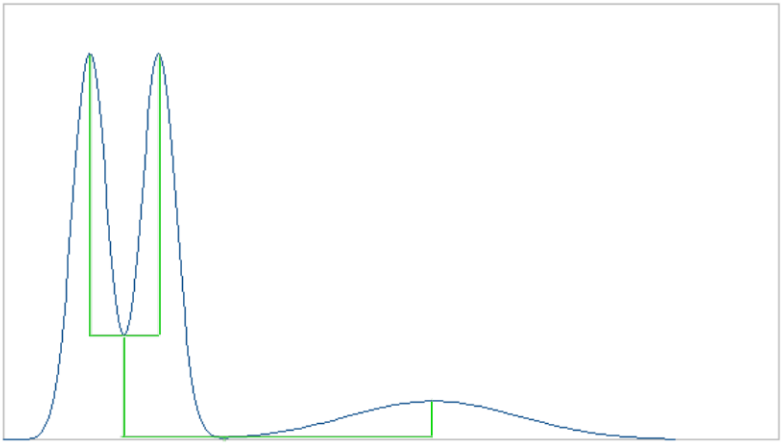

Figure 4.1: An example of density (blue) with 3 modes, when applying level set clustering results is clustering in 1 or 2 groups while the cluster tree (green) make the 3 groups associated to the 3 modes more visible

In practice the cluster tree is an improvement of the level set clustering when the density as "heterogeneous" components that can be problematic when estimating the density with kernel methods. Indeed the number of observation require to separate the denser modes (small window size) may create artificial irregularities for the sparser modes. An attempt to solve this problem has been proposed in [C1] has same asymptotics than the classical kernel method (proved) with better constant (empirically observed).

A way to overcome this heterogogeneity problem for density estimation is to consider nearest neighbors instead of kernel and, with no drastic revolution the density estimator proposed in 2.3.1 may help having level set trees that converges even when the density is

almost uniform and supported by a compact set and practically adapts to heterogeneous groups.

4.3 Classification : SVM with no Kernel Trick

Suppose that we observe $\mathbb{X}^+ = \{X_1^+, \dots, X_{n^+}^+\} \subset \mathbb{R}^d$ and $\mathbb{X}^- = \{X_1^-, \dots, X_{n^-}^-\} \subset \mathbb{R}^d$) two sets of iid observations drawn according to a distribution $\mathbb{P}_+$ supported by S_+ (resp. $\mathbb{P}_-$ supported by S_-). The classical two class classification problem consist in decide whether a new point X belongs to the "+" group or the "−" given its location. The Support Vector Machines (S.V.M.) (see [Vapnik 2000a] or [Vapnik 2000b]) is a method, based on margin maximization, that propose a solution to this problem. This method roughly relies on the following steps. First, consider the case where $\mathbb{X}^+$ and $\mathbb{X}^-$ are "linearly separable". Notice that this problem can be solved with only the knowledge of all the scalar products $\langle X_i^\pm, X_j^\pm \rangle$. Second, if the two groups are separable (but not linearly separable) transform the scalar product via a kernel function, hoping that it will send the data into an higher dimensional space in which the sets are separable (this step known as "the kernel trick"). Third, if the groups are not separable one can "soften" the margins.

Notice that, despite some debatable points: if linearly separable the problem may admit many solutions (See Figure 4.3 part 1), if linearly separable a linear separation might not be the best separation (see Figure 4.3 part 2), and mainly the lack of clear and rigorous justification of the the kernel trick step, the SVM appears to be a popular and efficient method.

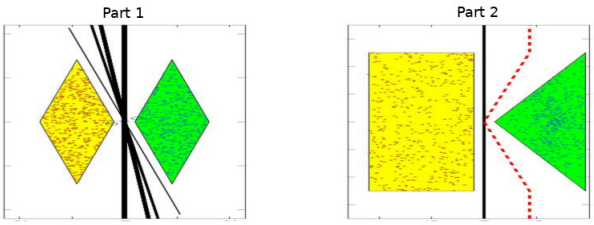

Figure 4.2: Two cases where the groups (yellow and green) are linearly separable, Part 1 (left) there is no uniqueness of the linear separation that maximizes the margin, Part 2 (right) we may prefer a non linear separation.

We would like to propose a new classification method that keeps the advantages of the SVM and that, we hope, has less fuzzy points.

Let us re-write the SVM method, under the separability hypothesis, in a geometrical purpose. Let φ be the transformation on the data induced by the change of scalar product. It is expected that we found a φ function such that $\varphi(S_+)$ and $\varphi(S_-)$ can be linearly separated. That is, $\mathcal{H}(\varphi(S_+)) \cap \mathcal{H}(\varphi(S_-)) = \emptyset$. The SVM algorithm then looks for the **hyper-plan** Δ that maximizes the margin $m = \min(d_{\min}(\mathcal{H}(\varphi(\mathbb{X}_+)), \Delta), d_{\min}(\mathcal{H}(\varphi(\mathbb{X}_-)), \Delta))$ (where $d_{\min}(A, B) = \min_{a \in A, b \in B}(||a - b||)$). Morally considering SVM we use a linear separation after a non linear transformation of the data. We propose to study a dual approach

where the separator might be not linear but there is non initial transformation on the data. Namely we propose to looks for the **hyper-surface** Δ that maximizes the margin $m = \min(d_{\min}(C_{r_+}(\mathbb{X}_+), \Delta), d_{\min}(C_{r_-}(\mathbb{X}_-), \Delta))$

Which lead to decide that x belongs to the $+$ (resp. $-$ group) if $d(x, C_{r_+}(\mathbb{X}_+)) < d(x, C_{r_-}(\mathbb{X}_-))$ (resp. $d(x, C_{r_+}(\mathbb{X}_+)) > d(x, C_{r_-}(\mathbb{X}_-))$)

Remarks:

