Bernoulli **26**(4), 2020, 3017–3050
https://doi.org/10.3150/20-BEJ1214

A k-points-based distance for robust geometric inference

CLAIRE BRÉCHETEAU[1,2] and CLÉMENT LEVRARD[3]

[1]*École Centrale de Nantes, 1 Rue de la Noë, 44300 Nantes, France.*
E-mail: claire.brecheteau@ec-nantes.fr; url: http://www.math.sciences.univ-nantes.fr/~brecheteau/
[2]*Laboratoire de Mathématiques Jean Leray, Faculté des Sciences et des Techniques, 2 Chemin de la Houssinière, 44322 Nantes, France*
[3]*Laboratoire de Probabilités, Statistiques et Modélisation, Bâtiment Sophie Germain, Université Paris-Diderot, 75013 Paris, France.*
E-mail: clement.levrard@lpsm.paris; url: http://www.normalesup.org/~levrard

Analyzing the sub-level sets of the distance to a compact submanifold of $\mathbb{R}^d$ is a common method in topological data analysis, to understand its topology. Therefore, topological inference procedures usually rely on a distance estimate based on n sample points (*Discrete Comput. Geom.* **33** (2005) 249–274). In the case where sample points are corrupted by noise, the distance-to-measure function (DTM, *Found. Comput. Math.* **11** (2011) 733–751) is a surrogate for the distance-to-compact-set function. In practice, approximating the homology of its sub-level sets requires to compute the homology of unions of n balls (*Discrete Comput. Geom.* **49** (2013) 22–45; In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* (2015) 168–180 SIAM), that might become intractable whenever n is large. To simultaneously face the two problems of a large number of points and noise, we introduce the k-power-distance-to-measure function (k-PDTM). This new surrogate for the distance-to-compact is a k-points-based approximation of the DTM. These k points are minimizers of a robustified version of the classical k-means criterion (In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66)* (1967) 281–297 Univ. California Press). The sublevel sets of the k-PDTM consist in unions of k balls, and this distance is also proved robust to noise. We assess the quality of this approximation for k possibly drastically smaller than n, and provide an algorithm to compute this k-PDTM from a sample. Numerical experiments illustrate the good behavior of this k-points approximation in a noisy topological inference framework.

Keywords: minimax rates; quantization; robust distance estimation; topological inference

1. Introduction

Geometric and topological inference consist in recovering geometric and topological features of a compact set (e.g., a compact submanifold in $\mathbb{R}^d$) such as its intrinsic dimension, its curvature or its homology (number of connected components, loops, voids etc.), from a set of n points sampled nearby. In statistics, such information is relevant, both to identify structures in datasets and to post process the data, be it in shape matching, classification or reconstruction etc. Methods of topological data analysis apply to numerous domains such as biology [30] or materials science [21], to name a few.

The information supplied by the homology is varied. The number of connected components is strongly related to the number of clusters in which the sample could be split. Other homolog-

ical features may help determining the true number of parameters required to describe the data. In particular, a widespread objective in topological inference consists in the construction, from a dataset, of approximations of the (possibly fictive) underlying compact set, with the correct homology. In line with current questions in statistics, in the domain of signal compression for instance, with wavelet compression, image segmentation etc., more interest should be payed to approximations that have a low storage cost. This paper provides results in this direction, in a noisy framework, via distance functions.

Let $M \subset \mathbb{R}^d$ be a compact set whose geometry and topology are to be inferred from an n-sample $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ drawn on M. The pioneering work [39] has paved the way of topological inference, showing that the Devroye-Wise estimator $\bigcup_{i=1}^n \overline{B}(X_i, r)$, the union of closed Euclidean balls of radius r centered at sampled points, has the same homology as M, provided that M is a compact submanifold and r is well-chosen according to n. This result can be thought of as a particular instance of topological inference based on distance estimation: if d_M denotes the distance to M, then d_M is inferred via $\mathrm{d}_{\mathbb{X}_n}$, the distance to sample points. The method exposed in [39] then boils down inferring the homology of the 0-sublevel set of d_M, $\mathrm{d}_M^{-1}((-\infty, 0])$ from the homology of the r-sublevel set of $\mathrm{d}_{\mathbb{X}_n}$. A general framework for geometric inference based on distance function estimation can be found in [16]. In a nutshell, [16], Proposition 4.3, states that if $\hat{\mathrm{d}}$ is an estimator for d_M and M is smooth enough, then, for some $r > 0$, the r-sublevel set of $\hat{\mathrm{d}}$ has the same homology as M provided that $\|\mathrm{d}_M - \hat{\mathrm{d}}\|_\infty$ is small enough.

This distance estimation problem has been thoroughly investigated through the lens of Hausdorff set estimation: indeed, if $\hat{M}$ is a set estimator, $\hat{\mathrm{d}} = \mathrm{d}_{\hat{M}}$, and d_H denotes the Hausdorff distance, then $\mathrm{d}_H(\hat{M}, M) = \|\hat{\mathrm{d}} - \mathrm{d}_M\|_\infty$. Optimal rates of convergence for $\|\hat{\mathrm{d}} - \mathrm{d}_M\|_\infty$, in terms of sample size n have been derived under various types of regularity assumptions on M and noise conditions. In the noise-free case, optimal rates for $\|\hat{\mathrm{d}} - \mathrm{d}_M\|_\infty$ are given in [41] whenever M satisfies some convexity-type assumptions, whereas [1,2,28,31] provide optimal rates when M is a smooth compact manifold. Note that, in the smooth manifold case with noisy observations, additional results on optimal rates for Hausdorff estimation can be found in [28]. All of these bounds can be combined with the aforementioned result [16], Proposition 4.3, to assess that the homology of M may be retrieved from the sublevel sets of $\hat{\mathrm{d}}$, provided that n is large enough.

However, when the sample size n is large, computing the homology from an n-points-based distance estimator $\hat{\mathrm{d}}$ may be computationally intractable. For instance, in the simplest case where $\hat{\mathrm{d}} = \mathrm{d}_{\mathbb{X}_n}$, a standard way to compute the homology of a sublevel set of $\mathrm{d}_{\mathbb{X}_n}$ is to build a Rips complex based on $\mathbb{X}_n$ whose homology can be efficiently computed [43]. The construction of such a simplicial complex requires the computation of pairwise distances, that is n^2 distances. To reduce this computational cost, a practical solution is to extract a *coreset* $\mathbb{X}_k \subset \mathbb{X}_n$ such that $\|\mathrm{d}_{\mathbb{X}_k} - \mathrm{d}_M\|_\infty$ is small enough to ensure topological correctness, then to compute a Rips complex based on $\mathbb{X}_k$. In the noise-free case, extracting such a coreset boils down to find an ε-covering of $\mathbb{X}_n$, where ε is the desired sup-norm precision. Using a uniform grid shows that an ε-covering with $k(\varepsilon) \lesssim \varepsilon^{-\frac{1}{d}}$ points at most exists, and can be found in practice using farthest point sampling algorithm for instance [24]. Such a coreset may also be used to compute more involved estimates for d_M, as in [36].

In noisy settings, with observations of the type $X_i = Y_i + N_i$, Y_i on M and N_i denoting Gaussian noise, using a covering of sample points as base points for a coreset can lead to arbitrarily

poor estimation of d_M. The goal of this paper is to nonetheless provide a coreset in such noisy situations, that is to build an approximation of d_M, based on the computation of a distance to k points, that may be proved close enough to d_M to allow further geometric inference.

To be more precise, we will build our k-points distance approximation as an approximation of the *distance-to-measure* [16], that may be thought of as a robust surrogate for d_M. Namely, for a Borel probability measure P on $\mathbb{R}^d$, a mass parameter $h \in [0, 1]$ and $x \in \mathbb{R}^d$, the distance of x to the measure P (DTM), $\mathrm{d}_{P,h}(x)$ is defined by

$$\mathrm{d}^2_{P,h}(x) = P_{x,h}\|x - \cdot\|^2,$$

where $P_{x,h}$ is the probability distribution defined as the restriction of the distribution P to the ball centered at x, with P-mass h, and with the notation Qf for the expectation of the function f with respect to the distribution Q. When P is uniform enough on M and M is regular enough, this distance is proved to approximate well the distance to M ([16], Proposition 4.9) and is robust to noise ([16], Theorem 3.5).

The distance-to-measure is usually inferred from $\mathbb{X}_n$ via its empirical counterpart, called *empirical DTM*, replacing P by the empirical distribution $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, where δ_x is the Dirac mass on x. As noted in [29], the sublevel sets of empirical DTM are unions of around $\binom{n}{q}$ balls, with $q = hn$, which makes their computation intractable in practice. To bypass this issue, approximations of the empirical DTM have been proposed in [29] (q-witnessed distance) and [13] (power distance). Up to our knowledge, these are the only available approximations of the empirical DTM. The sublevel sets of these two approximations are union of n balls. Thus, it makes the computation of topological invariants more tractable for small data sets, from alpha-shapes for instance (see, e.g., [22]). Nonetheless, when n is large, there is still a need for an optimal set of points allowing to efficiently compute an approximation of the DTM, as pointed out in [40]. Up to our knowledge, the only results on such a reduction are on the negative side, exposing a lower bound on the number of points $k(\varepsilon)$ that are needed to build an ε-approximation of the empirical DTM [38].

The main contribution of this paper is the construction (Section 2.3), for a distribution P and a mass parameter h, of a k-power distance $\mathrm{d}_{P,h,k}$ of the form

$$\mathrm{d}_{P,h,k}(x) = \sqrt{\min_{i \in [\![1,k]\!]} \|x - \tau_i\|^2 + \omega^2_{P,h}(\tau_i)},$$

that we call *k-power-distance-to-measure*, k-PDTM for short. We will prove that this k-points power distance is robust to noise (Proposition 17), and is a provably good approximation of the distance-to-measure (Proposition 14). This will allow us to give bounds on $\|\mathrm{d}_{P,h,k} - \mathrm{d}_M\|_\infty$ (Proposition 18) that can be used for further topological inference based on the sublevel sets of $\mathrm{d}_{P,h,k}$. We then prove that its empirical counterpart $\mathrm{d}_{P_n,h,k}$ is an optimal approximation of $\mathrm{d}_{P,h,k}$ from an n-sample (Theorem 19 and Proposition 21). At last we provide a Lloyd's type algorithm [34] to compute in practice such a k-power distance based on n sample points (Section 3.3), and numerically illustrate its good performance in a framework of topological inference (Section 4).

The paper is organized as follows. Section 2 introduces definitions, notations and base results that are required for the construction of the k-PDTM. A proper definition of $\mathrm{d}_{P,h,k}$ is given in Section 2.3, along with some basic properties. Section 3 exposes the main theoretical results of

the paper, that consist in guarantees for the k-PDTM in a topological inference framework (Section 3.1), optimality of the sample approximation of the k-PDTM (Section 3.2), and an algorithm to compute it (Section 3.3). Numerical illustrations are given in Section 4, and Sections 5 and 6 gather the derivations of the main results. Proof of technical intermediate results as well as additional figures are deferred to the Appendix.

2. Notations, definitions and first results

2.1. Notations for the distance-to-measure

Throughout the paper, observations will be elements of the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. The ball centered at c with radius r is denoted by $\mathrm{B}(c,r) = \{x \in \mathbb{R}^d \mid \|x-c\| < r\}$ and its closure by $\overline{\mathrm{B}}(c,r)$. The sphere is denoted by $\mathrm{S}(c,r) = \{x \in \mathbb{R}^d \mid \|x-c\| = r\}$. As well, if $A \subset \mathbb{R}^d$, $\overline{A}$ denotes the closure of A, A° its interior, $\partial A = \overline{A} \setminus A^\circ$ its boundary and $A^c = \mathbb{R}^d \setminus A$ its complementary set in $\mathbb{R}^d$. For any positive integer k, $[\![1,k]\!] = \{1,2,\ldots,k\}$. For any set A, $A^{(k)}$ stands for equivalence classes of $\{\mathbf{t} = (t_1,t_2,\ldots,t_k) \mid \forall i \in [\![1,k]\!], t_i \in A\}$, where two elements are identified whenever they are equal up to a permutation of the coordinates. Following the quantization terminology, elements of $(\mathbb{R}^d)^{(k)}$ are called *codebooks* and their k elements *codepoints*. For any distribution P and any integrable function f, the integral of f with respect to P is denoted by Pf or $\int f(u)P(\mathrm{d}u)$. We also denote $\sup_x |f(x)|$ by $\|f\|_\infty$. For $a, b \in \mathbb{R}$ the maximum and minimum of a and b will be denoted by $a \vee b$ and $a \wedge b$.

We consider probability distributions P with support $\mathrm{Supp}(P) \subset \mathbb{R}^d$. The family of these distributions is denoted by $\mathcal{P}(\mathbb{R}^d)$. The subset of distributions P in $\mathcal{P}(\mathbb{R}^d)$ with finite moment of order 2 ($P\|\cdot\|^2 < \infty$) is denoted by $\mathcal{P}_{(2)}(\mathbb{R}^d)$. The distribution whose support is to be inferred is an element of $\mathcal{P}^K(\mathbb{R}^d) = \{P \in \mathcal{P}(\mathbb{R}^d) \mid \mathrm{Supp}(P) \subset \overline{\mathrm{B}}(0,K)\}$ for $K > 0$. To infer $\mathrm{Supp}(P)$, we use a modified version Q of P. This measure Q is assumed to be *sub-Gaussian* with variance $V^2 > 0$. That is, Q is a distribution in $\mathcal{P}(\mathbb{R}^d)$ such that

$$Q\big(\mathrm{B}(0,t)^c\big) \le \exp\left(-\frac{t^2}{2V^2}\right)$$

for all $t > V$. The set of such measures is denoted by $\mathcal{P}^{(V)}(\mathbb{R}^d)$. Given $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ an n-sample from P, we denote by $P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ the corresponding empirical distribution.

For $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$ and $h \in (0,1]$, we use the notation $\mathcal{P}_h(P)$ for the set of distributions $P_h = \frac{1}{h}\mu$ with μ a submeasure of P (i.e., such that $\mu(B) \le P(B)$ for every Borel set $B \subset \mathbb{R}^d$) satisfying $\mu(\mathbb{R}^d) = h$. The set of all of their expectations is defined by

$$\tilde{\mathcal{M}}_h(P) = \big\{m(P_h) \mid P_h \in \mathcal{P}_h(P)\big\},$$

with the notation $m(P_h) = \int u P_h(\mathrm{d}u)$ for the mean of P_h, $v(P_h) = \int \|u - m(P_h)\|^2 P_h(\mathrm{d}u)$ for its variance and $M(P_h) = \|m(P_h)\|^2 + v(P_h)$ for its order 2 moment.

Some distributions in $\mathcal{P}_h(P)$ will be of special interest. Denote by

$$\delta_{P,h}(x) = \inf\big\{r > 0 \mid P\big(\overline{\mathrm{B}}(x,r)\big) > h\big\}$$

the smallest radius of a ball centered at $x \in \mathbb{R}^d$ of P-mass h. Then, *local distributions* are defined as restrictions of P to these balls.

Definition 1. Let $P \in \mathcal{P}(\mathbb{R}^d)$. The set of local distributions at a point x with mass parameter $h \in (0, 1]$, denoted by $\mathcal{P}_{x,h}(P)$ is the set of distributions $P_{x,h}$ defined by $P_{x,h} = \frac{1}{h}\mu$, where μ satisfies:

1. μ is a submeasure of P with P-mass h: $\frac{1}{h}\mu \in \mathcal{P}_h(P)$.
2. μ coincides with P on $\mathrm{B}(x, \delta_{P,h}(x))$.
3. $\mathrm{Supp}(\mu) \subset \overline{\mathrm{B}}(x, \delta_{P,h}(x))$.

