Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari, Clément Levrard

Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari and Clément Levrard

Abstract

In this paper we consider the problem of optimality in manifold reconstruction. A random sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ composed of points lying on a d-dimensional submanifold M, with or without outliers drawn in the ambient space, is observed. Based on the tangential Delaunay complex [3], we construct an estimator $\hat{M}$ that is ambient isotopic and Hausdorff-close to M with high probability. $\hat{M}$ is built from existing algorithms. In a model without outliers, we show that this estimator is asymptotically minimax optimal for the Hausdorff distance over a class of submanifolds with reach condition. Therefore, even with no *a priori* information on the tangent spaces of M, our estimator based on tangential Delaunay complexes is optimal. This shows that the optimal rate of convergence can be achieved through existing algorithms. A similar result is also derived in a model with outliers. A geometric interpolation result is derived, showing that the tangential Delaunay complex is stable with respect to noise and perturbations of the tangent spaces. In the process, a denoising procedure and a tangent space estimator both based on local principal component analysis (PCA) are studied.

1 Introduction

In various settings coming from experimental science, it is usual to manipulate point samples that can be modelled as lying on a d-dimensional submanifold $M \subset \mathbb{R}^D$. As M may carry a lot of information about the studied phenomenon, it is then natural to consider the problem of either approximating M geometrically, recovering it topologically, or both from a point sample $\mathbb{X}_n = \{X_1, \dots, X_n\}$. It is of particular interest in high codimension ($d \ll D$) where it can be used as a preliminary processing of the data for reducing its dimension, and then avoiding the curse of dimensionality. This problem is usually referred to as *manifold reconstruction* in the computational geometry community, and rather called *set/support estimation* or *manifold learning* in the statistics literature.

The computational geometry community has now been active on manifold reconstruction for many years, mainly in deterministic frameworks. In dimension 3, [12] provides a survey of the state of the art. In higher dimension, the employed methods rely on variants of the ambient Delaunay triangulation [8, 3]. The geometric and topological guarantees are derived under the assumption that the point cloud - fixed and non random - densely samples M at scale ε, with ε small enough or going to 0.

In the statistics literature, most of the attention has been paid to approximation guarantees, rather than topological ones. The approximation bounds are given in terms of the sample size n, that is assumed to be large enough or in the asymptotic $n \to \infty$. To derive these bounds, a broad scope of assumptions on M have been considered. For instance, if M is a bounded convex set and $\mathbb{X}_n$ does not contain outliers, a natural idea is to consider the convex hull $\hat{M} = Conv(\mathbb{X}_n)$ to be the

estimator. $Conv(\mathbb{X}_n)$ provides optimal rates of approximation for several loss functions [22, 14]. These rates depend crudely on the regularity of the boundary of the convex M. In addition, $Conv(\mathbb{X}_n)$ is clearly ambient isotopic to M so that it has both good geometric and topological properties. Generalisations of the notion of convexity based on rolling ball-type assumptions such as r-convexity and reach conditions [10, 18] yield rich classes of sets with good geometric properties. In particular, the reach, as introduced by H. Federer [16], has been proved to be a key regularity and scale parameter [7, 18].

This paper mainly follows up the two articles [3, 18], both dealing with the case of a d-dimensional manifold $M \subset \mathbb{R}^D$ under reach condition and where the dimension d is known.
On one hand, [3] focuses on a deterministic analysis and proposes a provably faithful reconstruction. These authors introduce a weighted Delaunay triangulation restricted to tangent spaces, the so-called Tangential Delaunay Complex. They give ambient isotopy guarantee and approximation bounds for the Hausdorff distance along with computational complexity bounds. This work provides a simplicial complex based on the input point cloud and tangent spaces. However, it lacks stability up to now, in the sense that the assumptions used in the proofs of [3] do not resist ambient perturbations. Indeed, it intensively uses the knowledge of the tangent spaces at each point and the absence of noise.
On the other hand, [18] takes a statistical approach in a model containing possibly outlier points. The authors derive an estimator that is proved to be minimax optimal for the Hausdorff loss $\mathrm{d_H}$. Although theoretically optimal, the proposed estimator appears to be intractable in practice.