1. Note that we still have the Support Vector idea since all the information is carried by boundary observations of $\mathbb{X}_+$ and $\mathbb{X}_-$.
2. Due to [Rodríguez-Casal & Saavedra-Nieves. 2019a] there is a fully data-driven way to tune parameters r_+ and r_- under r-convexity assumptions and almost uniformity.
3. That might be interesting to add some weights to the decision rule and to decide that x belongs to the $+$ (resp. $-$ group) if $(\mu_+)d(x, C_{r_+}(\mathbb{X}_+)) < (\mu_-)d(x, C_{r_-}(\mathbb{X}_-))$ (resp. $(\mu_+)d(x, C_{r_+}(\mathbb{X}_+)) > (\mu_-)d(x, C_{r_-}(\mathbb{X}_-))$)

We also have to extend the notion of soft margin. We nowadays see too possibilities

1. Introduce "inner depth", let introduce the scores

$$s(x, +) = d(x, C_{r_+}(\mathbb{X}_+))\mathbb{I}_{C_{r_+}(\mathbb{X}_+)}(x) - d(x, C_{r_+}(\mathbb{X}_+))\mathbb{I}_{C^c_{r_+}(\mathbb{X}_+)}(x)$$

$$s(x, -) = d(x, C_{r_-}(\mathbb{X}_-))\mathbb{I}_{C_{r_-}(\mathbb{X}_-)}(x) - d(x, C_{r_-}(\mathbb{X}_-))\mathbb{I}_{C^c_{r_-}(\mathbb{X}_-)}(x)$$

   And use a decision rule based on this (possibly weighted) scores.

2. Work on the "Level Set" instead of the support estimator, i.e. classify according to

$$S(x, \lambda_+, +) = d(x, C_{r_+}(\mathbb{X}_{+,\lambda_+}))\mathbb{I}_{C_{r_+}(\mathbb{X}_{+,\lambda_+})}(x) - d(x, C_{r_+}(\mathbb{X}_{+,\lambda_+}))\mathbb{I}_{C^c_{r_+}(\mathbb{X}_{+,\lambda_+})}(x)$$

$$S(x, \lambda_-, -) = d(x, C_{r_-}(\mathbb{X}_{-,\lambda_-})\mathbb{I}_{C_{r_-}(\mathbb{X}_{-,\lambda_-})}(x) - d(x, C_{r_-}(\mathbb{X}_{-,\lambda_-}))\mathbb{I}_{C^c_{r_-}(\mathbb{X}_{-,\lambda_-})}(x)$$

   with $\mathbb{X}_{+,\lambda_+} = \mathbb{X}_+ \cap \{x, \hat{f}_+(x) \geq \lambda\}$ and $\mathbb{X}_{-,\lambda_-} = \mathbb{X}_- \cap \{x, \hat{f}_-(x) \geq \lambda\}$ according to [Rodríguez-Casal & Saavedra-Nieves. 2019b].

4.4 Robust Fusion for big data

As mentioned all along this document most of the proposed set/manifold estimators suffer of a lack of robustness. They also have an heavy computational time. In [J3] it is noticed that, if parallelization may have small impact on results $\mathcal{L}^2$ error it may have impact more deeply the robustness, it is so necessary to find strategies to reduce this loss (applied to some far from geometric inference methods).

Bibliography

Scientific production

Prepints

[P1] Eddie Aamari, Catherine Aaron and Clément Levrard. *Minimax Boundary Estimation and Estimation with Boundary*. https://hal.uca.fr/UMR6620/hal-03317051v1. (Cited on pages 4, 15, 31, 35, 46 and 47.)

[P2] Catherine Aaron, Alejandro Cholaquidis and Ricardo Fraiman. *Surface Area Estimation*. https://hal.uca.fr/UMR6620/hal-02907297v4. (Cited on pages 2, 10, 26 and 38.)

Journal Papers

[J1] Catherine Aaron and Alejandro Cholaquidis. *On boundary detection*. Ann. Inst. Henri Poincaré Probab. Stat., vol. 56, no. 3, pages 2028–2050, 2020. (Cited on pages 3, 37, 38 and 46.)

[J2] Catherine Aaron. *Convergence rate for the λ-medial axis estimation under regularity condition*. Electron. J. Statist., vol. 13, pages 2686–2716, 2019. (Cited on pages 4 and 52.)

[J3] Catherine Aaron, Alejandro Cholaquidis, Ricardo Fraiman and Badih Ghattas. *Multivariate and functional robust fusion methods for structured Big Data*. J. Multivar. Anal., vol. 170, pages 149–161, 2019. (Cited on pages 5 and 61.)

[J4] Catherine Aaron and Olivier Bodart. *Convergence rates for estimators of geodesic distances and Fréchet expectations*. Journal of Applied Probability, vol. 55, page 1001–1013, 2018. (Cited on pages 5 and 56.)

[J5] Catherine Aaron, Alejandro Cholaquidis and Antonio Cuevas. *Detection of low dimensionality and data denoising via set estimation techniques*. Electron. J. Statist., vol. 11, pages 4596–4628, 2017. (Cited on pages 3, 4, 36 and 48.)

[J6] Catherine Aaron, Alejandro Cholaquidis and Ricardo Fraiman. *A generalization of the maximal-spacings in several dimensions and a convexity test*. Extremes., vol. 10, pages 605–634, 2017. (Cited on pages 1, 2, 7 and 12.)

[J7] Catherine Aaron and Olivier Bodart. *Local convex hull support and boundary estimation*. J. Multivariate Anal., vol. 147, pages 82–101, 2016. (Cited on pages 2, 17 and 18.)

[J8] Catherine Aaron. *Graph-based normalization and whitening for non-linear data analysis*. Neural networks, vol. 19, pages 864–876, 2006. (Cited on page 55.)

[J9] Catherine Aaron and Isabelle Bilon and Sebastien Galanti and Yamina Tadjeddine *Les styles de gestion de portefeuille existent-ils ?.* Revue d'Economie Financière, vol. 81, pages 171–188, 2005. (Not cited.)