Note that when $P(\partial \mathrm{B}(x, \delta_{P,h}(x))) = 0$, the set of distributions $\mathcal{P}_{x,h}(P)$ is reduced to a singleton $\{P_{x,h}\}$ with $P_{x,h}(B) = \frac{1}{h}P(B \cap \mathrm{B}(x, \delta_{P,h}(x)))$, for any Borel set B. Accordingly, we may define a *local mean* as the expectation of a local distribution, $m(P_{x,h}) = \int u P_{x,h}(\mathrm{d}u)$ associated to some $x \in \mathbb{R}^d$ and $h \in (0, 1]$. The *set of local means* of P with parameter $h \in (0, 1]$ is defined by

$$\mathcal{M}_h(P) = \left\{ m(P_{x,h}) \mid x \in \mathbb{R}^d, P_{x,h} \in \mathcal{P}_{x,h}(P) \right\}.$$

Example 2 (Uniform distribution on a circle). Let $P = \mathcal{U}_{\mathrm{S}(0,1)}$, the uniform distribution on the unit sphere $\mathrm{S}(0, 1) \subset \mathbb{R}^2$. Then, for every $x \neq 0$, $\mathcal{P}_{x,h}(P)$ is the singleton $\{P_{x,h}\}$, $P_{x,h}$ being the uniform distribution on the arc centered at $\frac{x}{\|x\|}$ subtending an angle $2\pi h$. For $x = 0$, $\mathcal{P}_{0,h}(P)$ coincides with $\mathcal{P}_h(P)$. As a consequence, for $x \neq 0$, $m(P_{x,h}) = \mathrm{sinc}(h\pi)\frac{x}{\|x\|}$ and $v(P_{x,h}) = 1 - \mathrm{sinc}(h\pi)^2$ where $\mathrm{sinc} : x \mapsto \frac{\sin(x)}{x}$ is the sinus cardinal function. For $x = 0$, the set $\{m(P_{0,h}) \mid P_{0,h} \in \mathcal{P}_{0,h}(P)\}$ coincides with the ball $\overline{\mathrm{B}}(0, \mathrm{sinc}(h\pi))$ and $v(P_{0,h}) \geq 1 - \mathrm{sinc}(h\pi)^2$ with equality if and only if $P_{0,h} = P_{x,h}$ for some $x \neq 0$. Note that for such an example, $\tilde{\mathcal{M}}_h(P)$ coincides with the set of local means $\mathcal{M}_h(P)$.

These notions of local distributions and local means are required to define the notion of distance-to-measure.

2.2. Definition of the distance-to-measure (DTM)

In the framework of geometric inference, to face the non-robustness to noise of the function distance to a compact set, the notion of *distance-to-measure* (*DTM*) has been introduced in [16]. The DTM $\mathrm{d}_{P,h}$ is a function defined on $\mathbb{R}^d$, associated with a probability distribution P and depending on a mass parameter $h \in [0, 1]$. Two equivalent definitions of the DTM in terms of submeasures are given in [16], Proposition 3.3. For every $x \in \mathbb{R}^d$, $h \in (0, 1]$ and $P_{x,h} \in \mathcal{P}_{x,h}(P)$,

$$\mathrm{d}^2_{P,h}(x) = \inf_{P_h \in \mathcal{P}_h(P)} P_h\|x - \cdot\|^2 = P_{x,h}\|x - \cdot\|^2. \tag{1}$$

Note that $P_{x,h}\|x - \cdot\|^2$ does not depend on the choice of $P_{x,h} \in \mathcal{P}_{x,h}(P)$. Indeed, if P_1, P_2 are in $\mathcal{P}_{x,h}(P)$, then they coincide on $\mathrm{B}(x, \delta_{P,h}(x))$, and the function $\|x - \cdot\|^2$ is constant on

$\partial \mathrm{B}(x, \delta_{P,h}(x))$. Whenever h is small, the DTM provably approximates well the distance to the support of P [16], Corollary 4.8 and Proposition 4.9, when P is a uniform distribution on a submanifold for instance.

If a small h allows to approximate the distance to the support of P, larger values for h make the DTM robust to small variations of the distribution P, in terms of the Wasserstein metric. Indeed, according to [16], if P and Q are two probability distributions on the space $(\mathbb{R}^d, \|\cdot\|)$ with finite second moment, then

$$\|\mathrm{d}_{P,h} - \mathrm{d}_{Q,h}\|_\infty \leq \frac{1}{\sqrt{h}} W_2(P, Q). \tag{2}$$

Let us recall that the Wasserstein metric W_p is defined, for $p \geq 1$, by

$$W_p^p(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}\big[\|X - Y\|^p\big], \tag{3}$$

where $\Pi(P, Q)$ denotes the set of distributions on $\mathbb{R}^d \times \mathbb{R}^d$ of random vectors (X, Y) such that $X \sim P$ (i.e., X is a random variable with distribution P) and $Y \sim Q$.

According to (1) and the bias-variance decomposition $P_h\|x - \cdot\|^2 = \|x - m(P_h)\|^2 + v(P_h)$, for $P_h \in \mathcal{P}_h(P)$, the distance-to-measure can be expressed as a *power distance* in the following Equation (4), that is as the square root of a function $f_{\tau,\omega} : x \mapsto \inf_{i\in I} \|x - \tau_i\|^2 + {\omega_i}^2$ for some set I, a family of centers $\boldsymbol{\tau} = (\tau_i)_{i\in I}$ and weights $\boldsymbol{\omega} = (\omega_i)_{i\in I}$. Namely, for every $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$ and $x \in \mathbb{R}^d$, we have

$$\mathrm{d}^2_{P,h}(x) = \inf_{P_h \in \mathcal{P}_h(P)} \big\|x - m(P_h)\big\|^2 + v(P_h), \tag{4}$$

where the minimum is attained at any measure $P_h = P_{x,h}$ in $\mathcal{P}_{x,h}(P)$. In this case, the centers $m(P_h)$ are elements of $\tilde{\mathcal{M}}_h(P)$. Note that according to the second equality of Equation (1), $\mathcal{P}_h(P)$ can be replaced by $\bigcup_{x\in\mathbb{R}^d} \mathcal{P}_{x,h}(P)$ in (4), so that the centers are actually elements of $\mathcal{M}_h(P)$.

Special instances of power distances that will be of particular interest in the following are *k-power distances*, indexed on a finite set of cardinality $|I| = k$. The following example gives some intuition on why the distance-to-measure can be a convenient tool for geometric inference in noisy settings, compared to classical quantization-based approaches.

Example 3 (Uniform distribution on a circle with noise). Let $Q_\beta = \beta\mathcal{U}_{\mathrm{S}(0,1)} + (1-\beta)\mathcal{U}_{\mathrm{B}(0,1)}$ be a noisy version of $P = \mathcal{U}_{\mathrm{S}(0,1)}$, the uniform distribution on the circle, for some $\beta \in (0, 1)$. According to [16], Theorem 3.5, Corollary 4.8, since $W_2(Q_\beta, P) \leq \sqrt{1-\beta}$, we have $\|\mathrm{d}_{Q_\beta,h} - \mathrm{d}_{\mathrm{S}(0,1)}\|_\infty \leq Ch + \sqrt{\frac{1-\beta}{h}}$ for some $C > 0$. Thus, for $h > 81(1-\beta)$ and $1-\beta$ small enough, [16], Theorem 4.6, ensures that the r-sublevel sets of $\mathrm{d}_{Q_\beta,h}$ are homotopy equivalent to $\mathrm{S}(0, 1)$, for a range of r's.

Now let $\boldsymbol{\tau}^*$ be a minimizer of the k-means criterion $Q_\beta \min_{j\in[\![1,k]\!]} \|\cdot - \tau_j\|^2$, in other words, an optimal k-points codebook for Q_β. The following lemma shed some light on the approximation properties of the distance to $\boldsymbol{\tau}^*$ function.

Lemma 4. *Consider the setup in Example 3. Let* $\mathrm{d}_{\boldsymbol{\tau}^*}$ *denote the distance to* $\boldsymbol{\tau}^*$ *function. Then, for k large enough,*

$$\sup_{x\in \mathrm{S}(0,1)} \left|\mathrm{d}_{\boldsymbol{\tau}^*}(x) - \mathrm{d}_{\mathrm{S}(0,1)}(x)\right| \leq C\left(\frac{1}{k^2} + (1-\beta)\right)^{\frac{1}{3}},$$

for some constant $C > 0$. *On the other hand, for every* $\rho > 0$, *there exists* $k_{\rho,\beta}$ *such that, for all* $k \geq k_{\rho,\beta}$, $\boldsymbol{\tau}^*$ *has at least one codepoint in* $\mathrm{B}(0,\rho)$.

A direct consequence of Lemma 4 is

$$\sup_{k\geq 0} \|\mathrm{d}_{\mathrm{S}(0,1)} - \mathrm{d}_{\boldsymbol{\tau}^*}\|_\infty \geq \sup_{k\geq 0} \left|\mathrm{d}_{\mathrm{S}(0,1)}(0) - \mathrm{d}_{\boldsymbol{\tau}^*}(0)\right| = 1.$$

A proof of Lemma 4 is given in Section A.1 of the Appendix. The intuition behind Lemma 4 is that though optimal codebooks designed via classical quantization can yield provably good covering of topological structures such as manifolds, they are also likely to have codepoints far from the structure in some noisy cases. In this case, geometric inference based on the sublevel sets of $\mathrm{d}_{\boldsymbol{\tau}^*}$ might lead to poor results.

2.3. Definition of the *k*-PDTM

As illustrated above, in Example 3, the distance-to-measure may be thought of as a robustified version of the distance-to-compact-set, designed for geometric inference in noisy settings. According to (4), its sub-level sets are unions of balls centered at elements of $\tilde{\mathcal{M}}_h(P)$. As noted in [29], in general, for empirical distributions based on n points $\{X_1, \ldots, X_n\}$, this amount of balls is finite but may be large (of order $\binom{n}{nh}$, where h is the mass parameter of the DTM). Approximations of the DTM consisting in reducing this number of balls to the sample size n are exposed in [13,29]. In this paper, we propose to reduce this number of balls to some $k \in \mathbb{N}^*$ possibly much smaller than the sample size, resulting in an approximation of the distance-to-measure that we prove accurate enough for further topological inference. This section is devoted to the introduction of such an approximation, namely the k-PDTM.

The k-PDTM is an approximation of the DTM obtained after reducing the set of submeasures $\mathcal{P}_h(P)$ (or equivalently, the set of centers $\tilde{\mathcal{M}}_h(P)$) to a set of k well-chosen submeasures (or k centers) in the definition of the DTM (4). As an answer to [40], such a set of k centers may be considered as a coreset for the DTM. These k submeasures are obtained by minimizing the following criterion R.

Definition 5. For $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$ and $\mathbf{P} = (P_i)_{i\in[\![1,k]\!]} \in \mathcal{P}_h(P)^{(k)}$, we define $R(\mathbf{P})$ by

$$R(\mathbf{P}) = P \min_{i\in[\![1,k]\!]} \left\|\cdot - m(P_i)\right\|^2 + v(P_i).$$

The following Proposition 6 ensures that there exist optimal submeasures with respect to the risk R.

Proposition 6. *If $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$, then the minimum of R is attained in $\mathcal{P}_h(P)^{(k)}$. We denote by $\mathbf{P}^*$ any such minimizer.*

The proof of Proposition 6 is to be found in Section 5.1. These optimal submeasures allow us to define the k-PDTM as follows.

Definition 7. The k-PDTM is any function $\mathrm{d}_{P,h,k} : \mathbb{R}^d \mapsto \mathbb{R}$ defined by

$$\mathrm{d}^2_{P,h,k}(x) = \min_{i \in [\![1,k]\!]} \left\| x - m\left(P_i^*\right) \right\|^2 + v\left(P_i^*\right),$$

for some $\mathbf{P}^* = (P_1^*, \ldots, P_k^*) \in \arg\min_{\mathbf{P} \in \mathcal{P}_h(P)^{(k)}} R(\mathbf{P})$.

The k-PDTM is a k-power distance whose graph lies above the graph of the DTM. It is not necessarily uniquely defined, since several minimizers of R may exist. Its sublevel sets are unions of k balls. Besides, the k centers of the k-PDTM yield a decomposition of the space $\mathbb{R}^d$ into k cells, and consequently, a decomposition of P into k weighted Voronoi measures.

Definition 8. A set of *weighted Voronoi measures* associated to a distribution $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$, k submeasures $(P_i)_{i \in [\![1,k]\!]} \in \mathcal{P}_h(P)^{(k)}$ and $h \in (0, 1]$ is a collection $\{\tilde{P}_{1,h}, \tilde{P}_{2,h}, \ldots \tilde{P}_{k,h}\}$ of $k \in \mathbb{N}^*$ non-negative submeasures of P such that $\sum_{i=1}^k \tilde{P}_{i,h} = P$ and

$$\forall x \in \mathrm{Supp}(\tilde{P}_{i,h}), \quad \left\| x - m(P_i) \right\|^2 + v(P_i) \leq \left\| x - m(P_j) \right\|^2 + v(P_j), \quad \forall j \in [\![1, k]\!].$$

Note that a set of weighted Voronoi measures can always be assigned to any $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$ and $(P_i)_{i \in [\![1,k]\!]} \in \mathcal{P}_h(P)^{(k)}$. Indeed, $\mathbb{R}^d$ may be split into weighted Voronoi cells associated to the centers $(m(P_i))_{i \in [\![1,k]\!]}$ and weights $(v(P_i))_{i \in [\![1,k]\!]}$ ([9], Section 4.4.2), with ties arbitrarily broken. The following key property of weighted Voronoi measures implies that minimizers $\mathbf{P}^*$ of the criterion R are actually elements of $(\bigcup_{t \in \mathbb{R}^d} \mathcal{P}_{t,h}(P))^{(k)}$.

Proposition 9. *Let $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$, and $(P_i)_{i \in [\![1,k]\!]} \in \mathcal{P}_h(P)^{(k)}$. Let $Q_1, \ldots, Q_k$ be such that $Q_i \in \mathcal{P}_{m(\tilde{P}_{i,h}),h}(P)$, for $i = 1, \ldots, k$. Then*

$$R(Q_1, \ldots, Q_k) \leq R(P_1, \ldots, P_k),$$

with equality only if, for all $i \in [\![1, k]\!]$ such that $\tilde{P}_{i,h}(\mathbb{R}^d) \neq 0$, we have $P_i \in \mathcal{P}_{m(\tilde{P}_{i,h}),h}(P)$.

The proof of Proposition 9 is deferred to Section 5.2. Now assume that, for any $t \in \mathbb{R}^d$, a choice of $P_{t,h} \in \mathcal{P}_{t,h}$ is given by $f(t)$. For $\mathbf{t} \in (\mathbb{R}^d)^k$, set $f(\mathbf{t}) = (f(t_1), \ldots, f(t_k))$. We may then define a risk $R_f(\mathbf{t})$ via the quantity

$$R_f(\mathbf{t}) = R\big(f(\mathbf{t})\big). \tag{5}$$

Proposition 9 shows that minimizing R over the set of k submeasures boils down to minimize $\mathbf{t} \mapsto R_f(\mathbf{t})$ over $(\mathbb{R}^d)^{(k)}$. It also provides a natural and tractable procedure for computing local

optima of the criterion R_f, cf. Algorithm 1 in Section 3.3. An alternative definition of the k-PDTM in terms of local distributions may be stated accordingly.

Corollary 10. *Let* $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$. *The k-PDTM is any function* $\mathrm{d}_{P,h,k} : \mathbb{R}^d \mapsto \mathbb{R}$ *defined by*

$$\mathrm{d}^2_{P,h,k}(x) = \min_{i \in [\![1,k]\!]} \left\| x - m\left(f\left(t_i^*\right)\right)\right\|^2 + v\left(f\left(t_i^*\right)\right),$$

for some $\mathbf{t}^* = (t_1^*, \ldots, t_k^*) \in \arg\min_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} R_f(\mathbf{t})$.

A proof of Corollary 10 is given in Section 5.3. An interesting consequence of Corollary 10 is that whatever the choice of f (choice of $P_{t,h} \in \mathcal{P}_{t,h}$), minimizers $\mathbf{t}^*$ of R_f always provide minimizers of R via $f(\mathbf{t}^*)$. Thus, in what follows, we assume that such a f is given, denote by $P_{t,h}$ the corresponding choice $f(t)$, and, with a slight abuse of notation, denote by $R(\mathbf{t})$ the corresponding risk $R_f(\mathbf{t})$.