Our main contribution (Theorem 6) makes a two-way link between the approaches of [3] and [18]. It shows that with an additional tangent space estimation procedure that we propose, the tangential complex of [3] achieves the optimal rate of convergence in the model without outliers of [18]. Conversely, Theorem 6 points out that the rate of [18] can be achieved with a tractable estimator $\hat{M}$ that is a simplicial complex of which vertices are the data points and such that $\hat{M}$ is ambient isotopic to M with high probability. In the presence of outliers, a similar result (Theorem 7) is also provided. Moreover, a stability result for the tangential Delaunay complex (Theorem 12) is proved.

Outline

This paper deals with the case where a sample $\mathbb{X}_n = \{X_1, \dots, X_n\} \subset \mathbb{R}^D$ of size n is randomly drawn on/around M. First, the statistical framework is described (Section 2). Two models are studied, one where $\mathbb{X}_n \subset M$ and the other where $\mathbb{X}_n$ contains outliers. We build a simplicial complex $\hat{M}_{\mathtt{TDC}}(\mathbb{X}_n)$ ambient isotopic to M and achieving a rate of approximation for the Hausdorff loss $\mathrm{d_H}(M, \hat{M}_{\mathtt{TDC}})$, with bounds holding uniformly over a class of submanifolds satisfying a reach condition. The derived rate of convergence is minimax optimal (Section 2) in the model without outliers. With outliers, a similar estimator $\hat{M}_{\mathtt{TDC+}}$ is built. $\hat{M}_{\mathtt{TDC}}$ and $\hat{M}_{\mathtt{TDC+}}$ are based on the tangential Delaunay complex (Section 3), that is first proved to be stable (Section 4) *via* an interpolation result. For this purpose, a method to estimate tangent spaces and to remove outliers based on local Principal Component Analysis (PCA) is proposed (Section 5). We conclude with general remarks and possible extensions (Section 6).

Notation

In what follows, we consider a smooth d-dimensional submanifold without boundary $M \subset \mathbb{R}^D$ to be reconstructed. For all $p \in M$, T_pM designates the tangent space of M at p. Tangent spaces will either be considered vectorial or affine depending on the context. The standard inner product in $\mathbb{R}^D$ is denoted by $\langle \cdot, \cdot \rangle$ and the Euclidean distance $\|\cdot\|$. We let $\mathcal{B}(p, r)$ denote the closed Euclidean ball of radius $r > 0$ centered at p. As introduced in [16], the reach of M, denoted by $\mathrm{reach}(M)$ is the maximal offset radius for which the projection π_M onto M is well defined. Denoting by $\mathrm{d}(\cdot, M)$ the distance to M, the *medial axis* of M $\mathrm{med}(M) = \{x \in \mathbb{R}^D | \exists a \neq b \in M, \|x-a\| = \|x-b\| = \mathrm{d}(x, M)\}$ is the set of points which have at least two nearest neighbors on M. Then, $\mathrm{reach}(M) = \inf_{p \in M} \mathrm{d}(p, \mathrm{med}(M))$. We simply write π for π_M when there is no possibility of confusion. For any smooth function $\Phi : \mathbb{R}^D \to \mathbb{R}^D$, we let $\mathrm{d}_a\Phi$ and $\mathrm{d}^2_a\Phi$ denote the first and second order differentials of Φ at $a \in \mathbb{R}^D$. For a linear map A, A^t designates its transpose. Let $\|A\|_{\mathrm{op}} = \sup_x \frac{\|Ax\|}{\|x\|}$ and $\|A\|_{\mathcal{F}} = \sqrt{\mathrm{trace}\,(A^tA)}$ denote respectively the operator norm induced by the Euclidean norm and the Frobenius norm. The distance between two linear subspaces $U, V \subset \mathbb{R}^D$ of same dimension is measured by the principal angle $\angle(U, V) = \max_{u \in U} \max_{v' \in V^\perp} \frac{\langle u, v' \rangle}{\|u\| \, \|v'\|} = \|\pi_U - \pi_V\|_{\mathrm{op}}$. The Hausdorff distance in $\mathbb{R}^D$ is denoted by d_{H}. We let $\cong$ denote the ambient isotopy relation in $\mathbb{R}^D$.
Throughout this paper, C_α will denote a generic constant depending on the parameter α. For clarity sakeness, c_α and K_α may also be used when several constants are involved.