[J10] Catherine Aaron and Sebastien Galanti and Yamina Tadjeddine *La gestion collective dans un marché agité : la dynamique des styles à partir des cartes de Kohonen.* Revue d'Economie Politique, vol. 81, pages 507–526, 2004. (Not cited.)

[J11] Catherine Aaron *Clustering under connectivity hypothesis.* Student, vol. 5, pages 43–58, 2004. (Not cited.)

Conferences (with proceedings)

[C1] Catherine Aaron. *Estimation de densité via un algorithme EM-Kernel.* In 42èmes Journées de Statistique, 2010. (Cited on page 59.)

[C2] Catherine Aaron and Corinne Perraudin and Joseph Rynkiewicz *Adaptation de l'algorithme SOM à l'analyse de données temporelles et spatiales: application à l'étude de l'évolution des performances en matière d'emploi.* proceedings of the Conference ASMDA, 480–488, 2005 (Not cited.)

[C3] Catherine Aaron and Corinne Perraudin and Joseph Rynkiewicz *Curves based Kohonen map and adaptative classification: an application to the convergence of the European Union countries.* proceedings of the Conference WSOM, 324–330, 2003 (Not cited.)

PhD Thesis

[T] Catherine Aaron. *Connexité et analyse des données non linéaires.* Université Paris I, 2005 (Not cited.)

Bibliography

[Aamari & Levrard 2018] Eddie Aamari and Clément Levrard. *Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction.* Discrete Comput. Geom., vol. 59, no. 4, pages 923–971, 2018. (Cited on pages 3, 31 and 42.)

[Aamari & Levrard 2019] Eddie Aamari and Clément Levrard. *Nonasymptotic rates for manifold, tangent space and curvature estimation.* Ann. Statist., vol. 47, no. 1, pages 177–204, 2019. (Cited on pages 2, 31, 35, 36, 37, 39, 42, 43 and 44.)

[Aamari *et al.* 2019] Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo and Larry Wasserman. *Estimating the reach of a manifold.* Electron. J. Stat., vol. 13, no. 1, pages 1359–1399, 2019. (Cited on pages 34 and 36.)

[Aizenbud & Sober 2021] Yariv Aizenbud and Barak Sober. *Non-Parametric Estimation of Manifolds from Noisy Data.* arXiv e-prints, page arXiv:2105.04754, 2021. (Cited on pages 48 and 52.)

[Alesker 2018] Seymon Alesker. *Some conjectures on intrinsic volumes of Riemannian manifolds and Alexandrov spaces.* Arnold Mathematical Journal, vol. 4, pages 1–17, 2018. (Cited on page 27.)

[Amenta *et al.* 2002] Nina Amenta, Sunghee Choi, Tamal K. Dey and Naveen Leekha. *A simple algorithm for homeomorphic surface reconstruction.* Internat. J. Comput. Geom. Appl., vol. 12, no. 1-2, pages 125–141, 2002. 16th Annual Symposium on Computational Geometry (Kowloon, 2000). (Cited on page 31.)

[Andea *et al.* 2004] Aleodor A. Andea, David L. Bouwman, Tracie Wallis and Daniel W. Visscher. *Correlation of tumor volume and surface area with lymph node status in patients with multifocal/multicentric breast carcinoma.* Cancer, vol. 100, 2004. (Cited on page 25.)

[Arias-Castro & Rodríguez-Casal 2017] Ery Arias-Castro and A. Rodríguez-Casal. *On estimating the perimeter using the alpha-shape.* Ann. Inst. H. Poincaré Probab. Statist., vol. 53, pages 1051–1068, 2017. (Cited on pages 2, 26 and 29.)

[Arias-Castro *et al.* 2019] Ery Arias-Castro, B. Pateiro-López and A. Rodríguez-Casal. *Minimax estimation of the volume of a set with smooth boundary.* Journal of the American Statistical Association, vol. 114, pages 1162–1173, 2019. (Cited on pages 10 and 25.)

[Armendáriz *et al.* 2009] I. Armendáriz, A. Cuevas and R. Fraiman. *Nonparametric estimation of boundary measures and related functionals: asymptotic results.* Adv. in Appl. Probab., vol. 41, pages 311–322, 2009. (Cited on page 45.)

[Attali & Montanvert 1996] Dominique Attali and Annick Montanvert. *Modeling noise for a better simplification of skeletons.* In Proc. of 3rd IEE Internat. Conf. Image Process, 1996. (Cited on page 51.)

[Attali *et al.*] D. Attali, J. Boissonnat and E. Edelsbrunner. Mathematical foundations of scientific visualization, computer graphics, and massive data exploration. Springer. (Cited on page 51.)

[Baíllo *et al.* 2001] Amparo Baíllo, Juan A. Cuesta-Albertos and Antonio Cuevas. *Convergence rates in nonparametric estimation of level sets.* Statistics and Probability Letters, vol. 53, no. 1, pages 27–35, May 2001. (Cited on page 21.)

[Baíllo 2003] Amparo Baíllo. *Total error in a plug-in estimator of level sets.* Statistics and Probability Letters, vol. 65, no. 4, pages 411–417, 2003. (Cited on page 21.)

[Balakrishnan *et al.* 2013] Sivaraman Balakrishnan, Srivatsan Narayanan, Alessandro Rinaldo, Aarti Singh and Larry Wasserman. *Cluster Trees on Manifolds.* In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 26. Curran Associates, Inc., 2013. (Cited on page 59.)

[Baldin & Reiß 2016] Nikolay Baldin and Markus Reiß. *Unbiased estimation of the volume of a convex body.* Stochastic Processes and their Applications, vol. 126, no. 12, pages 3716–3732, 2016. In Memoriam: Evarist Giné. (Cited on pages 10, 23 and 25.)

[Baíllo *et al.* 2000] Amparo Baíllo, Antonio Cuevas and Ana Justel. *Set Estimation and Nonparametric Detection.* The Canadian Journal of Statistics / La Revue Canadienne de Statistique, vol. 28, no. 4, pages 765–782, 2000. (Cited on pages 1 and 16.)