The above definition of the k-PDTM in terms of local means and variances is very convenient for the purpose of its computation, but the more general definition of the k-PDTM in terms of submeasures (Definition 7) is also crucial. Indeed, from Definition 7 we may also state an alternative parametrization of the k-PDTM by elements $\tau = m(P_h)$ of $\tilde{\mathcal{M}}_h(P)$ that will allow for a geometric interpretation of the k-PDTM. The corresponding variances $v(P_h)$ will be obtained as images $\omega_{P,h}(\tau)$ of the function $\omega_{P,h}$ defined for every τ in $\mathbb{R}^d$ by

$$\omega^2_{P,h}(\tau) = \sup_{x \in \mathbb{R}^d} \mathrm{d}^2_{P,h}(x) - \|x - \tau\|^2.$$

Lemma 11. *If* $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$, *then* $\omega_{P,h}(\tau) < +\infty$ *if and only if* $\tau \in \tilde{\mathcal{M}}_h(P)$. *Moreover, if* $\tau \in \tilde{\mathcal{M}}_h(P)$, *then there exists* $P_h \in \mathcal{P}_h(P)$ *such that* $\tau = m(P_h)$ *and* $\omega^2_{P,h}(\tau) = v(P_h)$. *More precisely,* $\omega^2_{P,h}(\tau) = \min_{P_h \in \mathcal{P}_h(P), m(P_h) = \tau} v(P_h)$.

The proof of Lemma 11 is deferred to Section 5.4. The natural reparametrization of the criterion R with the set of centers follows.

Theorem 12. *Let* $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$. *Then* $\mathbf{t}^* \in \arg\min_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} R(\mathbf{t})$ *if and only if*

$$\left(m(P_{t_1^*,h}), \ldots, m(P_{t_k^*,h})\right) \in \arg\min_{\boldsymbol{\tau} \in (\mathbb{R}^d)^{(k)}} P \min_{i \in [\![1,k]\!]} \|\cdot - \tau_i\|^2 + \omega^2_{P,h}(\tau_i).$$

The proof of Theorem 12 is to be found in Section 5.5. Theorem 12 states that the k-PDTM is a solution of a weighted k-means-type criterion. According to Lemma 11, the regularization terms $\omega_{P,h}(\tau_i)$ force the optimal codebooks $\boldsymbol{\tau}^*$ to be in $\tilde{\mathcal{M}}_h(P)^{(k)}$. Intuitively, elements τ such that $\omega_{P,h}(\tau)$ is small will be favoured. Such τ's gather a proportion h of the mass of P on their neighborhood. On the contrary, for elements τ such that $\omega_{P,h}(\tau)$ is large, the corresponding weighted Voronoi measures will not be massive, and the ball associated to such τ's will appear in the r-sublevel set of the function $x \mapsto \|x - \tau\|^2 + \omega^2_{P,h}(\tau)$ for large r's only. A direct consequence of Theorem 12 is that the squared k-PDTM may be interpreted as the closest squared

k-power distance to the squared DTM from above, in terms of $L_1(P)$ norm. This interpretation comes from the straightforward inequality $\|x-\tau\|^2+\omega_{P,h}^2(\tau)\geq \mathrm{d}_{P,h}^2(x)$. The resulting inequality $\mathrm{d}_{P,h,k}^2\geq \mathrm{d}_{P,h}^2$ allows for further comparison with k-means approximation of the distance-to-compact-set in noisy settings.

Example 13 (Noisy distribution on the circle). For the distribution $Q_\beta=\beta\mathcal{U}_{\mathrm{S}(0,1)}+(1-\beta)\mathcal{U}_{\mathrm{B}(0,1)}$. If $h>1-\beta$, since $\mathrm{d}_{Q_\beta,h,k}^2(0)\geq \mathrm{d}_{Q_\beta,h}^2(0)$, we have $\mathrm{d}_{Q_\beta,h,k}^2(0)\geq 1-\frac{1-\beta}{h}$. As a consequence, $\inf_{k\geq 0}\mathrm{d}_{Q_\beta,h,k}^2(0)\geq 1-\frac{1-\beta}{h}$, whereas $\inf_{k\geq 0}\mathrm{d}_{\boldsymbol{\tau}^*}^2(0)=0$, where $\boldsymbol{\tau}^*$ denotes an optimal k-points codebook.

The above example shows that we can expect the k-PDTM to approximate well the distance-to-compact-set in remote areas, contrary to the distance based on k-means, $\mathrm{d}_{\boldsymbol{\tau}^*}$. To check whether the k-PDTM provides also an efficient covering of the targeted structure is investigated in the following section.

3. Theoretical results for the k-PDTM

3.1. Geometric inference with the k-PDTM

Let M be a compact subset of $\mathbb{R}^d$, and P a distribution with support M. Here we show that the k-PDTM approximates the DTM, provided that the covering number of M with respect to the Euclidean norm and the continuity modulus $\zeta_{P,h}$ of the map $x\mapsto m(P_{x,h})$ are not too large.

For a subset $\mathcal{F}$ of a Banach space $\mathcal{B}$ endowed with the norm $\|\cdot\|_{\mathcal{B}}$, the ε-covering number of $\mathcal{F}$, $N'_{\|\cdot\|_{\mathcal{B}}}(\varepsilon,\mathcal{F})$ is defined as the minimum number of balls with radius ε that are needed to cover $\mathcal{F}$. In what follows, we adopt the shortcut $f_M(\varepsilon)$ to denote $N'_{\|\cdot\|}(\varepsilon,M)$, that is the ε-covering number of M considered as a subset of $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$. For every $\epsilon>0$, the continuity modulus $\zeta_{P,h}(\epsilon)$ is defined by

$$\zeta_{P,h}(\epsilon)=\sup_{x,y\in M,\|x-y\|\leq\varepsilon}\ \sup_{P_{x,h}\in\mathcal{P}_{x,h}(P),P_{y,h}\in\mathcal{P}_{y,h}(P)}\big\{\big|m(P_{x,h})-m(P_{y,h})\big|\big\}.$$

In what follows, $C_{l_1,\ldots,l_s}$ and $c_{l_1,\ldots,l_s}$ denote quantities depending on $l_1,\ldots,l_s$ only.

Proposition 14. *Let $K>0$, $P\in\mathcal{P}^K(\mathbb{R}^d)$ (a distribution whose support is included in $\overline{\mathrm{B}}(0,K)$), and let $M\subset\mathrm{B}(0,K)$ be such that $P(M)=1$. Let $f_M(\varepsilon)$ denote the ε-covering number of M. Then we have*

$$0\leq P\big(\mathrm{d}_{P,h,k}^2-\mathrm{d}_{P,h}^2\big)\leq 2f_M^{-1}(k)\zeta_{P,h}\big(f_M^{-1}(k)\big),\quad \text{with } f_M^{-1}(k)=\inf\big\{\varepsilon>0\mid f_M(\varepsilon)\leq k\big\}.$$

A proof of Proposition 14 is given in Section 6.2. Whenever P is roughly uniform on its support, the quantities $f_M^{-1}(k)$ and $\zeta_{P,h}$ mostly depend on the dimension and radius of M. We illustrate this point with two instances of particular interest for geometric inference. First, the case where the distribution P has an ambient-dimensional support is investigated.

Corollary 15. *Assume that P has a density f satisfying $0 < f_{\min} \leq f \leq f_{\max}$ on its support. Then*

$$0 \leq P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) \leq C_{f_{\max},K,d,h} k^{-2/d}.$$

The proof of Corollary 15 is given in Section 6.3. Note that no assumptions on the geometric regularity of M are required for Corollary 15 to hold. In the case where M has a lower-dimensional structure, more regularity is required, as for instance, in the following corollary.

Corollary 16. *Suppose that P is supported on $N \subset \mathrm{B}(0,K)$, a compact d'-dimensional $\mathcal{C}^2$-submanifold. Assume that P has a density $0 < f_{\min} \leq f \leq f_{\max}$ with respect to the volume measure on N. Moreover, suppose that P satisfies, for all $x \in N$ and positive r,*

$$P\big(\mathrm{B}(x,r)\big) \geq c f_{\min} r^{d'} \wedge 1. \tag{6}$$

Then, for $k \geq c_{N,f_{\min}}$ and $h \leq C_{N,f_{\min}}$, we have

$$0 \leq P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) \leq C_{N,f_{\min},f_{\max}} k^{-2/d'}.$$

Note that (6), also known as $(cf_{\min}, d')$-standard assumption, is a usual assumption in the set estimation framework (see, e.g., [17]). In the submanifold case, it may be thought of as a condition preventing the boundary from being arbitrarily narrow. This assumption is satisfied for instance in the case where ∂N is empty or is a $\mathcal{C}^2$ $(d'-1)$-dimensional submanifold (see, e.g., [3], Corollary 1). An important feature of Corollary 16 is that this approximation bound does not depend on the ambient dimension. The proof of Corollary 16 may be found in Section 6.4. Next we assess that our k-PDTM shares with the DTM the key property of robustness to noise.

Proposition 17. *Let $P \in \mathcal{P}^K(\mathbb{R}^d)$ for some $K > 0$ and $Q \in \mathcal{P}_{(2)}(\mathbb{R}^d)$. Let $\mathrm{d}^2_{Q,h,k}$ denote a k-PDTM for Q. Then,*

$$P\big(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h,k}\big) \leq 4W_1(P,Q) \sup_{s\in\mathbb{R}^d} \big\|m(P_{s,h})\big\| + \big\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\big\|_{\infty,\mathrm{B}(0,K)}.$$

Further, we have $P|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}| \leq B_{P,Q,h,k}$, where

$$B_{P,Q,h,k} = 3\big\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\big\|_{\infty,\mathrm{B}(0,K)} + P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) + 4W_1(P,Q) \sup_{s\in\mathbb{R}^d} \big\|m(P_{s,h})\big\|.$$

The proof of Proposition 17 can be found in Section 6.5. Note that Lemma 23 provides a bound on $\|m(P_{s,h})\|$ whenever P is sub-Gaussian. Moreover, [16], Theorem 3.5, ensures that $\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\|_{\infty,\mathrm{B}(0,K)}$ can be bounded in terms of $W_2(P,Q)$, up to a constant dependent on K. Combining these results can assess the stability of the k-PDTM, via a bound on $P(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h,k})$. Note at last that bounds on $P(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h})$ may be derived using Proposition 14, leading to the global bound $B_{P,Q,h,k}$.

It is worth mentioning that bounds on $P|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}|$ involving $Q(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{Q,h})$ may be stated as well. However, if the support of Q is not compact, then Proposition 14 cannot be used.

Also, if the support of Q is compact but has a dimension larger than the support of P (in the case of bounded additive noise for instance), Corollary 16 illustrates that $Q(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{Q,h})$ is likely to decrease slower than $P(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h})$ with respect to k. Therefore, whenever Q is thought of as a perturbation of P, bounds on $P|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}|$ in terms of $P(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h})$ lead to better dependencies in k.

Proposition 17 can provide guarantees on $P\mathrm{d}^2_{Q,h,k}$. In turn, provided that M is regular enough, these bounds can be turned into L_∞ bounds between $\mathrm{d}_{Q,h,k}$ and d_M. Following [16], Section 4, these L_∞ bounds can guarantee that the sublevels sets of the k-PDTM are homotopy equivalent to M, under suitable assumptions.

Proposition 18. *Let M be a compact set in $\mathrm{B}(0, K)$ such that $P(M) = 1$. Moreover, assume that there exists d' such that, for every $p \in M$ and $r \geq 0$,*

$$P\big(\mathrm{B}(p, r)\big) \geq C(P) r^{d'} \wedge 1. \tag{7}$$

Let Q be a Borel probability measure (thought of as a perturbation of P), and let Δ_P^2 denote $P\mathrm{d}^2_{Q,h,k}$. Then, we have

$$\|\mathrm{d}_{Q,h,k} - \mathrm{d}_M\|_\infty \leq \max\Big\{C(P)^{-\frac{1}{d'+2}} \Delta_P^{\frac{2}{d'+2}}, 2\Delta_P, W_2(P, Q)h^{-\frac{1}{2}}\Big\},$$

where W_2 denotes the Wasserstein distance.

The proof of Proposition 18 can be found in Section 6.6. According to [16], Corollary 4.8, if P satisfies (7), then

$$\|\mathrm{d}_{Q,h} - \mathrm{d}_M\|_\infty \leq \left(\frac{h}{C(P)}\right)^{\frac{1}{d'}} + W_2(P, Q)h^{-\frac{1}{2}}.$$

Hence, Proposition 18 ensures that the k-PDTM achieves roughly the same performance as the distance-to-measure provided that $\mathrm{d}^2_{Q,h,k}$ is small enough on the support M to be inferred. As will be shown in the following section, this will be the case if Q is an empirical measure drawn close to the targeted support.

3.2. Approximation of the k-PDTM from point clouds

In this section, $P \in \mathcal{P}^K(\mathbb{R}^d)$ is a distribution supported on a compact set M to be inferred. We have at our disposal an n-sample $\mathbb{X}_n = \{X_1, X_2, \ldots, X_n\}$ from a modification $Q \in \mathcal{P}_{(2)}(\mathbb{R}^d)$ of P. An approximation of the k-PDTM $\mathrm{d}_{Q,h,k}$, is given by the *empirical k-PDTM* $\mathrm{d}_{Q_n,h,k}$, where $Q_n = \sum_{i=1}^n \frac{1}{n}\delta_{X_i}$ is the empirical measure from $\mathbb{X}_n$.

An approximation of the distance to empirical measure, $\mathrm{d}_{Q_n,h}$, is given by the so-called *q-witnessed distance*, introduced in [29]. Namely, if $Q_{n,i} \in \mathcal{P}_{X_i,\frac{q}{n}}(Q_n)$, $i = 1, \ldots, n$, are n distributions that are uniform on sets of q-nearest neighbors of the points in $\mathbb{X}_n$, the (squared)

q-witnessed distance is defined as follows:

$$\left(\mathrm{d}^W_{Q_n,q}\right)^2 : x \mapsto \min_{i\in[\![1,n]\!]} Q_{n,i}\,\|x-\cdot\|^2.$$

Note that when $k=n$, any q-witnessed distance $\mathrm{d}^W_{Q_n,q}$, for $q=nh$, is an empirical k-PDTM $\mathrm{d}_{Q_n,h,n}$. Indeed, the criterion of Definition 5 is minimal for any such family of n distributions $(Q_{n,i})_{i\in[\![1,n]\!]}$. For any point x whose set of q nearest neighbors in $\mathbb{X}_n$ is not a set of q-nearest neighbors of elements in $\mathbb{X}_n$, the q-witnessed distance may differ from the distance to the empirical measure. Our empirical k-PDTM may be thought of as a slight generalization of the q-witnessed distance. This is a k-power distance, for an arbitrary k.

We investigate the quality of approximation of the DTM $\mathrm{d}_{P,h}$ with the empirical k-PDTM $\mathrm{d}_{Q_n,h,k}$, when Q is defined as the convolution of P with a sub-Gaussian distribution with variance σ^2. Within this context, according to Lemma 24, Q is sub-Gaussian with variance $V^2=(K+\sigma)^2$.

Theorem 19. *Let P be supported on $M\subset \mathrm{B}(0,K)$. Assume that we observe $X_1,\dots,X_n$ such that $X_i=Y_i+Z_i$, where the Y_i's and Z_i's are all independent, Y_i is sampled from P and Z_i is sub-Gaussian with variance σ^2, with $\sigma\le K$. Let Q_n denote the empirical distribution associated with the X_i's. Then, for any $p>0$, with probability larger than $1-10n^{-p}$, we have*

$$\left|P\left(\mathrm{d}^2_{Q_n,h,k}-\mathrm{d}^2_{Q,h,k}\right)\right| \le C\sqrt{k\log(k)d}\,\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}.$$

A proof of Theorem 19 is given in Section 6.7. The $\sqrt{kd}/\sqrt{n}$ term is in line with the rate of convergence for the k-means method (see, e.g., [8]), as well as with the rate of convergence for $\|\mathrm{d}_{P_n,h}-\mathrm{d}_{P,h}\|_\infty$ exposed in [18]. The $K\sigma$ term is due to the expectation with respect to P (instead of Q). Theorem 19, combined with Proposition 17, allows us to choose k in order to minimize $|P(\mathrm{d}^2_{Q_n,h,k}-\mathrm{d}^2_{P,h})|$. Indeed, in the framework of Corollaries 15 and 16 where the support has intrinsic dimension d', such a minimization boils down to optimizing a quantity of the form $\frac{C\sqrt{k\log(k)d}K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C_{P,h}k^{-\frac{2}{d'}}$. Choosing $k\sim n^{\frac{d'}{d'+4}}$ achieves the desired tradeoff between bias and variance. From the point of view of geometric inference, this leads to computing the distance to $n^{d'/(d'+4)}$ points rather than n, which might save some time. Note that when d' is large, smaller choices of k, though suboptimal for our bounds, would nonetheless give the right topology for large n. In some sense, Theorem 19 advocates only an upper bound on k, above which no increase of precision can be expected. Combining Theorem 19 and Proposition 14 leads to the following result.