2 Minimax Risk and Optimality

2.1 Statistical Model

Let us describe the general statistical setting we will use to define optimality for manifold reconstruction. A *statistical model* $\mathcal{D}$ is a set of probability distributions on $\mathbb{R}^D$. In any statistical experiment, $\mathcal{D}$ is fixed and known. We observe an independent and identically distributed sample of size n (or i.i.d. n-sample) $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ drawn according to some unknown distribution $P \in \mathcal{D}$. If no noise is allowed, the problem is to recover the *support* of P, that is, the smallest closed set $C \subset \mathbb{R}^D$ such that $P(C) = 1$. Let us give two examples of such models $\mathcal{D}$ by describing those of interest in this paper.

Let $\mathcal{M}_{D,d,\rho}$ be the set of all the d-dimensional connected submanifolds $M \subset \mathbb{R}^D$ without boundary satisfying $\mathrm{reach}(M) \geq \rho$. The reach assumption is crucial to avoid arbitrarily curved and pinched shapes [10]. From a reconstruction point of view, ρ gives a minimal feature size on M, and then a minimal scale for geometric information. Every $M \in \mathcal{M}_{D,d,\rho}$ inherits a measure induced by the d-dimensional Hausdorff measure $\mathrm{d}\mathcal{H}_d(x)$ on $\mathbb{R}^D \supset M$. We denote this induced measure $\mathrm{d}v_M(x)$. Beyond the geometric restrictions induced by the lower bound ρ on the reach, it also requires the natural measure v_M to behave like a d-dimensional measure, up to uniform constants. Namely, v_M satisfies the (a, d)-standard property of [6], with $a = a_{d,\rho}$. Denote by $\mathcal{U}_M(f_{min}, f_{max})$ the set of probability distributions Q having a density f with respect to v_M such that $0 < f_{min} \leq f(x) \leq f_{max} < \infty$ for all $x \in M$. Roughly speaking, when $Q \in \mathcal{U}_M(f_{min}, f_{max})$, points are drawn almost uniformly on M since f is bounded away from 0 and ∞. This is to ensure that the sample visits all the areas of M with high probability. The model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$ consists of the set of all these almost uniform measures on submanifolds of

dimension d having reach greater than a fixed value $\rho > 0$.

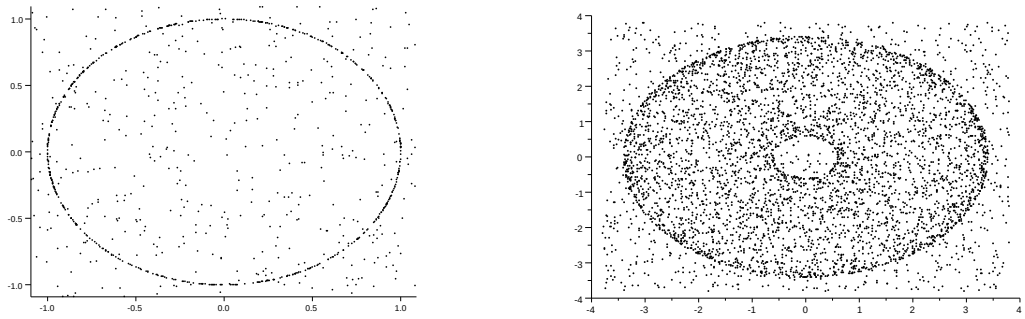
Definition 1 (Model without outliers)**.** $\mathcal{G}_{D,d,f_{min},f_{max},\rho} = \bigcup_{M \in \mathcal{M}_{D,d,\rho}} \mathcal{U}_M(f_{min}, f_{max})$.