[Beermann & Reitzner 2015] M. Beermann and M. Reitzner. *Beyond the Efron–Buchta Identities: Distributional Results for Poisson Polytopes.* Discrete Comput Geom, vol. 53, page 226–244, 2015. (Cited on page 7.)

[Berenfeld & Hoffmann 2021] Clément Berenfeld and Marc Hoffmann. *Density estimation on an unknown submanifold.* Electron. J. Statist., vol. 15, 2021. (Cited on page 47.)

[Berenfeld *et al.* 2021] Clément Berenfeld, John Harvey, Marc Hoffmann and Krishnan Shankar. *Estimating the Reach of a Manifold via its Convexity Defect Function.* Discrete & Computational Geometry, Jun 2021. (Cited on page 36.)

[Berrendero *et al.* 2014] José R. Berrendero, Alejandro Cholaquidis, Antonio Cuevas and Ricardo Fraiman. *A geometrically motivated parametric model in manifold estimation.* Statistics, vol. 48, pages 1004 – 983, 2014. (Cited on page 45.)

[Berry & Sauer 2017] Tyrus Berry and Timothy Sauer. *Density estimation on manifolds with boundary.* Comput. Statist. Data Anal., vol. 107, pages 1–17, 2017. (Cited on page 47.)

[Bhattacharya & Bhattacharya. 2008] Abhishek Bhattacharya and Rabi Bhattacharya. *Statistics on Riemannian Manifolds: Asymptotic Distribution and Curvature*. In Proceedings of the American Mathematical Society,, pages 2959–2964, 2008. (Cited on page 58.)

[Bhattacharya & Lin 2016] Abhishek Bhattacharya and Lizhen. Lin. *Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces*. In Proceedings of the American Mathematical Society,, pages 413–428, 2016. (Cited on page 58.)

[Bhattacharya & Patrangenaru 2005] Rabi Bhattacharya and Vic Patrangenaru. *Large sample theory of intrinsic and extrinsic sample means on manifolds—II*. The Annals of Statistics, vol. 33, no. 3, pages 1225 – 1259, 2005. (Cited on page 58.)

[Bhattacharya & Patrangenaru 2014] R. Bhattacharya and V. Patrangenaru. *Statistics on manifolds and landmarks based image analysis: A nonparametric theory with applications*. Journal of Statistical Planning and Inference, vol. 145, pages 1–22, 2014. (Cited on page 58.)

[Biau *et al.* 2007] Gérard Biau, Benoît Cadre and Bruno Pelletier. *A graph-based estimator of the number of clusters*. ESAIM: Probability and Statistics, vol. 11, pages 272–280, 2007. (Cited on pages 22 and 59.)

[Biau *et al.* 2008] Gérard Biau, Benoît Cadre and Bruno Pelletier. *Exact rates in density support estimation*. Journal of Multivariate Analysis, vol. 99, no. 10, pages 2185–2207, 2008. (Cited on pages 1 and 16.)

[Biau *et al.* 2009] Gérard Biau, Benoit Cadre, David Mason and Bruno Pelletier. *Asymptotic Normality in Density Support Estimation*. Electronic Journal of Probability, vol. 14, no. none, pages 2617 – 2635, 2009. (Cited on pages 1 and 16.)

[Blum 1967] Harry Blum. *A Transformation for Extracting New Descriptors of Shape*. In Weiant Wathen-Dunn, editor, Models for the Perception of Speech and Visual Form, pages 362–380. MIT Press, Cambridge, 1967. (Cited on page 50.)

[Boissonnat & Ghosh 2014] Jean-Daniel Boissonnat and Arijit Ghosh. *Manifold reconstruction using tangential Delaunay complexes*. Discrete Comput. Geom., vol. 51, no. 1, pages 221–267, 2014. (Cited on pages 31 and 42.)

[Boissonnat *et al.* 2018] Jean-Daniel Boissonnat, Ramsay Dyer, Arijit Ghosh and Nikolay Martynchuk. *An Obstruction to Delaunay Triangulations in Riemannian Manifolds*. Discrete & Computational Geometry, vol. 59, pages 226–237, 2018. (Cited on page 53.)

[Bräker & Hsing 1998] H. Bräker and T. Hsing. *On the area and perimeter of a random convex hull in a bounded convex set*. Probability Theory and Related Fields, vol. 111, pages 517–550, 1998. (Cited on page 10.)

[Brandt & Algazi 1992] Jonathan W. Brandt and V. Ralph Algazi. *Continuous skeleton computation by Voronoi diagram.* CVGIP Image Underst., vol. 55, pages 329–338, 1992. (Cited on page 51.)

[Bredon 1993] Glen E. Bredon. Topology and geometry, volume 139 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1993. (Cited on page 33.)

[Brito *et al.* 2013] M. R. Brito, A. Quiroz and J. Yukich. *Intrinsic dimension identification via graph-theoretic methods.* J. Multivar. Anal., vol. 116, pages 263–277, 2013. (Cited on page 56.)

[Brunel *et al.* 2021] Victor-Emmanuel Brunel, Jason M. Klusowski and Dana Yang. *Estimation of convex supports from noisy measurements.* Bernoulli, vol. 27, no. 2, pages 772 – 793, 2021. (Cited on page 21.)

[Brunel 2013] Victor-Emmanuel Brunel *A universal deviation inequality for random polytopes.* https://hal.archives-ouvertes.fr/hal-00903687v2/document, 2013. (Cited on pages 7 and 11.)

[Brunel 2016a] Victor-Emmanuel Brunel. *Adaptive estimation of convex and polytopal density support.* Probability Theory and Related Fields, vol. 164, pages 1–16, 2016. (Cited on page 11.)