Proposition 20. *With the same setting as Theorem 19, if M is a submanifold with intrinsic dimension $d'\ge 1$, then*:

$$\left|P\left(\mathrm{d}^2_{Q_n,h,k}-\mathrm{d}^2_{P,h}\right)\right| \le C\sqrt{k\log(k)d}\,\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}} + C_{P,h}k^{-\frac{2}{d'}}.$$

Thus, choosing $k \sim n^{\frac{d'}{d'+4}}$ *leads to*

$$\left|P\left(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{P,h}\right)\right| \leq C_{P,h}\sqrt{d}n^{-\frac{2}{d'+4}}\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h} + C\frac{K\sigma}{\sqrt{h}}.$$

The proof of Proposition 20 is to be found in Section 6.8. Noting that $P\mathrm{d}^2_{P,h} \leq C_P h^{\frac{1}{d'}}$ in this case (see, e.g., [16], Proposition 4.9), Proposition 20 can be combined with Proposition 18 to yield a bound on $\|\mathrm{d}_{Q_n,h,k} - \mathrm{d}_M\|_\infty$.

To assess optimality of Theorem 19 in terms of sample size dependency, a lower bound on the best k-points approximation of the DTM that is achievable on the set of distributions supported in $\mathrm{B}(0, K)$ may be derived from [7], Theorem 1, or [33], Proposition 3.1.

Proposition 21. *For* $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$ *and* P *a probability measure, denote*

$$\mathrm{d}^2_{P,h,\mathbf{t}} : x \mapsto \min_{j\in[\![1,k]\!]}\left[\left\|x - m(P_{t_j,h})\right\|^2 + v(P_{t_j,h})\right],$$

where, for $j = 1, \ldots, k$, $P_{t_j,h} \in \mathcal{P}_{t_j,h}$. *For* $k \geq 3$, $n \geq \frac{3k}{2}$ *and* $h \leq \frac{1}{2k}$, *we have*

$$\inf_{\hat{\mathbf{t}}} \sup_{P|\,\mathrm{Supp}(P)\subset \mathrm{B}(0,K)} \mathbb{E}P\left(\mathrm{d}^2_{P,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\right) \geq c_0\frac{K^2k^{\frac{1}{2}-\frac{2}{d}}}{\sqrt{n}}, \tag{8}$$

where c_0 *is a constant and* $\hat{\mathbf{t}}$ *denotes an empirically designed vector* $(\hat{t}_1, \ldots, \hat{t}_k)$ *in* $(\mathbb{R}^d)^{(k)}$. *Moreover, if* $n \geq 14k$, *then*

$$\inf_{\hat{\mathbf{t}}} \sup_{P|\,\mathrm{Supp}(P)\subset \mathrm{B}(0,K)} \mathbb{E}P\left(\mathrm{d}^2_{P_n,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\right) \geq c_0\frac{K^2k^{\frac{1}{2}-\frac{2}{d}}}{\sqrt{n}} - 32K^2ke^{-\frac{n}{72k^2}}. \tag{9}$$

Thus, Proposition 21 confirms that the sample size dependency of Theorem 19 is optimal in the noise-free case, up to $\log(n)$ factors. A proof is given in Section 6.9.

3.3. Algorithm

In this section, we expose a Lloyd-type algorithm to compute a local minimizer for the cost function associated with the empirical k-PDTM. For an n-sample $\mathbb{X}_n$ with empirical distribution Q_n, Proposition 9 suggests a procedure to minimize the empirical risk $\mathbf{t} \mapsto R_n(\mathbf{t}) = Q_n \min_{i\in[\![1,k]\!]} \|\cdot - m(Q_{nt_i,h})\|^2 + v(Q_{nt_i,h})$. Indeed, given some codebook $\mathbf{t}$, replacing $\mathbf{t}$ with the means of the weighted Voronoi measures $(\tilde{Q}_{nt_i,h})_{i\in[\![1,k]\!]}$ can only decrease the empirical risk R_n. For a sample $\mathbb{X}_n$, this boils down to compute the weighted Voronoi cells $(\mathcal{C}(t_i))_{i\in[\![1,k]\!]}$ (i.e. the support of the measures $(\tilde{Q}_{nt_i,h})_{i\in[\![1,k]\!]}$), and to replace t_i with the mean of the points of $\mathbb{X}_n$ in $\mathcal{C}(t_i)$. We use the notation $|\mathcal{C}(t)|$ for the cardinality number of $\mathcal{C}(t)$, $m(t)$ for $m(Q_{nt,h})$ and $v(t)$ for $v(Q_{nt,h})$. The procedure is described in Algorithm 1.

Algorithm 1: Local minimum algorithm

Input : $\mathbb{X}_n$ an n-sample from Q, h and k ;
Initialization
Sample $t_1, t_2, \ldots, t_k$ from $\mathbb{X}_n$ without replacement ;
while $R_n(\mathbf{t})$ decreases make the following two steps:
  # Decomposition into weighted Voronoi cells.
  for j **in** $1..n$:
    Add X_j to the $\mathcal{C}(t_i)$ (for i as small as possible) satisfying
    $\|X_j - m(t_i)\|^2 + v(t_i) \leq \|X_j - m(t_l)\|^2 + v(t_l)\ \forall l \neq i$;
  # Computation of the new centers.
  for i **in** $1..k$:
    $t_i = \frac{1}{|\mathcal{C}(t_i)|} \sum_{X \in \mathcal{C}(t_i)} X$;
Output : $(t_1, t_2, \ldots, t_k)$

Proposition 22. *Algorithm 1 converges to a local minimum of*

$$R_n : \mathbf{t} \mapsto Q_n \min_{i \in [\![1,k]\!]} \left\| \cdot - m(Q_{nt_i,h}) \right\|^2 + v(Q_{nt_i,h}).$$

This result is a direct consequence of Proposition 9. Therefore, Algorithm 1 provides an approximation of the k-PDTM. Since the algorithm does not converge to the optimal centers, we suggest running the algorithm several times and storing the best solution in terms of the empirical cost R_n, as for k-means.

As mentioned above, Algorithm 1 may be thought of as a special instance of Lloyd's algorithm [34]. This algorithm consists in repeatedly decomposing the space $\mathbb{R}^d$ into cells associated to the t_i's, and then replacing the t_i's by the means of P restricted to the cells. This kind of algorithm provably outputs local minimizers of risks of the form $R_{\mathrm{d}} : \boldsymbol{\tau} \mapsto P \min_{i \in [\![1,k]\!]} \mathrm{d}(\cdot, \tau_i)$, for any Bregman divergence d (see, e.g., [6]). We recall that a Bregman divergence is defined by $\mathrm{d}(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle$, for some convex function ϕ. Actually, Lloyd's algorithm only works for Bregman divergences [5], since they are the only functionals d such that $c \mapsto P\mathrm{d}(\cdot, c)$ attains its minimum at $c = P\cdot$, the expectation of P. This suggests that our criterion R may be expressed in terms of some Bregman divergence.

For $P \in \mathcal{P}_{(2)}(\mathbb{R}^d)$, according to [16], Proposition 3.6, the function $\psi_{P,h} : x \mapsto \|x\|^2 - \mathrm{d}^2_{P,h}(x)$ is convex, and its set of subgradients at x is given by $\Delta_{x,h} = \{2m(P_{x,h}) \mid P_{x,h} \in \mathcal{P}_{x,h}(P)\}$. A simple computation based on (1) shows that the Bregman divergence associated with $\psi_{P,h}$ is defined for every $x, t \in \mathbb{R}^d$ by

$$\mathrm{d}_{\psi_{P,h}}(x, t) = \left\| x - m(P_{t,h}) \right\|^2 + v(P_{t,h}) - \mathrm{d}^2_{P,h}(x). \tag{10}$$

Since $\mathrm{d}^2_{P,h}(x)$ does not depend on t, our criterion R has the same minimizers as R_{d} for the Bregman divergence $\mathrm{d} = \mathrm{d}_{\psi_{P,h}}$. Thus, Proposition 9 is a consequence of the fact that $\psi_{P,h}$ is a Bregman divergence.

4. Numerical illustrations

4.1. Topological inference from noisy pointclouds

Let M be a compact subset of $\mathbb{R}^d$. Geometric and topological information about M can be recovered from some r-sublevel sets of the function distance to M, d_M (see, e.g., [16], Proposition 4.3). To tackle the tough question of the selection of r, or simply to track multiscale information, the concept of persistent homology has been introduced in [23]. It consists in describing the evolution of the homology (number of connected components, holes, etc.) of the sublevel sets of d_M. Persistent homology can be encoded via persistence diagrams. A persistence diagram is a multiset of points (b, d). Each point (b, d) is associated to one topological feature (a connected component, a hole, a void, etc.) that appears when $r = b$ (its birth time) and disappears when $r = d$ (its death time). As well, if $\|\hat{\mathrm{d}} - \mathrm{d}_M\|_\infty$ is small enough, then the persistence diagrams associated with $\hat{\mathrm{d}}$ and d_M will be provably close [19], that is the set of pairs (b, d) build from the sublevel sets of d_M and $\hat{\mathrm{d}}$ will be roughly similar. In particular, the lifetimes $d - b$ of the topological features will be close. To assess the relevancy of our approach in a noisy topological inference setting, we will compute the persistence diagrams associated with the empirical k-PDTM and its trimmed and truncated versions, and compare them with the outputs of other methods.

Following [29], we choose for M the infinity symbol embedded in $\mathbb{R}^2$. The persistence diagram associated to d_M is depicted in Figure 3. This diagram contains one red point $(0, \infty)$, that corresponds to the connected component (0-dimensional topological feature), and two green points that correspond to the two holes (1-dimensional topological features).

We generated a sample of 200 points, uniformly on the infinity symbol, with an additional additive Gaussian noise, with standard deviation $\sigma = 0.02$. This sample is corrupted by 80 outliers – 40 points generated according to the uniform distribution on the rectangle $[-2, 5] \times [-2, 2]$ and 40 points on the rectangle $[-4, 7] \times [-4, 4]$. This results in a corrupted sample $\mathbb{X}_n$ of 280 points. The persistence diagram associated with the sublevel sets of the empirical DTM function is represented in Figure 1 (left). Approximations of the persistence diagrams of the DTM for the uniform distribution P on M and the sampling distribution Q, from samples of size 5000 are also represented in Figure 1. The diagrams are computed with the function `gridDiag` of the `R` package `TDA` [25].

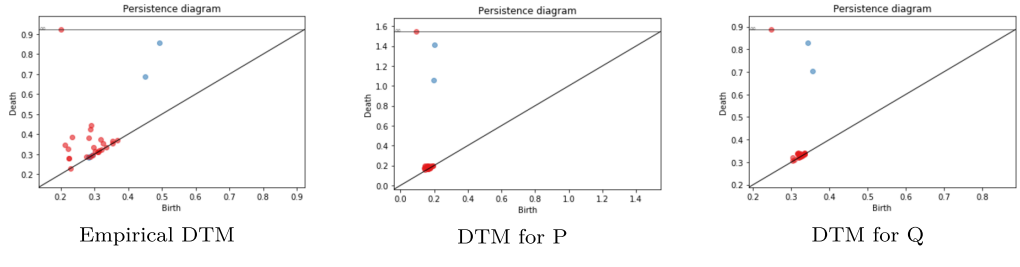

Figure 1. Persistence diagrams for the DTM and empirical DTM functions.

We compare several methods to recover relevant features of M from $\mathbb{X}_n$. Each method boils down to building an approximation f of d_M. These functions are of the type $f : x \mapsto \sqrt{\min_{i\in I} \|x - \tau_i\|^2 + \omega_i^2}$, for some finite set I, centers $\tau_i \in \mathbb{R}^d$ and weights $\omega_i \geq 0$. The first function we consider is derived from the k-means algorithm [37] ($|I| = k$, centers τ_i are given by the optima of the k-means criterion and $\omega_i = 0$), the second is the q-witnessed distance [29] ($|I| = n = 280$, it coincides with the k-PDTM for $k = n$, with mass parameter $h = q/n$), the third one is the k-PDTM ($|I| = k$, with mass parameter $h = q/n$). We also compare with the *power distance* [13], that chooses $|I| = n = 280$, τ_i as the i-th point of the sample $\mathbb{X}_n$, and ω_i^2 as the squared DTM at τ_i to the empirical distribution on $\mathbb{X}_n$, with mass parameter $h = q/n$. At last, we include in our comparative study the distance function to the *decluttered sample* $\tilde{\mathbb{X}}$, that is $\{\tau_i\}_{i\in I} = \tilde{\mathbb{X}}$, $\omega_i = 0$. This decluttered sample $\tilde{\mathbb{X}}$ is issued from the denoising procedure exposed in [14], with parameters 5.4 and 7.95, so that on average 200 points are considered as signal points by the procedure (i.e., $|I| \approx 200$ on average).

Most of these methods depend on two parameters q and k. Providing a method to calibrate q and k in general is beyond the scope of the paper. Here, we choose $q = 10$ and $k = 50$. Roughly, q is chosen small enough so that the distance to the q-th nearest neighbor remains small compared to the curvature of M but large enough to deal with noise, and k is chosen large enough so that a uniform grid with k points has grid size small compared to the curvature of M. More details on this heuristic can be found in the Appendix.

We implemented these methods with the `R` software. To be more specific, we used the function `kmeans`, the `FNN` library to compute nearest neighbors and the function `dtm` from the `TDA` package to compute the DTM. In Figure 2, we plotted the points of $\mathbb{X}_n$. Points are represented with the same color when they lie in the same weighted Voronoi cell (for the centers τ_i and weights ω_i^2). Centers τ_i are represented by triangles and colored in function of the weights ω_i^2 (black centers correspond to $\omega_i^2 = 0$). The second row of Fig-

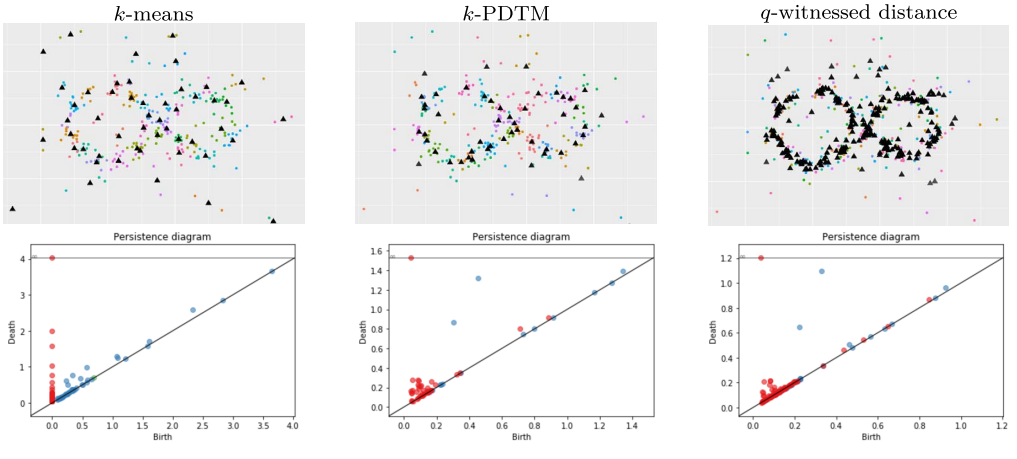

Figure 2. Comparison of the basic methods.

ure 2 depicts the corresponding persistence diagrams. They were obtained using the function `weighted_alpha_complex_3d_persistence` in the `Gudhi C++` library, based on alpha-shapes [22]. We observe on Figure 2 that the three main features of the symbol infinity (one connected component, two holes) are recovered for the k-PDTM and the q-witnessed distance, but not for k-means. As exposed in Lemma 4, this is due to the "void-filling" drawback of k-means. The persistence diagrams built from the power distance [13] and the distance to decluttered sample [14] are also quite similar to the k-PDTM and q-witnessed ones, succeeding in recovering the topological features of the infinity symbol. The corresponding illustrations can be found in Section D.1 of the Appendix.

4.2. Outliers detection

Another possible interest of the proposed method is outlier detection, based on the following principle: if an observation X_i is such that $\mathrm{d}_M(X_i)$ is large, then it can be considered as an outlier. Our denoising scheme consists in replacing d_M with an approximation $\hat{\mathrm{d}}$, then to remove points X_i such that $\hat{\mathrm{d}}(X_i)$ is large.