We do not explicitly impose a bound on the diameter of M. Actually, a bound is implicitly present in the model, as stated in the next lemma, the proof of which follows from a volume argument.

Lemma 2. *There exists $C_d > 0$ such that for all $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$ with associated M,*

$$\operatorname{diam}(M) \leq \frac{C_d}{\rho^{d-1} f_{min}} =: K_{d,f_{min},\rho}.$$

Eventually, we may include distributions contaminated with outliers uniformly drawn in a ball $\mathcal{B}_0$ containing M, as illustrated in Figure 1. Up to translation, we can always assume that $M \ni 0$. To avoid boundary effects, we will take $\mathcal{B}_0$ to contain widely M so that the outlier distribution surrounds M everywhere. Since M has at most diameter $K_{d,f_{min},\rho}$ from Lemma 2 we arbitrarily fix $\mathcal{B}_0 = \mathcal{B}(0, K_0)$, where $K_0 = K_{d,f_{min},\rho} + \rho$. Notice that the larger the radius of $\mathcal{B}_0$, the easier to label the outlier points since they should be very far away from each other.



(a) Circle with outliers: $d = 1$,$D = 2$.

(b) Torus with outliers: $d = 2$,$D = 3$.

Figure 1: Point clouds $\mathbb{X}_n$ drawn from distributions of two instances of models $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ with different parameters and $\beta < 1$.

Definition 3 (Model with outliers/Clutter noise model)**.** *For $0 < f_{min} \leq f_{max} < \infty$, $0 < \beta \leq 1$, and $\rho > 0$, we define $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ to be the set of mixture distributions*

$$P = \beta Q + (1 - \beta) U_{\mathcal{B}_0},$$

where $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$ has support M such that $0 \in M$, and $U_{\mathcal{B}_0}$ is the uniform distribution on $\mathcal{B}_0 = \mathcal{B}(0, K_0)$.

Alternatively, a random variable X with distribution $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ can be represented as $X = VX' + (1 - V)X''$, where $V \in \{0, 1\}$ is a Bernoulli random variable with parameter β, X' has distribution in $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$ and X'' has a uniform distribution over $\mathcal{B}_0$, and such that V, X', X'' are independent. In particular for $\beta = 1$, $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta=1} = \mathcal{G}_{D,d,f_{min},f_{max},\rho}$.

2.2 Minimax Risk

For a probability measure $P \in \mathcal{D}$, denote by $\mathbb{E}_P$ - or simply $\mathbb{E}$ - the expectation with respect to the product measure $P^{(n)}$. The quantity we will be interested in is the *minimax risk* associated to the model $\mathcal{D}$. For $n \geq 0$,

$$R_n(\mathcal{D}) = \inf_{\hat{M}} \sup_{P \in \mathcal{D}} \mathbb{E}_P \left[\mathrm{d_H} \left(M, \hat{M} \right) \right],$$

where the *infimum* is taken over all the estimators $\hat{M} = \hat{M}(X_1, \ldots, X_n)$ computed over a n-sample. $R_n(\mathcal{D})$ is the best risk that an estimator based on a n sample can achieve uniformly over the class $\mathcal{D}$. It is clear from the definition that if $\mathcal{D}' \subset \mathcal{D}$ then $R_n(\mathcal{D}') \leq R_n(\mathcal{D})$. It follows the intuition that the broader the class of considered manifolds, the more difficult it is to estimate them uniformly well. Studying $R_n(\mathcal{D})$ for a fixed n is a difficult task that can rarely be carried out. We will focus on the semi-asymptotic behaviour of this risk. For this purpose, we will say that two sequences $(a_n)_n$ and $(b_n)_n$ are asymptotically comparable, denoted by $a_n \asymp b_n$, if there exist $c, C > 0$ such that for n large enough, $cb_n \leq a_n \leq Cb_n$.