[Brunel 2016b] Victor-Emmanuel Brunel. *Concentration of the empirical level sets of Tukey's halfspace depth.* Probability Theory and Related Fields, vol. 173, pages 1165–1196, 2016. (Cited on page 20.)

[Brunel 2020] Victor-Emmanuel Brunel. *Deviation inequalities for random polytopes in arbitrary convex bodies.* Bernoulli, vol. 26, no. 4, pages 2488 – 2502, 2020. (Cited on pages 7, 11 and 24.)

[Brécheteau & Levrard 2020] Claire Brécheteau and Clément Levrard. *A k-points-based distance for robust geometric inference.* Bernoulli, vol. 26, no. 4, pages 3017 – 3050, 2020. (Cited on page 21.)

[Bubenik *et al.* 2009] Peter Bubenik, Gunnar E. Carlsson, Peter T. Kim and Zhiming Luo. *Statistical topology via Morse theory, persistence and nonparametric estimation.* arXiv: Statistics Theory, 2009. (Cited on page 55.)

[Butucea *et al.* 2007] Cristina Butucea, Mathilde Mougeot and Karine Tribouley. *Functional approach for excess mass estimation in the density model.* Electronic Journal of Statistics, vol. 1, no. none, pages 449 – 472, 2007. (Cited on page 22.)

[Bárány 1992] Imre Bárány. *Random polytopes in smooth convex bodies.* Mathematika, vol. 39, no. 1, page 81–92, 1992. (Cited on pages 7 and 10.)

[Cadre *et al.* 2013] Benoît Cadre, Bruno Pelletier and Pierre Pudlo. *Estimation of density level sets with a given probability content.* Journal of Nonparametric Statistics, vol. 25, no. 1, pages 261–272, 2013. (Cited on page 21.)

[Cadre 2005] Benoît Cadre. *Kernel estimation of density level sets.* Journal of Multivariate Analysis, vol. 97, pages 999–1023, 2005. (Cited on page 22.)

[Carlsson *et al.* 2005] Gunnar E. Carlsson, Afra Zomorodian, Anne D. Collins and Leonidas J. Guibas. *Persistence Barcodes for Shapes.* Int. J. Shape Model., vol. 11, pages 149–188, 2005. (Cited on page 55.)

[Carlsson 2009] Gunnar E. Carlsson. *Topology and data.* Bulletin of the American Mathematical Society, vol. 46, pages 255–308, 2009. (Cited on page 55.)

[Charpentier & Gallic 2015] Arthur Charpentier and Ewen Gallic. *Kernel density estimation based on Ripley's correction.* GeoInformatica, vol. 20, pages 95–116, 2015. (Cited on page 10.)

[Chaudhuri & Dasgupta 2010] Kamalika Chaudhuri and Sanjoy Dasgupta. *Rates of Convergence for the Cluster Tree.* In Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1, NIPS'10, page 343–351, Red Hook, NY, USA, 2010. Curran Associates Inc. (Cited on pages 5 and 59.)

[Chazal & Lieutier 2005] Frédéric Chazal and André Lieutier. *The "λ-medial axis".* Graphical Models, vol. 67, no. 4, pages 304–331, 2005. (Cited on page 51.)

[Chazal *et al.* 2014] Frédéric Chazal, Marc Glisse, Catherine Labruère and Bertrand Michel. *Convergence rates for persistence diagram estimation in Topological Data Analysis.* In Eric P. Xing and Tony Jebara, editors, Proceedings of the 31st International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, pages 163–171, Bejing, China, 22–24 Jun 2014. PMLR. (Cited on page 55.)

[Chen *et al.* 2015] Yen-Chi Chen, Christopher R. Genovese and Larry A. Wasserman. *Density Level Sets: Asymptotics, Inference, and Visualization.* Journal of the American Statistical Association, vol. 112, pages 1684 – 1696, 2015. (Cited on page 22.)

[Chevalier 1976] Jacques Chevalier. *Estimation du support et du contour du support d'une loi de probabilité.* Annales de l'I.H.P. Probabilités et statistiques, vol. 12, no. 4, pages 339–364, 1976. (Cited on pages 1 and 15.)

[Cholaquidis & Cuevas 2020] Alejandro Cholaquidis and Antonio Cuevas. *Set estimation under biconvexity restrictions.* ESAIM: PS, vol. 24, pages 770–788, 2020. (Cited on pages 2 and 17.)

[Cholaquidis *et al.* 2014] Alejandro Cholaquidis, Antonio Cuevas and Ricardo Fraiman. *ON POINCARÉ CONE PROPERTY.* Annals of Statistics, vol. 42, pages 255–284, 2014. (Cited on pages 2 and 17.)

[Cuevas & Rodríguez-Casal 2004] Antonio Cuevas and Alberto Rodríguez-Casal. *On boundary estimation.* Adv. in Appl. Probab., vol. 36, no. 2, pages 340–354, 2004. (Cited on page 15.)

[Cuevas & Rodríguez-Casal 2004] Antonio Cuevas and Alberto Rodríguez-Casal. *On Boundary Estimation.* Advances in Applied Probability, vol. 36, no. 2, pages 340–354, 2004. (Cited on page 7.)

[Cuevas *et al.* 2007] Antonio Cuevas, Ricardo Fraiman and Alberto Rodríguez-Casal. *A Nonparametric Approach to the Estimation of Lengths and Surface Areas.* The Annals of Statistics, vol. 35, no. 3, pages 1031–1051, 2007. (Cited on pages 25 and 45.)

[Cuevas *et al.* 2012] A. Cuevas, R. Fraiman and B. Pateiro-López. *On Statistical Properties of Sets Fulfilling Rolling-Type Conditions.* Advances in Applied Probability, vol. 44, pages 311 – 329, 2012. (Cited on page 45.)

[Cuevas *et al.* 2013] Antonio Cuevas, Ricardo Fraiman and László Györfi. *Towards a universally consistent estimator of the Minkowski content.* Esaim: Probability and Statistics, vol. 17, pages 359–369, 2013. (Cited on pages 25 and 45.)