Such a procedure needs as an input a *level* α, that is the proportion of points that will be considered as signal points. Note that there exist heuristics to empirically design such an α (see, e.g., [11]). A level α being given, a straightforward approach consists in removing the $n(1-\alpha)$ points that corresponds to the largest values of $\hat{\mathrm{d}}$, for an estimate $\hat{\mathrm{d}}$ of d_M. In the following, we refer to this method as *truncation*, resulting in *truncated* k-means, *truncated* power distance, *truncated* q-witnessed distance and *truncated* k-PDTM.

However, it is possible to combine compression and denoising, by looking simultaneously for a subset of $n\alpha$ points (*trimming set*) and a set of k points that approximates the best the trimming set. For a non-negative function d, this corresponds to the minimization of the criterion $\boldsymbol{\tau} \mapsto \inf_{P_\alpha \in \mathcal{P}_\alpha(P)} P_\alpha \min_{i \in [\![1,k]\!]} \mathrm{d}(\cdot, \tau_i)$. Whenever d is a Bregman divergence, minimizers of such a criterion may be obtained via a Lloyd-type algorithm (see, e.g., [11]). Fortunately, since the k-means distance and the k-PDTM may be expressed via Bregman divergences, namely the squared Euclidean norm and (10), the procedure exposed in [11] applies. The outputs of the aforementioned procedure will be called *trimmed* k-means [20] and *trimmed* k-PDTM.

We experiment each of these methods for the dataset of the previous section (200 signal points around the infinity symbol, 80 ambient noise points). We choose $\alpha = 200$, $q = 10$ and $k = 50$. Figure 3 depicts, for the trimmed versions of the algorithms, the resulting partition signal/outliers along with the k centers and weights (shade of triangles) in the first row. The second row exposes the corresponding persistence diagrams. Similar illustrations for the *truncated* algorithms may be found in the Appendix.

The trimmed k-PDTM globally succeeds in identifying noise points and providing a relevant geometric approximation of the signal. To be more precise on the topological performances of the aforementioned methods, we repeated the experiment 100 times. At each time, we computed the lifetimes of the topological features and sorted them in decreasing order. Figure 4 below exposes the means of these lifetimes.

Methods based on the k-PDTM, the q-witnessed and the power distance, as well as the decluttering method of [14] are close to the ground truth. This is not the case for k-means-based

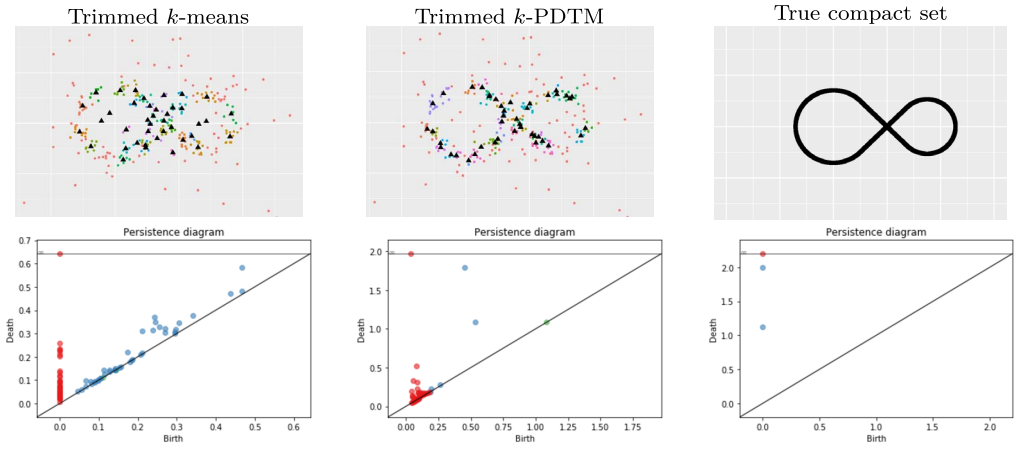

Figure 3. Comparison of the trimmed versions of the methods.

methods that add spurious holes (corresponding to the three last 1-dimensional features), and add many spurious connected components. In this case, spurious connected components are caused by centers located far from the support, whereas spurious holes are caused by centers located inside loops, breaking large loops into smaller loops. This phenomenon does not occur for k-PDTM-based methods since potentially damaging centers have a large weight. Consequently, such centers are either removed by truncation or appear lately in the sublevel set of the function. Thus, their impact on lifetimes of relevant features is weak. The first 0-dimensional feature, corresponding to the infinite connected component, has been removed in Figure 4. On the whole, our method compares well with the q-witnessed and the power-distance-based methods.

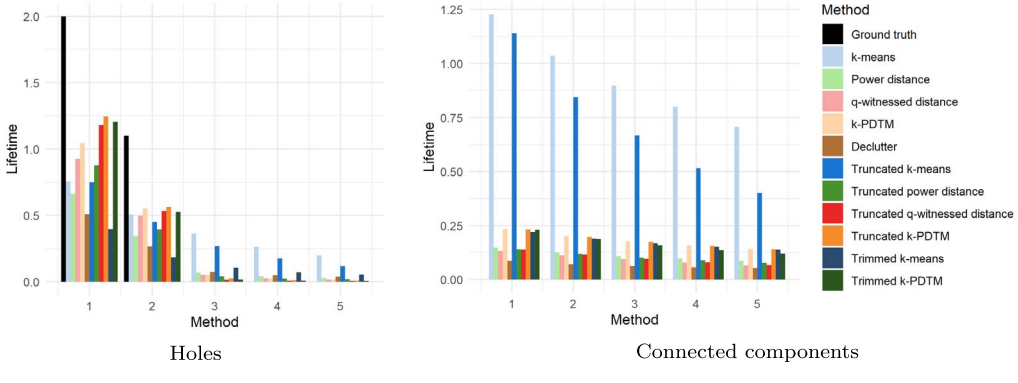

Figure 4. Features lifetimes.

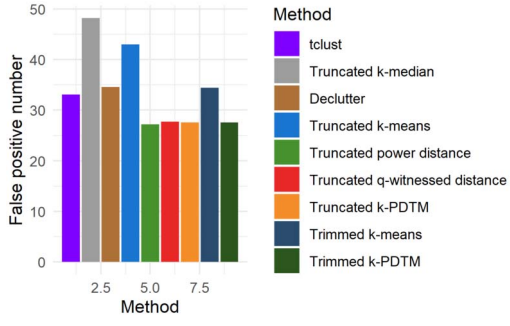

Figure 5. False positive number.

The diagram on Figure 5 depicts the mean amount of False positive over the 100 repetitions, that is the number of signal points that are labeled as outliers by the different algorithms. We also include comparison with other trimming approaches for outlier detection, such as tclust [27] (`tclust` function in `trimcluster R` library) and the truncated version of k-median [15] (`kGmedian` function of the `Gmedian R` library). Again, our method compares well with q-witnessed and power-distance-based denoising, contrary to the other methods.

5. Proofs for Section 2

5.1. Proof of Proposition 6

It suffices to prove that $\mathcal{P}_h(P)$ is a compact set (for the weak convergence metric) and that R is continuous (i.e., that $P_h \mapsto m(P_h)$ and $P_h \mapsto M(P_h)$ are continuous on $\mathcal{P}_h(P)$).

The set $h\mathcal{P}_h(P)$ is tight. Prokhorov's theorem entails that, for any sequence $(\mu_n/h)_{n\in\mathbb{N}}$ in $\mathcal{P}_h(P)$, up to a subsequence, there exists μ a Borel positive measure on $\mathbb{R}^d$ such that μ_n converges weakly to μ. The dominated convergence lemma applied to the functions $\mathbb{1}_{\mathbb{R}^d}$ and $\mathbb{1}_O$, for an open set O, ensures that $\mu(\mathbb{R}^d) = h$ and $\mu(O) \leq P(O)$, proving that $\mu \in h\mathcal{P}_h(P)$. Then, [4], page 438, yields that μ and P are regular measures. Thus, μ is a submeasure of P of mass h, and $\mathcal{P}_h(P)$ is compact.

We will now prove that the maps $P_h \mapsto m(P_h)$ and $P_h \mapsto M(P_h)$ are continuous on $\mathcal{P}_h(P)$. For $M > 0$ and $u \in \mathbb{R}^d$, denote by $u \wedge M$ the vector $(u_1 \wedge M, \ldots, u_d \wedge M)$, where $u_i \wedge M$ denotes $\min(u_i, M)$. Consider $(P_{h,n})_{n\in\mathbb{N}}$ a sequence in $\mathcal{P}_h(P)$ converging to some distribution P_h. Then there exists $M_\varepsilon > 0$ such that for every $P'_h \in \mathcal{P}_h(P)$,

$$\left\| P'_h(\cdot \wedge M_\varepsilon) - P'_h \cdot \right\| \leq \frac{P(\|\cdot\| \mathbb{1}_{\|\cdot\|_\infty > M_\varepsilon})}{h} \leq \varepsilon.$$

On the other hand, since $\cdot \mapsto \cdot \wedge M_\varepsilon$ is bounded and continuous, $\|P_{h,n}(\cdot \wedge M_\varepsilon) - P_h(\cdot \wedge M_\varepsilon)\|$ converges to 0. This proves the continuity of $P_h \mapsto m(P_h)$. We also have that $P_{h,n}\|\cdot\|^2 \wedge M \to$

$P_h\|\cdot\|^2 \wedge M$. Since for every $P'_h \in \mathcal{P}_h(P)$,

$$\left|P'_h\left(\|\cdot\|^2 \wedge M - \|\cdot\|^2\right)\right| \leq \frac{P\|\cdot\|^2 \mathbb{1}_{\|\cdot\|^2 > M}}{h}$$

and P has a finite second order moment, we deduce as well that $M(P_{h,n}) \to M(P_h)$.

5.2. Proof of Proposition 9

For short, we use the notation $m_i = m(\tilde{P}_{i,h})$, $v_i = v(\tilde{P}_{i,h})$ and $Q(du)f(u)$ for the expectation of f with respect to the Borel measure Q. Then, a bias-variance decomposition yields

$$\begin{aligned} R(P_1,\dots,P_k) &= P(du) \min_{i\in[\![1,k]\!]} P_i(dz)\|u-z\|^2 = \sum_{i=1}^k \tilde{P}_{i,h}(du)P_i(dz)\|u-z\|^2 \\ &= \sum_{i=1}^k \tilde{P}_{i,h}\left(\mathbb{R}^d\right)P_i(dz)\left(\|z-m_i\|^2+v_i\right) \geq \sum_{i=1}^k \tilde{P}_{i,h}\left(\mathbb{R}^d\right)Q_i(dz)\left(\|z-m_i\|^2+v_i\right) \\ &= \sum_{i=1}^k \tilde{P}_{i,h}(du)Q_i(dz)\|z-u\|^2, \end{aligned}$$

where $Q_i \in \mathcal{P}_{m_i,h}(P)$, and equality holds if and only if $P_i \in \mathcal{P}_{m_i,h}(P)$ or $\tilde{P}_{i,h}(\mathbb{R}^d) = 0$. Thus, denoting by $(\tilde{P}_{m_i,h})_{i\in[\![1,k]\!]}$ the set of weighted Voronoi measures associated to the measures $(Q_i)_{i\in[\![1,k]\!]}$, we have

$$R(P_1,\dots,P_k) \geq \sum_{i=1}^k \tilde{P}_{m_i,h}(du)Q_i(dz)\|z-u\|^2 = R(Q_1,\dots,Q_k).$$

5.3. Proof of Corollary 10

Let $(P_1^*,\dots,P_k^*)$ be a minimizer of R, and $\tilde{P}_{1,h},\dots,\tilde{P}_{k,h}$ be the associated weighted Voronoi measures provided by Definition 8. Then, according to Proposition 9, $f(m(\tilde{P}_{1,h}),\dots,m(\tilde{P}_{k,h}))$ is also an R-minimizer.

5.4. Proof of Lemma 11

Let $g(x,P_h) = M(P_h) - \|\tau\|^2 + 2\langle x, \tau - m(P_h)\rangle$. Then (4) entails that

$$\omega^2_{P,h}(\tau) = \sup_{x\in\mathbb{R}^d} \inf_{P_h\in\mathcal{P}_h(P)} g(x,P_h).$$

According to Section 5.1, $\mathcal{P}_h(P)$ is a compact set, and $P_h \mapsto m(P_h)$ as well as $P_h \mapsto M(P_h)$ are continuous. Note also that $P_h \mapsto m(P_h)$ and $P_h \mapsto M(P_h)$ are linear functions on the space of positive measures. So, for every $x \in \mathbb{R}^d$, $g(x,\cdot)$ is continuous and linear. On the other hand, for every P_h in $\mathcal{P}_h(P)$, $g(\cdot, P_h)$ is linear and continuous. Sion's theorem [32] yields that

$$\omega^2_{P,h}(\tau) = \min_{P_h \in \mathcal{P}_h(P)} \sup_{x \in \mathbb{R}^d} M(P_h) - \|\tau\|^2 + 2\big\langle x, \tau - m(P_h)\big\rangle. \tag{11}$$

Therefore, $\omega^2_{P,h}(\tau) < \infty$ is equivalent to $\tau \in \tilde{\mathcal{M}}_h(P)$. Now let τ be in $\tilde{\mathcal{M}}_h(P)$. According to (11), we have

$$\omega^2_{P,h}(\tau) = \inf_{P_h \in \mathcal{P}_h(P), m(P_h)=\tau} M(P_h) - \|\tau\|^2 = \inf_{P_h \in \mathcal{P}_h(P), m(P_h)=\tau} v(P_h).$$

Since $P_h \mapsto v(P_h)$ is continuous on $\mathcal{P}_h(P)$ and $\mathcal{P}_h(P) \cap m^{-1}(\{\tau\})$ is compact, there exists P_h such that $m(P_h) = \tau$ and $\omega^2_{P,h}(\tau) = v(P_h)$.

5.5. Proof of Theorem 12

Let $\tilde{R}$ denote $\boldsymbol{\tau} \mapsto P \min_{j \in [\![1,k]\!]} \|\cdot - \tau_j\|^2 + \omega^2_{P,h}(\tau_j)$, for $\boldsymbol{\tau} \in (\mathbb{R}^d)^{(k)}$. Lemma 11 ensures that $\min_{\mathcal{P}_h(P)^k} R(P_1, \dots, P_k) = \min_{(\mathbb{R}^d)^{(k)}} \tilde{R}(\boldsymbol{\tau})$.

Assume that $\mathbf{t}$ is such that $(m(P_{t_1,h}), \dots, m(P_{t_k,h})) \in \arg\min_{\boldsymbol{\tau}} \tilde{R}(\boldsymbol{\tau})$, and, for short, denote by $\omega_i^2 = \omega^2_{P,h}(m(P_{t_i,h}))$. For a fixed i, if $\omega_i^2 < v(P_{t_i,h})$, then Lemma 11 provides $P_i' \in \mathcal{P}_h(P)$ such that $m(P_i') = m(P_{t_i,h})$ and $v(P_i') = \omega_i^2 < v(P_{t_i,h})$. Thus

$$P_i'(du)\|u - t_i\|^2 = \big\|t_i - m(P_{t_i,h})\big\|^2 + \omega_i^2 < P_{t_i,h}(du)\|u - t_i\|^2,$$

hence the contradiction. Thus, $\omega_i^2 = v(P_{t_i,h})$, $\tilde{R}(m(P_{t_1,h}), \dots, m(P_{t_k,h})) = R(P_{t_1,h}, \dots, P_{t_k,h})$, and $(P_{t_1,h}, \dots, P_{t_k,h})$ minimizes R.

Conversely, assume that $(P_{t_1,h}, \dots, P_{t_k,h})$ minimizes R. Then $\tilde{R}(m(P_{t_1,h}), \dots, m(P_{t_k,h})) \le R(P_{t_1,h}, \dots, P_{t_k,h}) = \min_{(\mathbb{R}^d)^{(k)}} \tilde{R}(\boldsymbol{\tau})$, according to Lemma 11. Thus, $(m(P_{t_1,h}), \dots, m(P_{t_k,h}))$ minimizes $\tilde{R}$.

6. Proofs for Section 3

6.1. Intermediate results

The proofs of the Section 3 results will make intensive use of the following lemmas, whose proofs are postponed to the Appendix. We first mention some well-known results about sub-Gaussian distributions.