[Cuevas *et al.* 2014] Antonio Cuevas, Pamela Llop and Beatriz Pateiro-López. *On the estimation of the medial axis and inner parallel body.* Journal of Multivariate Analysis, vol. 129, pages 171–185, 2014. (Cited on page 51.)

[Cuevas 1990] Antonio Cuevas. *On Pattern Analysis in the Non-Convex Case.* Kybernetes, vol. 19, pages 26–33, 1990. (Cited on pages 8 and 9.)

[Delicado *et al.* 2014] Pedro Delicado, Adolfo Hernández and Gábor Lugosi. *Data-based decision rules about the convexity of the support of a distribution.* Electronic Journal of Statistics, vol. 8, no. 1, pages 96 – 129, 2014. (Cited on page 11.)

[Delicado 2001] P. Delicado. *Another Look at Principal Curves and Surfaces.* Journal of Multivariate Analysis, vol. 77, pages 84–116, 2001. (Cited on page 56.)

[Demartines & Herault 1997] P. Demartines and J. Herault. *Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets.* IEEE Transactions on Neural Networks, vol. 8, no. 1, pages 148–154, 1997. (Cited on page 56.)

[Devroye & Wise 1980] Luc Devroye and Gary L. Wise. *Detection of Abnormal Behavior Via Nonparametric Estimation of the Support.* SIAM Journal on Applied Mathematics, vol. 38, no. 3, pages 480–488, 1980. (Cited on pages 1 and 15.)

[Divol 2020] Vincent Divol. *Minimax adaptive estimation in manifold inference.* https://arxiv.org/abs/2001.04896, 2020. (Cited on pages 3, 18 and 31.)

[Divol 2021] Vincent Divol. *Reconstructing measures on manifolds: an optimal transport approach.* https://arxiv.org/abs/2102.07595, 2021. (Cited on page 47.)

[Dümbgen & Walther 1996] Lutz Dümbgen and Günther Walther. *Rates of convergence for random approximations of convex sets.* Advances in Applied Probability, vol. 28, no. 2, page 384–393, 1996. (Cited on pages 7, 8, 15, 35 and 42.)

[Edelsbrunner & Shah 1994] Herbert Edelsbrunner and Nimish R. Shah. *Triangulating Topological Spaces*. In Proceedings of the Tenth Annual Symposium on Computational Geometry, SCG '94, page 285–292, New York, NY, USA, 1994. Association for Computing Machinery. (Cited on page 31.)

[Edelsbrunner *et al.* 1983] H. Edelsbrunner, D. Kirkpatrick and R. Seidel. *On the shape of a set of points in the plane*. IEEE Transactions on Information Theory, vol. 29, no. 4, pages 551–559, 1983. (Cited on pages 1 and 55.)

[Efron 1965] Bradley Efron. *The Convex Hull of a Random Set of Points*. Biometrika, vol. 52, no. 3/4, pages 331–343, 1965. (Cited on pages 7 and 10.)

[Ellingson *et al.* 2013] Leif Ellingson, V. Patrangenaru and F. Ruymgaart. *Nonparametric estimation of means on Hilbert manifolds and extrinsic analysis of mean shapes of contours*. J. Multivar. Anal., vol. 122, pages 317–333, 2013. (Cited on page 58.)

[Eltzner & Huckemann 2018] Benjamin Eltzner and Stephan F. Huckemann. *A smeary central limit theorem for manifolds with application to high-dimensional spheres*. arXiv: Statistics Theory, 2018. (Cited on page 58.)

[Erba *et al.* 2019] Vittorio Erba, Marco Gherardi and Pietro Rotondo. *Intrinsic dimension estimation for locally undersampled data*. Scientific Reports, vol. 9, 2019. (Cited on page 56.)

[Federer 1959] Herbert Federer. *Curvature measures*. Trans. Amer. Math. Soc., vol. 93, pages 418–491, 1959. (Cited on pages 34 and 35.)

[Federer 1969] Herbert Federer. Geometric measure theory. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969. (Cited on page 26.)

[Fefferman *et al.* 2013] Charles Fefferman, Sanjoy K. Mitter and Hariharan Narayanan. *Testing the Manifold Hypothesis*. arXiv: Statistics Theory, 2013. (Cited on page 36.)

[Funke & Kawka 2015a] Benedikt Funke and Rafael Kawka. *Nonparametric density estimation for multivariate bounded data using two non-negative multiplicative bias correction methods*. Computational Statistics & Data Analysis, vol. 92, pages 148–162, 2015. (Cited on page 10.)

[Funke & Kawka 2015b] Benedikt Funke and Rafael Kawka. *Nonparametric density estimation for multivariate bounded data using two non-negative multiplicative bias correction methods*. Computational Statistics & Data Analysis, vol. 92, pages 148–162, 2015. (Cited on page 10.)

[Genovese *et al.* 2012a] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli and Larry Wasserman. *The Geometry of Nonparametric Filament Estimation*. Journal of the American Statistical Association, vol. 107, no. 498, pages 788–799, 2012. (Cited on pages 4 and 49.)

[Genovese *et al.* 2012b] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli and Larry Wasserman. *Minimax manifold estimation.* J. Mach. Learn. Res., vol. 13, pages 1263–1291, 2012. (Cited on pages 4, 31, 35, 48 and 52.)

[Getz & Wilmers 2004] Wayne M. Getz and Christopher C. Wilmers. *A local nearest-neighbor convex-hull construction of home ranges and utilization distributions.* Ecography, vol. 27, pages 489–505, 2004. (Cited on pages 2, 18 and 22.)

[Giné & Guillou 2002] Evarist Giné and Armelle Guillou. *Rates of strong uniform consistency for multivariate kernel density estimators.* Annales de l'Institut Henri Poincare (B) Probability and Statistics, vol. 38, no. 6, pages 907–921, 2002. (Cited on pages 13 and 14.)

[Grassberger & Procaccia 1983] Peter Grassberger and Itamar Procaccia. *Measuring the Strangeness of Strange Attractors.* Physica D: Nonlinear Phenomena, vol. 9, pages 189–208, 1983. (Cited on pages 55 and 56.)

[Härdle *et al.* 1995] Wolfgang K. Härdle, Byeong Uk Park and Alexandre B. Tsybakov. *Estimation of non-sharp support boundaries.* Journal of Multivariate Analysis, vol. 55, pages 205–218, 1995. (Cited on page 15.)

[Hendriks 1990] Harrie Hendriks. *Nonparametric Estimation of a Probability Density on a Riemannian Manifold Using Fourier Expansions.* Annals of Statistics, vol. 18, pages 832–849, 1990. (Cited on page 47.)

[Janson 1987] Svante Janson. *Maximal Spacings in Several Dimensions.* The Annals of Probability, vol. 15, no. 1, pages 274–280, 1987. (Cited on page 11.)

[Jimìenez & Yukich 2011] Raùl Jimìenez and J. E. Yukich. *Nonparametric estimation of surface integrals.* The Annals of Statistics, vol. 39, no. 1, pages 232 – 260, 2011. (Cited on pages 25, 45 and 57.)

[Jones *et al.* 1995] M. C. Jones, O. Linton and J. P. Nielsen. *A Simple Bias Reduction Method for Density Estimation.* Biometrika, vol. 82, no. 2, pages 327–338, 1995. (Cited on page 10.)

[Karunamuni & Zhang 2008] R.J. Karunamuni and S. Zhang. *Some improvements on a boundary corrected kernel density estimator.* Statistics & Probability Letters, vol. 78, no. 5, pages 499–507, 2008. (Cited on page 10.)

[Kim & Park 2013] Yoon Tae Kim and Hyun Suk Park. *Geometric structures arising from kernel density estimation on Riemannian manifolds.* J. Multivar. Anal., vol. 114, pages 112–126, 2013. (Cited on page 47.)

[Kim & Zhou 2015] Arlene K. H. Kim and Harrison H. Zhou. *Tight minimax rates for manifold estimation under Hausdorff loss.* Electron. J. Stat., vol. 9, no. 1, pages 1562–1582, 2015. (Cited on pages 35 and 43.)

[Kohonen 2004] Teuvo Kohonen. *Self-organized formation of topologically correct feature maps.* Biological Cybernetics, vol. 43, pages 59–69, 2004. (Cited on pages 5 and 55.)

[Leblanc 2010] Alexandre Leblanc. *A bias-reduced approach to density estimation using Bernstein polynomials.* Journal of Nonparametric Statistics, vol. 22, pages 459 – 475, 2010. (Cited on page 10.)

[Lee *et al.* 2004] John Aldo Lee, Amaury Lendasse and Michel Verleysen. *Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis.* Neurocomputing, vol. 57, pages 49–76, 2004. (Cited on page 56.)

[Lee 2011] John M. Lee. Introduction to topological manifolds, volume 202 of *Graduate Texts in Mathematics*. Springer, New York, second édition, 2011. (Cited on page 32.)

[Lennon *et al.* 2002] Marc Lennon, Grégoire Mercier, Marie-Catherine Mouchot and Laurence Hubert-Moy. *Curvilinear component analysis for nonlinear dimensionality reduction of hyperspectral images.* In SPIE Remote Sensing, 2002. (Cited on page 56.)

[Maggioni *et al.* 2016] Mauro Maggioni, Stanislav Minsker and Nate Strawn. *Multiscale dictionary learning: non-asymptotic bounds and robustness.* J. Mach. Learn. Res., vol. 17, pages Paper No. 2, 51, 2016. (Cited on page 42.)

[Marron & Ruppert 1994] J. S. Marron and D. Ruppert. *Transformations to Reduce Boundary Bias in Kernel Density Estimation.* Journal of the Royal Statistical Society. Series B (Methodological), vol. 56, no. 4, pages 653–671, 1994. (Cited on page 10.)

[Mattila 1995] Pertti Mattila. Geometry of sets and measures in euclidean spaces: Fractals and rectifiability. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 1995. (Cited on page 45.)

[Nagel] W. Nagel. *Image Analysis and Mathematical Morphology. Volume 2: Theoretical Advances. Edited by Jean Serra.* Journal of Microscopy, vol. 152, pages 597–597. (Cited on page 50.)

[Niyogi *et al.* 2008] Partha Niyogi, Stephen Smale and Shmuel Weinberger. *Finding the Homology of Submanifolds with High Confidence from Random Samples.* Discrete & Computational Geometry, vol. 39, pages 419–441, 2008. (Cited on pages 2, 47 and 56.)

[Oudot 2008] Steve Oudot. *On the Topology of the Restricted Delaunay Triangulation and Witness Complex in Higher Dimensions.* https://arxiv.org/abs/0803.1296, 2008. (Cited on page 31.)

[Pateiro-López & Rodríguez-Casal 2008] Beatriz Pateiro-López and Alberto Rodríguez-Casal. *Length and surface area estimation under smoothness restrictions.* Advances in Applied Probability, vol. 40, no. 2, page 348–358, 2008. (Cited on page 25.)

[Patrangenaru & Ellingson 2015] Vic Patrangenaru and Leif Ellingson. *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis*. 2015. (Cited on page 58.)

[Pelletier 2005] Bruno Pelletier. *Kernel density estimation on Riemannian manifolds*. Statistics & Probability Letters, vol. 73, no. 3, pages 297–304, 2005. (Cited on page 47.)

[Pennec 2006] Xavier Pennec. *Intrinsic Statistics on Riemannian Manifolds: Basic Tools for Geometric Measurements*. Journal of Mathematical Imaging and Vision, vol. 25, pages 127–154, 2006. (Cited on page 58.)

[Penrose 1999] M. Penrose. *A Strong Law for the Largest Nearest-Neighbour Link between Random Points*. Journal of The London Mathematical Society-second Series, vol. 60, pages 951–960, 1999. (Cited on page 36.)