Lemma 23. *Let $Q \in \mathcal{P}^{(V)}(\mathbb{R}^d)$, a sub-Gaussian measure with variance $V^2 > 0$, and $Q_h \in \mathcal{P}_h(Q)$. Then we have*

$$Q_h \|\cdot\|^2 \leq \frac{3V^2}{h}.$$

The proof of Lemma 23 is given in Section A.2 of the Appendix. Next, Lemma 24 below ensures that the distributions involved in Theorem 19 are sub-Gaussian.

Lemma 24. *If Y is a random variable sampled from a distribution P in $\mathcal{P}^K(\mathbb{R}^d)$ and Z is independent from Y and sampled from a distribution Q' in $\mathcal{P}^{(\sigma)}(\mathbb{R}^d)$ for some $\sigma > 0$. Then, the distribution Q of the random variable $X = Y + Z$ is sub-Gaussian with variance $V^2 = (K+\sigma)^2$, that is in $\mathcal{P}^{(K+\sigma)}(\mathbb{R}^d)$.*

Moreover,

$$W_1(P, Q) \leq 3\sigma \quad \text{and} \quad W_2(P, Q) \leq \sqrt{3}\sigma.$$

A proof of Lemma 24 can be found in Section A.3, Appendix. In what follows, we let γ and $\hat{\gamma}$ denote the functions

$$\begin{aligned} \gamma(\mathbf{t}, x) &= \min_{i \in [\![1,k]\!]} -2\langle x, m(Q_{t_i,h})\rangle + \left\|m(Q_{t_i,h})\right\|^2 + v(Q_{t_i,h}), \\ \hat{\gamma}(\mathbf{t}, x) &= \min_{i \in [\![1,k]\!]} -2\langle x, m(Q_{nt_i,h})\rangle + \left\|m(Q_{nt_i,h})\right\|^2 + v(Q_{nt_i,h}), \end{aligned} \tag{12}$$

for $(\mathbf{t}, x) \in (\mathbb{R}^d)^{(k)} \times \mathbb{R}^d$ with $\mathbf{t} = (t_1, t_2, \ldots, t_k)$. We will use two deviation bounds, stated below.

Lemma 25. *If Q is sub-Gaussian with variance V^2, then, for every $p > 0$, with probability larger than $1 - n^{-p}$, we have*

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left|(Q - Q_n)\gamma(\mathbf{t}, \cdot)\right| \leq C \frac{V^2 \sqrt{k \log(k) d}(1+p)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}},$$

for some absolute positive constant C.

The proof of Lemma 25 is deferred to Section B.5 of the Appendix.

Lemma 26. *Assume that Q is sub-Gaussian with variance V^2, then, for every $p > 0$, with probability larger than $1 - 9n^{-p}$, we have*

$$\sup_{t \in \mathbb{R}^d} \left\|m(Q_{t,h}) - m(Q_{nt,h})\right\| \leq \frac{CV\sqrt{d}(p+1)\log(n)}{h\sqrt{n}},$$

$$\sup_{\mathbf{t} \in (\mathbb{R}^d)^{(k)}} \left|Q_n(\gamma - \hat{\gamma})(\mathbf{t}, \cdot)\right| \leq CV^2 \frac{\sqrt{d}(p+1)^{\frac{3}{2}} \log(n)^{\frac{3}{2}}}{h\sqrt{n}}.$$

As well, the proof of Lemma 26 is deferred to Section B.4 in the Appendix.

6.2. Proof of Proposition 14

The first inequality comes from (4). We now focus on the second bound. By definition of $\mathrm{d}_{P,h,k}$, for all $\mathbf{t} = (t_1, t_2, \ldots, t_k) \in (\mathbb{R}^d)^{(k)}$ we have $P\mathrm{d}^2_{P,h,k} \leq P \min_{i\in[\![1,k]\!]} \|\cdot - m(P_{t_i,h})\|^2 + v(P_{t_i,h})$. Thus,

$$P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) \leq P\Big(\min_{i\in[\![1,k]\!]} \big\|\cdot - m(P_{t_i,h})\big\|^2 + v(P_{t_i,h}) - \mathrm{d}^2_{P,h}\Big)$$
$$= P\Big(\min_{i\in[\![1,k]\!]} \big(\mathrm{d}^2_{P,h}(t_i) - \|t_i\|^2\big) - \big(\mathrm{d}^2_{P,h} - \|\cdot\|^2\big) + \big\langle \cdot - t_i, -2m(P_{t_i,h})\big\rangle\Big),$$

according to (1). Now [16], Corollary 3.7, ensures that, for x, y in $\mathbb{R}^d$,

$$\|y\|^2 - \mathrm{d}^2_{P,h}(y) - \big(\|x\|^2 - \mathrm{d}^2_{P,h}(x)\big) \geq 2\big\langle m(P_{x,h}), y - x\big\rangle. \tag{13}$$

We deduce that

$$P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) \leq P \min_{i\in[\![1,k]\!]} 2\big\langle \cdot - t_i, m(P_{\cdot,h}) - m(P_{t_i,h})\big\rangle$$
$$\leq 2P \min_{i\in[\![1,k]\!]} \|\cdot - t_i\| \big\|m(P_{\cdot,h}) - m(P_{t_i,h})\big\|.$$

Now choose $t_1, \ldots, t_k$ such that $M \subset \bigcup_{i\in[\![1,k]\!]} \mathrm{B}(t_i, f_M^{-1}(k))$. The result follows.

6.3. Proof of Corollary 15

The proof of Corollary 15 is based on the following bounds, in the case where P is absolutely continuous with respect to the Lebesgue measure, with density f satisfying $0 < f_{\min} \leq f \leq f_{\max}$.

$$f_M^{-1}(k) \leq 2K\sqrt{d}k^{-1/d}, \tag{14}$$

$$\zeta_{P,h}\big(f_M^{-1}(k)\big) \leq C_{f_{\max},K,d,h} k^{-1/d}. \tag{15}$$

First, note that since $M \subset \mathrm{B}(0, K)$, for any $\varepsilon > 0$, $f_M(\varepsilon) \leq f_{\mathrm{B}(0,K)}(\varepsilon) \leq (\frac{2K\sqrt{d}}{\varepsilon})^d$, hence (14). To prove the second inequality, we have to give a bound on the modulus of continuity $\zeta_{P,h}$. Let x, y be in M, and denote by $\delta = \|x - y\|$. Since P has a density,

$$P\partial\mathrm{B}\big(x, \delta_{P,h}(x)\big) = P\partial\mathrm{B}\big(y, \delta_{P,h}(y)\big) = 0.$$

We deduce that $P_{x,h} = \frac{1}{h} P_{|\mathrm{B}(x,\delta_{P,h}(x))}$ and $P_{y,h} = \frac{1}{h} P_{|\mathrm{B}(y,\delta_{P,h}(y))}$. Without loss of generality, assume that $\delta_{P,h}(x) \geq \delta_{P,h}(y)$. Then

$$\mathrm{B}\big(y, \delta_{P,h}(y)\big) \subset \mathrm{B}\big(x, \delta_{P,h}(x) + \delta\big).$$

We may bound $\|m(P_{x,h}) - m(P_{y,h})\|$ by

$$\begin{aligned}
\frac{1}{h}\big\| P \cdot (\mathbb{1}_{\mathrm{B}(x,\delta_{P,h}(x))} - \mathbb{1}_{\mathrm{B}(y,\delta_{P,h}(y))})\big\| &\le \frac{K}{h} P|\mathbb{1}_{\mathrm{B}(x,\delta_{P,h}(x))} - \mathbb{1}_{\mathrm{B}(y,\delta_{P,h}(y))}| \\
&= 2\frac{K}{h} P\big(\mathrm{B}\big(y,\delta_{P,h}(y)\big) \setminus \big(\mathrm{B}\big(x,\delta_{P,h}(x)\big) \cap \mathrm{B}\big(y,\delta_{P,h}(y)\big)\big)\big) \\
&\le 2\frac{K}{h} P\big(\mathrm{B}\big(x,\delta_{P,h}(x)+\delta\big) \cap \mathrm{B}\big(x,\delta_{P,h}(x)\big)^c\big) \\
&= 2\frac{K}{h}\omega_d\big[\big(\delta_{P,h}(x)+\delta\big)^d - \delta_{P,h}(x)^d\big] \le 2\frac{K^{d+1}\omega_d}{h}\left[\left(1+\frac{\delta}{\delta_{P,h}(x)}\right)^d - 1\right],
\end{aligned}$$

where ω_d denotes the Lebesgue volume of the ball $\mathrm{B}(0,1)$ in $\mathbb{R}^d$. Since $(1+v)^d \le 1 + d(1+v)^{d-1}v$, for $v \ge 0$, and $\delta_{P,h}(x) \ge (\frac{h}{f_{\max}\omega_d})^{1/d}$, we have $\zeta_{P,h}(\delta) \le C_{f_{\max},K,d,h}\delta$, hence (15). The result of Corollary 15 follows.

6.4. Proof of Corollary 16

Since N is a $\mathcal{C}^2$-submanifold, its reach ρ (as defined in [26], Definition 4.1) is positive. Without loss of generality we assume that N is connected. Since P has a density with respect to the volume measure on N, we have $P(N^\circ) = 1$. Thus, we take $M = N^\circ$, that is the set of interior points. Since P satisfies a $(cf_{\min}, d')$-standard assumption, we have

$$f_M(\varepsilon) \le 2^{d'}/\big(cf_{\min}r^{-d'}\big),$$

according to [17], Lemma 10. Hence,

$$f_M^{-1}(k) \le C_{f_{\min},N}k^{-1/d'}.$$

It remains to bound the continuity modulus of $x \mapsto m(P_{x,h})$. For any x in M, since $P(\partial N) = 0$ and P has a density with respect to the volume measure on N, we have $P_{x,h} = P_{|\mathrm{B}(x,\delta_{P,h}(x))}$. Besides, since for all $r > 0$, $P(\mathrm{B}(x,r)) \ge cf_{\min}r^{d'}$, we may write $\delta_{P,h}(x) \le c_{N,f_{\min}}h^{1/d'} \le \rho/12$, for h small enough. Now let x and y be in M so that $\|x-y\| = \delta \le \rho/12$, and without loss of generality assume that $\delta_{P,h}(x) \ge \delta_{P,h}(y)$. Then, proceeding as in the proof of (15), it comes

$$\big\|m(P_{x,h}) - m(P_{y,h})\big\| \le \frac{2K}{h} P\big(\mathrm{B}\big(x,\delta_{P,h}(x)+\delta\big) \cap \mathrm{B}\big(x,\delta_{P,h}(x)\big)^c\big).$$

Since $\delta_{P,h}(x) + \delta \le \rho/6$, for any u in $\mathrm{B}(x,\delta_{P,h}(x)+\delta) \cap M$ we may write $u = \exp_x(rv)$, where $v \in T_xM$ with $\|v\| = 1$ and $r = d_N(u,x)$ is the geodesic distance between u and x (see, e.g., [26], Theorem 4.18 or [1], Proposition 25). Note that, according to [1], Proposition 26, for any u_1 and u_2 such that $\|u_1 - u_2\| \le \rho/4$,

$$\|u_1 - u_2\| \le d_N(u_1,u_2) \le 2\|u_1 - u_2\|. \tag{16}$$

Now let $p_1, \dots, p_m$ be a δ-covering of the sphere $\mathrm{S}(x, \delta_{P,h}(x)) = \overline{B}(x, \delta_{P,h}(x)) \setminus B(x, \delta_{P,h}(x))$. According to (16), we may choose $m \leq c_{d'} \delta_{P,h}(x)^{d'-1} \delta^{-(d'-1)}$.

Now, for any u such that $u \in M$ and $\delta_{P,h}(x) \leq \|x - u\| \leq \delta_{P,h}(x) + \delta$, there exists $t \in \mathrm{S}(x, \delta_{P,h}(x))$ such that $\|t - u\| \leq 2\delta$. Hence,

$$P\big(\mathrm{B}\big(x, \delta_{P,h}(x) + \delta\big) \cap \mathrm{B}\big(x, \delta_{P,h}(x)\big)^c\big) \leq \sum_{j=1}^{m} P\big(\mathrm{B}(p_j, 2\delta)\big).$$

For any j, since $2\delta \leq \rho/6$, in local polar coordinates around p_j we may write,

$$P\big(\mathrm{B}(p_j, 2\delta)\big) \leq \int_{\{r,v \,|\, \exp_{p_j}(rv) \in M, r \leq 4\delta\}} f(r,v) J(r,v)\, dr\, dv \leq f_{\max} \int_{\{r,v | r \leq 4\delta\}} J(r,v)\, dr\, dv,$$

using (16), where $J(r,v)$ denotes the Jacobian of the volume form. According to [1], Proposition 27, we have $J(r,v) \leq C_{d'} r^{d'}$, hence $P(\mathrm{B}(p_j, 2\delta)) \leq C_{d'} f_{\max} \delta^{d'}$. Thus,

$$\big\|m(P_{x,h}) - m(P_{y,h})\big\| \leq \frac{2K}{h} m C_{d'} f_{\max} \delta^{d'} \leq C_{N, f_{\max}, f_{\min}} \delta.$$

Choosing k large enough so that $f_M^{-1}(k) \leq C_{f_{\min}, N} k^{-1/d'} \leq \rho/12$ gives the result.

6.5. Proof of Proposition 17

For all $x \in \mathrm{Supp}(P)$,

$$\begin{aligned} \mathrm{d}^2_{Q,h,k}(x) - \mathrm{d}^2_{P,h}(x) &= \mathrm{d}^2_{Q,h,k}(x) - \mathrm{d}^2_{Q,h}(x) + \mathrm{d}^2_{Q,h}(x) - \mathrm{d}^2_{P,h}(x) \\ &\geq -\big\|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\big\|_{\infty, \mathrm{Supp}(P)}. \end{aligned}$$

Thus, $(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h})_- \leq \|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\|_{\infty, \mathrm{Supp}(P)}$ on $\mathrm{Supp}(P)$, where $f_- : x \mapsto f(x) \mathbb{1}_{f(x) \leq 0}$ denotes the negative part of any function f on $\mathbb{R}^d$. Then,

$$\begin{aligned} P\big|\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\big| &= P\big(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\big) + 2\big(\mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h}\big)_- \\ &\leq P\Delta + P\big(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\big) + 2\big\|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\big\|_{\infty, \mathrm{Supp}(P)}, \end{aligned}$$

with $\Delta = \mathrm{d}^2_{Q,h,k} - \mathrm{d}^2_{P,h,k}$. To bound $P\Delta$ from above, let $\mathbf{s} \in (\overline{B}(0,K))^{(k)}$ be a k-points minimizer of R for P, such that when $\tilde{P}_{s_i,h}(\mathbb{R}^d) \neq 0$, $s_i = m(\tilde{P}_{s_i,h})$. Such an $\mathbf{s}$ exists according to Proposition 9 and Lemma 11. Set $f_{Q,t}(x) = -2\langle x, m(Q_{t,h})\rangle + M(Q_{t,h})$ for $t \in \mathbb{R}^d$, and let $\mathbf{t}$ be a k-points minimizer of R for Q.

$$\begin{aligned} P\Delta &= P\Big(\min_{i \in [\![1,k]\!]} f_{Q,t_i} - \min_{i \in [\![1,k]\!]} f_{P,s_i}\Big) \\ &\leq (P - Q) \min_{i \in [\![1,k]\!]} f_{Q,t_i} + (Q - P) \min_{i \in [\![1,k]\!]} f_{Q,s_i} + P\Big(\min_{i \in [\![1,k]\!]} f_{Q,s_i} - \min_{i \in [\![1,k]\!]} f_{P,s_i}\Big). \end{aligned}$$

For a transport plan π between P and Q, $(P-Q)\min_{i\in[\![1,k]\!]} f_{Q,t_i}$ is bounded by

$$\mathbb{E}_{(X,Y)\sim\pi}\Big[\min_{i\in[\![1,k]\!]} -2\big\langle X, m(Q_{t_i,h})\big\rangle + M(Q_{t_i,h}) - \min_{i\in[\![1,k]\!]} -2\big\langle Y, m(Q_{t_i,h})\big\rangle + M(Q_{t_i,h})\Big]$$
$$\leq 2\mathbb{E}_{(X,Y)\sim\pi}\Big[\sup_{t\in\mathbb{R}^d}\big\langle Y-X, m(Q_{t,h})\big\rangle\Big].$$

Thus, $(P-Q)\min_{i\in[\![1,k]\!]} f_{Q,t_i} \leq 2W_1(P,Q)\sup_{t\in\mathbb{R}^d}\|m(Q_{t,h})\|$, choosing for π the optimal transport plan for the W_1 distance between P and Q in (3). Also note that $P(\min_{i\in[\![1,k]\!]} f_{Q,s_i} - \min_{i\in[\![1,k]\!]} f_{P,s_i})$ is bounded from above by

$$\sum_{i=1}^{k} \tilde{P}_{s_i,h}\big(-2\big\langle\cdot, m(Q_{s_i,h})\big\rangle + M(Q_{s_i,h}) + 2\big\langle\cdot, m(P_{s_i,h})\big\rangle - M(P_{s_i,h})\big)$$
$$= \sum_{i=1}^{k} \tilde{P}_{s_i,h} 2\big\langle\cdot - s_i, m(P_{s_i,h}) - m(Q_{s_i,h})\big\rangle + \mathrm{d}^2_{Q,h}(s_i) - \mathrm{d}^2_{P,h}(s_i)$$
$$\leq \big\|\mathrm{d}^2_{P,h} - \mathrm{d}^2_{Q,h}\big\|_{\infty,\mathrm{B}(0,K)} + 2\sum_{i=1}^{k} \tilde{P}_{s_i,h}\big(\mathbb{R}^d\big)\big\langle m(\tilde{P}_{s_i,h}) - s_i, m(P_{s_i,h}) - m(Q_{s_i,h})\big\rangle.$$

Since $s_i = m(\tilde{P}_{s_i,h})$, the result follows.

6.6. Proof of Proposition 18

Let $\Delta_{\infty,K}$ denote $\sup_{x\in M} \mathrm{d}_{Q,h,k}(x)$, and let $x\in M$ achieving the maximum distance. Since $\mathrm{d}_{Q,h,k}$ is 1-Lipschitz, we deduce that $\mathrm{B}(x, \frac{\Delta_{\infty,K}}{2}) \subset \{y \mid \mathrm{d}_{Q,k,h}(y) \geq \frac{\Delta_{\infty,K}}{2}\}$. Since $P(\mathrm{B}(x, \frac{\Delta_{\infty,K}}{2})) \geq C(P)(\frac{\Delta_{\infty,K}}{2})^{d'} \wedge 1$, Markov inequality yields that

$$\Delta_P^2 \geq C(P)\left(\frac{\Delta_{\infty,K}}{2}\right)^{d'+2} \wedge \frac{\Delta^2_{\infty,K}}{4}.$$

Thus we have $\sup_{x\in M}(\mathrm{d}_{Q,h,k} - \mathrm{d}_M)(x) = \Delta_{\infty,K} \leq C(P)^{-\frac{1}{d'+2}} \Delta_P^{\frac{2}{d'+2}} \vee 2\Delta_P$. Now, for $x\in\mathbb{R}^d$, we let $p\in M$ such that $\|x-p\| = \mathrm{d}_M(x)$. Denote by $r = \|x-p\|$, and let t_j be such that $\mathrm{d}_{Q,h,k}(p) = \sqrt{\|p - m(Q_{t_j,h})\|^2 + v(Q_{t_j,h})}$. Then

$$\begin{aligned}
\mathrm{d}_{Q,h,k}(x) &\leq \sqrt{\big\|x - m(Q_{t_j,h})\big\|^2 + v(Q_{t_j,h})} \\
&\leq \sqrt{\mathrm{d}^2_{Q,h,k}(p) + r^2 + 2r\big\|p - m(Q_{t_j,h})\big\|} \\
&\leq \sqrt{\mathrm{d}^2_{Q,h,k}(p) + r^2 + 2r\mathrm{d}_{Q,h,k}(p)} \\
&= \mathrm{d}_M(x) + \big(\mathrm{d}_{Q,h,k}(p) - \mathrm{d}_M(p)\big).
\end{aligned}$$

Hence, $\sup_{x\in\mathbb{R}^d}(\mathrm{d}_{Q,h,k}-\mathrm{d}_M)(x)=\sup_{x\in M}(\mathrm{d}_{Q,h,k}-\mathrm{d}_M)(x)=\Delta_{\infty,K}$. On the other hand, we have $\mathrm{d}_{Q,h,k}\geq \mathrm{d}_{Q,h}$, along with $\|\mathrm{d}_{Q,h}-\mathrm{d}_{P,h}\|_\infty\leq h^{-\frac{1}{2}}W_2(P,Q)$ (see, e.g., [16], Theorem 3.5) as well as $\mathrm{d}_{P,h}\geq \mathrm{d}_M$. Hence $\mathrm{d}_{Q,h,k}\geq \mathrm{d}_M-h^{-\frac{1}{2}}W_2(P,Q)$.

6.7. Proof of Theorem 19

We recall that γ and $\hat{\gamma}$ are defined in (12). According to Lemma 24, $Q\in\mathcal{P}^{(V)}(\mathbb{R}^d)$ with $V=\sigma+K$. Let

$$\mathbf{s}=\arg\min\big\{Q\gamma(\mathbf{t},\cdot)\mid \mathbf{t}=(t_1,t_2,\dots,t_k)\in(\mathbb{R}^d)^{(k)}\big\},$$
$$\hat{\mathbf{s}}=\arg\min\big\{Q_n\hat{\gamma}(\mathbf{t},\cdot)\mid \mathbf{t}=(t_1,t_2,\dots,t_k)\in(\mathbb{R}^d)^{(k)}\big\},$$
$$\tilde{\mathbf{s}}=\arg\min\big\{Q_n\gamma(\mathbf{t},\cdot)\mid \mathbf{t}=(t_1,t_2,\dots,t_k)\in(\mathbb{R}^d)^{(k)}\big\}.$$

With these notations, for all $x\in\mathbb{R}^d$,

$$\mathrm{d}^2_{Q,h,k}(x)=\|x\|^2+\gamma(\mathbf{s},x)\quad\text{and}\quad \mathrm{d}^2_{Q_n,h,k}(x)=\|x\|^2+\hat{\gamma}(\hat{\mathbf{s}},x).$$

We intend to bound $l(\mathbf{s},\hat{\mathbf{s}})=Q(\mathrm{d}^2_{Q_n,h,k}-\mathrm{d}^2_{Q,h,k})=Q(\gamma(\hat{\mathbf{s}},\cdot)-\gamma(\mathbf{s},\cdot))$.

$$\begin{aligned}l(\mathbf{s},\hat{\mathbf{s}})&=Q\gamma(\hat{\mathbf{s}},\cdot)-Q_n\gamma(\hat{\mathbf{s}},\cdot)+Q_n\gamma(\hat{\mathbf{s}},\cdot)-Q_n\gamma(\tilde{\mathbf{s}},\cdot)+Q_n\gamma(\tilde{\mathbf{s}},\cdot)-Q\gamma(\mathbf{s},\cdot)\\&\leq\sup_{\mathbf{t}\in(\mathbb{R}^d)^{(k)}}(Q-Q_n)\gamma(\mathbf{t},\cdot)+Q_n(\gamma-\hat{\gamma})(\hat{\mathbf{s}},\cdot)\\&\quad+Q_n\big(\hat{\gamma}(\hat{\mathbf{s}},\cdot)-\hat{\gamma}(\tilde{\mathbf{s}},\cdot)\big)+Q_n(\hat{\gamma}-\gamma)(\tilde{\mathbf{s}},\cdot)+\sup_{\mathbf{t}\in(\mathbb{R}^d)^{(k)}}(Q_n-Q)\gamma(\mathbf{t},\cdot),\end{aligned}$$

where we used $Q_n\gamma(\tilde{\mathbf{s}},\cdot)\leq Q_n\gamma(\mathbf{s},\cdot)$. Now, since $Q_n(\hat{\gamma}(\hat{\mathbf{s}},\cdot)-\hat{\gamma}(\tilde{\mathbf{s}},\cdot))\leq 0$, we get

$$\begin{aligned}l(\mathbf{s},\hat{\mathbf{s}})\leq&\sup_{\mathbf{t}\in(\mathbb{R}^d)^{(k)}}(Q-Q_n)\gamma(\mathbf{t},\cdot)+\sup_{\mathbf{t}\in(\mathbb{R}^d)^{(k)}}(Q_n-Q)\gamma(\mathbf{t},\cdot)\\&+\sup_{\mathbf{t}\in(\mathbb{R}^d)^{(k)}}Q_n(\gamma-\hat{\gamma})(\mathbf{t},\cdot)+\sup_{\mathbf{t}\in\mathbb{R}^{d(k)}}Q_n(\hat{\gamma}-\gamma)(\mathbf{t},\cdot).\end{aligned}$$

Combining Lemma 25 and Lemma 26 entails, with probability larger than $1-10n^{-p}$,

$$l(\mathbf{s},\hat{\mathbf{s}})\leq CV^2\sqrt{k\log(k)d}\,\frac{(p+1)^{\frac{3}{2}}\log(n)^{\frac{3}{2}}}{h\sqrt{n}}.$$

It remains to bound $|P\mathrm{d}^2_{Q_n,h,k}-Q\mathrm{d}^2_{Q_n,h,k}|$ as well as $|P\mathrm{d}^2_{Q,h,k}-Q\mathrm{d}^2_{Q,h,k}|$. To this aim we recall that $X=Y+Z$, Z being sub-Gaussian with variance σ^2. Thus, denoting by $s_j(x)=\arg\min_{j\in[\![1,k]\!]}\|x-m(Q_{s_j,h})\|^2+v(Q_{s_j,h})$,

$$\begin{aligned}P\mathrm{d}^2_{Q,h,k}-Q\mathrm{d}^2_{Q,h,k}\leq\mathbb{E}_{(Y,Z)}\big[&\big\|Y-m(Q_{s_j(Y+Z),h})\big\|^2+v(Q_{s_j(Y+Z),h})\\&-\big(\big\|Y+Z-m(Q_{s_j(Y+Z),h})\big\|^2+v(Q_{s_j(Y+Z),h})\big)\big]\end{aligned}$$

$$\leq \mathbb{E}_Z\|Z\|^2 + 2\mathbb{E}_{(Y,Z)} \max_{j\in[\![1,k]\!]} \left|\langle Z, m(Q_{s_j,h}) - Y\rangle\right|$$
$$\leq 3\sigma^2 + 2\sqrt{3}\sigma\left(\max_{j\in[\![1,k]\!]} \left\|m(Q_{s_j,h})\right\| + K\right) \leq \frac{C\sigma K}{\sqrt{h}},$$

using Cauchy–Schwarz inequality, Lemma 23 and $\sigma \leq K$.

The converse bound on $Q\mathrm{d}^2_{Q,h,k} - P\mathrm{d}^2_{Q,h,k}$ may be proved the same way. Similarly, we may write

$$\begin{aligned}
&P\mathrm{d}^2_{Q_n,h,k} - Q\mathrm{d}^2_{Q_n,h,k}\\
&\quad\leq 3\sigma^2 + 2\sqrt{3}\sigma\left(\max_{j\in[\![1,k]\!]} \left\|m(Q_{ns_j,h})\right\| + K\right)\\
&\quad\leq 3\sigma^2 + 2\sqrt{3}\sigma\left(\max_{j\in[\![1,k]\!]} \left\|m(Q_{s_j,h})\right\| + \sup_{t\in\mathbb{R}^d} \left\|m(Q_{t,h}) - m(Q_{nt,h})\right\| + K\right)\\
&\quad\leq 3\sigma^2 + 2\sqrt{3}\sigma\left(\max_{j\in[\![1,k]\!]} \left\|m(Q_{s_j,h})\right\| + C(K+\sigma)\sqrt{d}\frac{(p+1)\log(n)}{h\sqrt{n}} + K\right)\\
&\quad\leq \frac{C\sigma K}{\sqrt{h}} + \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}},
\end{aligned}$$

according to Lemma 23 and Lemma 26. The bound on $Q\mathrm{d}^2_{Q_n,h,k} - P\mathrm{d}^2_{Q_n,h,k}$ derives from the same argument. Collecting all pieces, we get, using $\sigma \leq K$,

$$\begin{aligned}
\left|P\left(d^2_{Q_n,h,k} - d^2_{Q,h,k}\right)\right| &\leq \left|Q\left(d^2_{Q_n,h,k} - d^2_{Q,h,k}\right)\right| + \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}} + \frac{C\sigma K}{\sqrt{h}}\\
&\leq \frac{C\sigma K\sqrt{d}(p+1)\log(n)}{h\sqrt{n}} + \frac{CkK^2\sqrt{dk\log(k)}((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}}\\
&\quad + \frac{C\sigma K}{\sqrt{h}}.
\end{aligned}$$

6.8. Proof of Proposition 20

Combining bounds obtained in Theorem 19 and Proposition 17 yields

$$\begin{aligned}
\left|P\left(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{P,h}\right)\right| &\leq C\sqrt{k\log(k)d}\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}}\\
&\quad + 3\left\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\right\|_{\infty,\mathrm{B}(0,K)} + P\left(\mathrm{d}^2_{P,h,k} - \mathrm{d}^2_{P,h}\right)\\
&\quad + 4W_1(P,Q)\sup_{s\in\mathbb{R}^d}\left\|m(P_{s,h})\right\|.
\end{aligned}$$

Using Corollary 16 and Lemma 23 entails

$$\left|P\left(\mathrm{d}^2_{Q_n,h,k} - \mathrm{d}^2_{P,h}\right)\right| \leq C\sqrt{k\log(k)d}\frac{K^2((p+1)\log(n))^{\frac{3}{2}}}{h\sqrt{n}} + C\frac{K\sigma}{\sqrt{h}} + 3\left\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\right\|_{\infty,\mathrm{B}(0,K)} + C_P k^{-\frac{2}{d'}}.$$

At last, using (2), Lemmas 23 and 24 leads to

$$\begin{aligned}\left\|\mathrm{d}^2_{Q,h} - \mathrm{d}^2_{P,h}\right\|_{\infty,\mathrm{B}(0,K)} &\leq \|\mathrm{d}_{Q,h} - \mathrm{d}_{P,h}\|_{\infty,\mathrm{B}(0,K)}\left(\|\mathrm{d}_{Q,h}\|_{\infty,\mathrm{B}(0,K)} + \|\mathrm{d}_{P,h}\|_{\infty,\mathrm{B}(0,K)}\right)\\ &\leq \frac{W_2(P,Q)}{\sqrt{h}}\left(\sup_{x\in\mathrm{B}(0,K)}\sqrt{\left\|x - m(Q_{x,h})\right\|^2 + v(Q_{x,h})} + 2K\right)\\ &\leq \frac{\sqrt{3}\sigma}{\sqrt{h}}\left(\sqrt{K^2 + 2K\sqrt{3}\frac{\sigma+K}{\sqrt{h}} + 3\frac{(\sigma+K)^2}{h}} + 2K\right)\\ &\leq \frac{\sqrt{3}\sigma(3K + \sqrt{3}(K+\sigma))}{h}.\end{aligned}$$

6.9. Proof of Proposition 21

As in the proof of [7], Theorem 1, for the sake of simplicity we assume that k is divisible by 3, set $m = (2k)/3$, and let $z_1, \ldots, z_m$ be a 6Δ-net in $\mathrm{B}(0,K)$, with $\Delta = K/(6m^{\frac{1}{d}})$, so that such a net exists. We let as well $w_1, \ldots, w_m$ be in $\mathbb{R}^d$ such that $\|w_i\| = \Delta$ and $z_i + w_i \in \mathrm{B}(0,K)$. For $\sigma \in \{-1,+1\}^m$ such that $\sum_{i=1}^m \sigma_i = 0$ we denote by P_σ the distribution that satisfies, for $i \in [\![1,m]\!]$,

$$P_\sigma\left(\{z_i\}\right) = P_\sigma\left(\{z_i + w_i\}\right) = \frac{(1+\sigma_i\delta)}{2m},$$

with $\delta \leq \frac{1}{3}$. For $\tau \in \{-1,1\}^{\frac{m}{2}}$, $\sigma(\tau)$ is defined by $\sigma(\tau)_j = \tau_j$ and $\sigma(\tau)_{\frac{m}{2}+j} = -\tau_j$, for $j \in [\![1, m/2]\!]$. We define now a p-points quantizer F as a map from $\mathbb{R}^d$ such that $|F(\mathbb{R}^d)| = p$, and define F_σ as the k-points quantizer satisfying

$$\begin{aligned}&F_\sigma(z_i) = z_i, \qquad F_\sigma(z_i + w_i) = z_i + w_i \quad \text{if } \sigma_i = +1,\\ &F_\sigma(z_i) = F_\sigma(z_i + w_i) = z_i \quad \text{if } \sigma_i = -1.\end{aligned}$$

At last, for a quantizer F with images $q_1, \ldots, q_p$ and sets of preimages $V_1, \ldots, V_p$, we denote by $R(F, P_\sigma)$ the quantity

$$R(F,P_\sigma) = \sum_{i=1}^p P_\sigma\left[\left\|\cdot - m(P_{q_i,h})\right\|^2 + v(P_{q_i,h})\right]\mathbb{1}_{V_i},$$

where for $i \in [\![1, p]\!]$, $P_{q_i,h} \in \mathcal{P}_{q_i,h}(P_\sigma)$. With a slight abuse we call nearest-neighbor quantizer a quantizer whose sets of preimages are the set of Voronoi cells associated with $(m(P_{q_i,h}), v(P_{q_i,h}))$, with ties arbitrarily broken.