[Pinelis 1994] I. Pinelis. *Extremal Probabilistic Problems and Hotelling's T^2 Test Under a Symmetry Condition*. Annals of Statistics, vol. 22, pages 357–368, 1994. (Cited on page 38.)

[Polonik 1995] Wolfgang Polonik. *Measuring Mass Concentrations and Estimating Density Contour Clusters-An Excess Mass Approach*. The Annals of Statistics, vol. 23, no. 3, pages 855–881, 1995. (Cited on page 22.)

[Reitzner 2003] Matthias Reitzner. *Random Polytopes and the Efron-Stein Jackknife Inequality*. The Annals of Probability, vol. 31, no. 4, pages 2136–2166, 2003. (Cited on page 10.)

[Rényi & Slanke 1963] A. Rényi and R. Sulanke. *Uber die konvexe Hülle von n zufällig gewählten Punkten*. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete , vol. 2, pages 75–84, 1963. (Cited on page 1.)

[Rényi & Slanke 1964] A. Rényi and R. Sulanke. *Uber die konvexe Hülle von n zufällig gewählten Punkten II*. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete , vol. 3, pages 138–147, 1964. (Cited on page 1.)

[Rigollet & Vert 2009] Philippe Rigollet and Régis Vert. *Optimal rates for plug-in estimators of density level sets*. Bernoulli, vol. 15, no. 4, pages 1154–1178, 2009. (Cited on page 22.)

[Rinaldo & Wasserman 2010] Alessandro Rinaldo and Larry Wasserman. *GENERALIZED DENSITY CLUSTERING*. The Annals of Statistics, vol. 38, no. 5, pages 2678–2722, 2010. (Cited on page 59.)

[Rodríguez-Casal, A. & Saavedra-Nieves, P. 2016] Rodríguez-Casal, A. and Saavedra-Nieves, P. *A fully data-driven method for estimating the shape of a point cloud*. ESAIM: PS, vol. 20, pages 332–348, 2016. (Cited on pages 16 and 52.)

[Rodríguez Casal 2007] Alberto Rodríguez Casal. *Set estimation under convexity type assumptions.* Ann. Inst. H. Poincaré Probab. Statist., vol. 43, no. 6, pages 763–774, 2007. (Cited on pages 2, 7, 16, 17, 20, 28 and 38.)

[Rodríguez-Casal & Saavedra-Nieves. 2019a] A. Rodríguez-Casal and P. Saavedra-Nieves. *Extent of occurrence reconstruction using a new data-driven support estimator.* https://arxiv.org/abs/1907.08627, 2019. (Cited on pages 16, 52 and 61.)

[Rodríguez-Casal & Saavedra-Nieves. 2019b] A. Rodríguez-Casal and P. Saavedra-Nieves. *Minimax Hausdorff estimation of density level sets.* https://arxiv.org/abs/1905.02897, 2019. (Cited on pages 22 and 61.)

[Ruppert & Cline 1994] David Ruppert and Daren B. H. Cline. *Bias Reduction in Kernel Density Estimation by Smoothed Empirical Transformations.* The Annals of Statistics, vol. 22, no. 1, pages 185 – 210, 1994. (Cited on page 10.)

[Schneider 1988] Rolf Schneider. *Random approximation of convex sets.* Journal of Microscopy, vol. 151, no. 3, pages 211–227, 1988. (Cited on pages 7 and 10.)

[Schutt 1993] Carsten Schutt. *Random polytopes and affine surface area.* 1993. (Cited on page 10.)

[Singh *et al.* 2007a] Gurjeet Singh, Facundo Mémoli and Gunnar E. Carlsson. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.* In PBG@Eurographics, 2007. (Cited on pages 5 and 55.)

[Singh *et al.* 2007b] Gurjeet Singh, Facundo Mémoli and Gunnar E. Carlsson. *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition.* In PBG@Eurographics, 2007. (Cited on pages 5 and 55.)

[Srivastava *et al.* 2008] Anuj Srivastava, Chafik Samir, Shantanu H. Joshi and Mohamed Daoudi. *Elastic Shape Models for Face Analysis Using Curvilinear Coordinates.* Journal of Mathematical Imaging and Vision, vol. 33, pages 253–265, 2008. (Cited on pages 5 and 56.)

[Tenenbaum *et al.* 2000] Joshua B. Tenenbaum, Vin de Silva and John C. Langford. *A Global Geometric Framework for Nonlinear Dimensionality Reduction.* Science, vol. 290, no. 5500, pages 2319–2323, 2000. (Cited on pages 5 and 56.)

[Thäle & Yukich 2016] Christoph Thäle and J.E. Yukich. *Asymptotic theory for statistics of the Poisson–Voronoi approximation.* Bernoulli, vol. 22, no. 4, pages 2372 – 2400, 2016. (Cited on pages 25 and 57.)

[Tsybakov 1997] A. B. Tsybakov. *On nonparametric estimation of density level sets.* The Annals of Statistics, vol. 25, no. 3, pages 948 – 969, 1997. (Cited on page 22.)

[Vapnik 2000a] Vladimir Naumovich Vapnik. *The Nature of Statistical Learning Theory.* In Statistics for Engineering and Information Science, 2000. (Cited on page 60.)

[Vapnik 2000b] Vladimir Naumovich Vapnik. *The Nature of Statistical Learning Theory.* In Statistics for Engineering and Information Science, 2000. (Cited on page 60.)

[Walther 1999] Guenther Walther. *On a generalization of Blaschke's Rolling Theorem and the smoothing of surfaces.* Mathematical Methods in The Applied Sciences, vol. 22, pages 301–316, 1999. (Cited on pages 8 and 26.)

[Zomorodian & Carlsson 2004] Afra Zomorodian and Gunnar E. Carlsson. *Computing persistent homology.* In SCG '04, 2004. (Cited on page 55.)