The proof of Proposition 21 follows from the same arguments as [33], Proposition 3.1. We first use the following lemma.

Lemma 27. *Assume that $\delta \le \frac{1}{3}$ and $h \le \frac{1}{3m}$. Let σ and σ' be such that $\sum_{i=1}^m \sigma_i = \sum_{i=1}^m \sigma'_i = 0$, and let $\rho(\sigma, \sigma')$ denote the distance $\sum_{i=1}^m |\sigma_i - \sigma'_i|$. Then*

$$R\big(F_\sigma, P'_\sigma\big) = R(F_\sigma, P_\sigma) + \frac{\delta\Delta^2}{2m}\rho\big(\sigma, \sigma'\big).$$

Moreover, for every k-points nearest neighbor quantizer F there exists σ and τ such that

$$\forall P_{\sigma(\tau')} R(F, P_{\sigma(\tau')}) \ge R(F_\sigma, P_{\sigma(\tau')}) \ge \frac{1}{2} R(F_{\sigma(\tau)}, P_{\sigma(\tau')}).$$

The proof of Lemma 27 is a slight modification of that of [33], Proposition 4.2. For the sake of completeness it is given in Section C.1 of the Appendix. Let $\hat{\mathbf{t}}$ be an empirically designed vector in $(\mathbb{R}^d)^{(k)}$, and recall that $P\mathrm{d}^2_{P,h,\mathbf{t}}$ is defined as $P\sum_{j=1}^k [\|\cdot - m(P_{t_i,h})\|^2 + v(P_{t_i,h})]\mathbb{1}_{V_i}$. According to Lemma 27, we may write

$$\begin{aligned}
\inf_{\hat{\mathbf{t}}} \sup_{P\,|\,\mathrm{Supp}(P)\subset \mathrm{B}(0,K)} \mathbb{E}P\big(\mathrm{d}^2_{P,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\big) &\ge \inf_{\hat{\mathbf{t}}} \sup_{P_{\sigma(\tau')}} \mathbb{E}R(\hat{\mathbf{t}}) - R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) \\
&\ge \frac{1}{2} \inf_{\hat{\tau}} \sup_{P_{\sigma(\tau')}} \mathbb{E}R(F_{\sigma(\hat{\tau})}, P_{\sigma(\tau')}) - R(F_{\sigma(\tau')}, P_{\sigma(\tau')}) \\
&\ge \frac{\delta\Delta^2}{2m} \inf_{\hat{\tau}} \sup_{P_{\sigma(\tau')}} \mathbb{E}\rho\big(\hat{\tau}, \tau'\big), \qquad (17)
\end{aligned}$$

where $\hat{\tau}$ denotes an empirically designed element of $\{-1, +1\}^{\frac{m}{2}}$. Let μ denote the measure $\sum_{i=1}^m (\delta_{z_i} + \delta_{z_i+w_i})$. For any distribution P and Q having densities with respect to μ we denote by $H^2(P, Q)$ their Hellinger distance.

Lemma 28. *Let τ and τ' in $\{-1, 1\}^{\frac{m}{2}}$ such that $\rho(\tau, \tau') = 2$. Then*

$$H^2\big(P^{\otimes n}_{\sigma(\tau)}, P^{\otimes n}_{\sigma(\tau')}\big) \le \frac{4n\delta^2}{m} := \alpha.$$

The proof of Lemma 28 is a slight modification of the proof of [33], Lemma 4.5, and is given in Section C.2 in the Appendix. A direct application of Assouad's Lemma (see, e.g., [42], Theorem 2.12) entails that, for $\alpha \le 2$,

$$\inf_{\hat{\tau}} \sup_{\tau\in\{-1,1\}^{\frac{m}{2}}} \mathbb{E}\rho(\hat{\tau}, \tau) \ge \frac{m}{4}\big(1 - \sqrt{\alpha(1 - \alpha/4)}\big).$$

For $\delta = \frac{\sqrt{m}}{2\sqrt{n}}$, (17) yields

$$\inf_{\hat{\mathbf{t}}} \sup_{P|\,\mathrm{Supp}(P)\subset \mathrm{B}(0,K)} \mathbb{E} P\big(\mathrm{d}^2_{P,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\big) \geq c_0 \frac{K^2 k^{\frac{1}{2}-\frac{2}{d}}}{\sqrt{n}}.$$

This proves (8).

Now denote by A the event $\bigcup_{i=1}^m \{P_n(\{z_i\}) \leq h\} \cup \{P_n(\{z_i + w_i\}) \leq h\}$. If $\delta \leq \frac{1}{9}$ and $h \leq \frac{1}{2k}$, using $\frac{(1-\delta)}{2m} - h \geq \frac{1}{6k}$ and a union of bounded difference inequalities (see, e.g., [10], Theorem 6.2) leads to $\mathbb{P}_\sigma(A) \leq 2me^{-\frac{n}{72k^2}}$. Next, on the event A^c, we have, for any σ, $t \in \mathbb{R}^d$ and $P_{nt,h} \in \mathcal{P}_{h,t}(P_n)$, $P_{nt,h} \in \mathcal{P}_{t,h}(P)$. Thus, for $\mathbf{t} \in (\mathbb{R}^d)^{(k)}$, $\mathrm{d}^2_{P_n,h,\mathbf{t}} \mathbb{1}_{A^c} = (\min_{j=1,\dots,k} \|. - m(P_{t_j,h})\|^2 + v(P_{t_j,h}))\mathbb{1}_{A^c}$, where $P_{t_j,h} \in \mathcal{P}_{t_j,h}$. Therefore, since for every σ, $\mathrm{Supp}(P_\sigma) \subset \mathrm{B}(0,K)$, we may write

$$\begin{aligned}
&\inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P\big(\mathrm{d}^2_{P_n,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\big) \\
&\quad \geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P\big(\mathrm{d}^2_{P_n,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\big)\mathbb{1}_{A^c} - 16K^2me^{-\frac{n}{72k^2}} \\
&\quad \geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P\Big(\min_{j=1,\dots,k} \|. - m(P_{\hat{t}_j,h})\|^2 + v(P_{\hat{t}_j,h}) - \mathrm{d}^2_{P,h,k}\Big)\mathbb{1}_{A^c} - 16K^2me^{-\frac{n}{72k^2}} \\
&\quad \geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P\Big(\min_{j=1,\dots,k} \|. - m(P_{\hat{t}_j,h})\|^2 + v(P_{\hat{t}_j,h}) - \mathrm{d}^2_{P,h,k}\Big) - 32K^2me^{-\frac{n}{72k^2}} \\
&\quad \geq \inf_{\hat{\mathbf{t}}} \sup_{\sigma} \mathbb{E}_\sigma P\big(\mathrm{d}^2_{P,h,\hat{\mathbf{t}}} - \mathrm{d}^2_{P,h,k}\big) - 32K^2me^{-\frac{n}{72k^2}}.
\end{aligned}$$

Since $n \geq 14k$, $\delta = \frac{\sqrt{m}}{2\sqrt{n}} \leq \frac{1}{9}$ and (17) leads to the result again.

Acknowledgements

This work was partially supported by the ANR project TopData, GUDHI, the INRIA team DataShape and the Laboratoire de Mathématiques d'Orsay. The authors are grateful to Marc Glisse for his help to install and use the Gudhi C++ library.

Supplementary Material

Appendix: Additional figures and proofs of technical results (DOI: 10.3150/20-BEJ1214 SUPP; .pdf). Due to space constraints, we relegate technical details as well as additional figures pertaining to Section 4 to the supplement [12].

References

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

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

[3] Aaron, C. and Cholaquidis, A. (2019). On boundary detection. arXiv E-prints.

[4] Aliprantis, C.D. and Border, K.C. (2006). *Infinite Dimensional Analysis*: *A Hitchhiker's Guide*, 3rd ed. Berlin: Springer. MR2378491

[5] Banerjee, A., Guo, X. and Wang, H. (2005). On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. Inf. Theory* **51** 2664–2669. MR2246384 https://doi.org/10.1109/TIT.2005.850145

[6] Banerjee, A., Merugu, S., Dhillon, I.S. and Ghosh, J. (2005). Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6** 1705–1749. MR2249870

[7] Bartlett, P.L., Linder, T. and Lugosi, G. (1998). The minimax distortion redundancy in empirical quantizer design. *IEEE Trans. Inf. Theory* **44** 1802–1813. MR1664098 https://doi.org/10.1109/18.705560

[8] Biau, G., Devroye, L. and Lugosi, G. (2008). On the performance of clustering in Hilbert spaces. *IEEE Trans. Inf. Theory* **54** 781–790. MR2444554 https://doi.org/10.1109/TIT.2007.913516

[9] Boissonnat, J.-D., Chazal, F. and Yvinec, M. (2018). *Geometric and Topological Inference*. *Cambridge Texts in Applied Mathematics*. Cambridge: Cambridge Univ. Press. MR3837127 https://doi.org/10.1017/9781108297806

[10] Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities*: *A Nonasymptotic Theory of Independence*. Oxford: Oxford Univ. Press. MR3185193 https://doi.org/10.1093/acprof:oso/9780199535255.001.0001

[11] Brécheteau, C., Fischer, A. and Levrard, C. (2018). Robust Bregman clustering. arXiv E-prints.

[12] Brécheteau, C. and Levrard, C. (2020). Supplement to "A k-points-based distance for robust geometric inference." https://doi.org/10.3150/20-BEJ1214SUPP

[13] Buchet, M., Chazal, F., Oudot, S.Y. and Sheehy, D.R. (2015). Efficient and robust persistent homology for measures. In *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* 168–180. Philadelphia, PA: SIAM. MR3451037 https://doi.org/10.1137/1.9781611973730.13

[14] Buchet, M., Dey, T.K., Wang, J. and Wang, Y. (2018). Declutter and resample: Towards parameter free denoising. *J. Comput. Geom.* **9** 21–46. MR3866406

[15] Cardot, H., Cénac, P. and Zitt, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43. MR3019484 https://doi.org/10.3150/11-BEJ390

[16] Chazal, F., Cohen-Steiner, D. and Mérigot, Q. (2011). Geometric inference for probability measures. *Found. Comput. Math.* **11** 733–751. MR2859954 https://doi.org/10.1007/s10208-011-9098-0

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

[18] Chazal, F., Massart, P. and Michel, B. (2016). Rates of convergence for robust geometric inference. *Electron. J. Stat.* **10** 2243–2286. MR3541971 https://doi.org/10.1214/16-EJS1161

[19] Cohen-Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete Comput. Geom.* **37** 103–120. MR2279866 https://doi.org/10.1007/s00454-006-1276-5

[20] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997). Trimmed k-means: An attempt to robustify quantizers. *Ann. Statist.* **25** 553–576. MR1439314 https://doi.org/10.1214/aos/1031833664

[21] Dijksman, J.A., Kovalcinova, L., Ren, J., Behringer, R.P., Kramar, M., Mischaikow, K. and Kondic, L. (2018). Characterizing granular networks using topological metrics. *Phys. Rev. E* **97** 042903.

[22] Edelsbrunner, H. (1992). Weighted alpha shapes. Technical report, Champaign, IL, USA.
[23] Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002). Topological persistence and simplification **28** 511–533. MR1949898 https://doi.org/10.1007/s00454-002-2885-2
[24] Eldar, Y., Lindenbaum, M., Porat, M. and Zeevi, Y.Y. (1997). The farthest point strategy for progressive image sampling. *IEEE Trans. Image Process.* **6** 1305–1315. https://doi.org/10.1109/83.623193
[25] Fasy, B.T., Kim, J., Lecci, F. and Maria, C. (2014). Introduction to the R package TDA.
[26] Federer, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93** 418–491. MR0110078 https://doi.org/10.2307/1993504
[27] Fritz, H., Garcia-Escudero, L.A. and Mayo-Iscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *J. Stat. Softw.* **47** 1–26.
[28] Genovese, C.R., Perone-Pacifico, M., Verdinelli, I. and Wasserman, L. (2012). Manifold estimation and singular deconvolution under Hausdorff loss. *Ann. Statist.* **40** 941–963. MR2985939 https://doi.org/10.1214/12-AOS994
[29] Guibas, L., Morozov, D. and Mérigot, Q. (2013). Witnessed k-distance. *Discrete Comput. Geom.* **49** 22–45. MR3010216 https://doi.org/10.1007/s00454-012-9465-x
[30] Kanari, L., Dłotko, P., Scolamiero, M., Levi, R., Shillcock, J., Hess, K. and Markram, H. (2018). A topological representation of branching neuronal morphologies. *Neuroinformatics* **16** 3–13.
[31] Kim, A.K.H. and Zhou, H.H. (2015). Tight minimax rates for manifold estimation under Hausdorff loss. *Electron. J. Stat.* **9** 1562–1582. MR3376117 https://doi.org/10.1214/15-EJS1039
[32] Komiya, H. (1988). Elementary proof for Sion's minimax theorem. *Kodai Math. J.* **11** 5–7. MR0930413 https://doi.org/10.2996/kmj/1138038812
[33] Levrard, C. (2015). Nonasymptotic bounds for vector quantization in Hilbert spaces. *Ann. Statist.* **43** 592–619. MR3316191 https://doi.org/10.1214/14-AOS1293
[34] Lloyd, S.P. (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** 129–137. MR0651807 https://doi.org/10.1109/TIT.1982.1056489
[35] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (*Berkeley, Calif.*, 1965/66) 281–297. Berkeley, CA: Univ. California Press. MR0214227
[36] Maggioni, M., Minsker, S. and Strawn, N. (2016). Multiscale dictionary learning: Non-asymptotic bounds and robustness. *J. Mach. Learn. Res.* **17** Paper No. 2, 51. MR3482922
[37] McLachlan, G. and Peel, D. (2000). *Finite Mixture Models. Wiley Series in Probability and Statistics*: *Applied Probability and Statistics*. New York: Wiley Interscience. MR1789474 https://doi.org/10.1002/0471721182
[38] Mérigot, Q. (2013). Lower bounds for k-distance approximation. In *Computational Geometry* (*SoCG'*13) 435–440. New York: ACM. MR3208242 https://doi.org/10.1145/2462356.2462367
[39] Niyogi, P., Smale, S. and Weinberger, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39** 419–441. MR2383768 https://doi.org/10.1007/s00454-008-9053-2
[40] Phillips, J.M., Wang, B. and Zheng, Y. (2015). Geometric inference on kernel density estimates. In 31*st International Symposium on Computational Geometry. LIPIcs. Leibniz Int. Proc. Inform.* **34** 857–871. Wadern: Schloss Dagstuhl. Leibniz-Zent. Inform. MR3392827
[41] Rodríguez Casal, A. (2007). Set estimation under convexity type assumptions. *Ann. Inst. Henri Poincaré Probab. Stat.* **43** 763–774. MR3252430 https://doi.org/10.1016/j.anihpb.2006.11.001
[42] Tsybakov, A.B. (2009). *Introduction to Nonparametric Estimation. Springer Series in Statistics*. New York: Springer. MR2724359 https://doi.org/10.1007/b13794
[43] Zomorodian, A. and Carlsson, G. (2005). Computing persistent homology. *Discrete Comput. Geom.* **33** 249–274. MR2121296 https://doi.org/10.1007/s00454-004-1146-y

Received August 2019 and revised March 2020