Definition 4. *An estimator $\hat{M}$ is said to be (asymptotically) minimax optimal over $\mathcal{D}$ if*

$$\sup_{P \in \mathcal{D}} \mathbb{E}_P \left[\mathrm{d_H} \left(M, \hat{M} \right) \right] \asymp R_n(\mathcal{D}).$$

In other words, $\hat{M}$ is (asymptotically) minimax optimal if it achieves, up to constants, the best possible rate of convergence in the worst case.

Studying a minimax rate of convergence is twofold. On one hand, deriving an upper bound on R_n boils down to propose an estimator and to study its quality uniformly on $\mathcal{D}$. On the other hand, lower bounding R_n amounts to analyse the worst possible case in $\mathcal{D}$. This part is usually achieved with standard Bayesian techniques [21]. For the models considered in the present paper, the rates were given in [18, 20].

Theorem 5 (Theorem 3 of [20])**.** $R_n\left(\mathcal{G}_{D,d,f_{min},f_{max},\rho}\right) \asymp R_n\left(\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}\right) \asymp \left(\frac{\log n}{n}\right)^{2/d}$.

Beyond this theoretical result, an interesting question is to know whether this minimax rate can be achieved by a tractable algorithm. Indeed, that proposed in [18] especially relies on a minimization problem over the class of submanifolds $\mathcal{M}_{D,d,\rho}$, which seems computationally costly. In addition, the proposed estimator is itself a manifold, that raises storage problems. Moreover, no guarantee is given on the topology of the estimator.

Throughout the present paper, we will build an estimator $\hat{M}_{\texttt{TDC}}$ that addresses these issues. $\hat{M}_{\texttt{TDC}}$ is based on the tangential Delaunay complex (Section 3), with a tangent space estimation using a local PCA (Section 5). The result, derived in the model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$, is stated as follows.

Theorem 6. *$\hat{M}_{\texttt{TDC}} = \hat{M}_{\texttt{TDC}}(\mathbb{X}_n)$ is a simplicial complex of vertices $\mathbb{X}_n$ such that*

$$\lim_{n \to \infty} \mathbb{P} \left(\mathrm{d_H}(M, \hat{M}_{\texttt{TDC}}) \leq C_{d,f_{min},f_{max},\rho} \left(\frac{\log n}{n} \right)^{2/d} \text{ and } M \cong \hat{M}_{\texttt{TDC}} \right) = 1.$$

Moreover, $\hat{M}_{\texttt{TDC}}$ is minimax optimal over $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$: for n large enough,

$$\sup_{Q\in\mathcal{G}_{D,d,f_{min},f_{max},\rho}} \mathbb{E}_Q \mathrm{d_H}(M,\hat{M}_{\texttt{TDC}}) \leq C'_{d,f_{min},f_{max},\rho}\left(\frac{\log n}{n}\right)^{2/d}.$$

It is interesting to note that the constants appearing in Theorem 6 do not depend on the ambient dimension D.

Furthermore, in the model with outliers $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, with the same procedure used to derive Theorem 6 and an additional preprocessing of the data based on local PCA to remove outliers (Section 5), we get an estimator $\hat{M}_{\texttt{TDC+}}$ satisfying the following similar statement.

Theorem 7. *$\hat{M}_{\texttt{TDC+}} = \hat{M}_{\texttt{TDC+}}(\mathbb{X}_n)$ is a simplicial complex of vertices contained in $\mathbb{X}_n$ such that*

$$\lim_{n\to\infty}\mathbb{P}\left(\mathrm{d_H}(M,\hat{M}_{\texttt{TDC+}}) \leq C_{d,f_{min},f_{max},\rho}\left(\frac{\log n}{\beta n}\right)^{2/(d+1)} \text{ and } M\cong\hat{M}_{\texttt{TDC+}}\right)=1.$$

Moreover, for n large enough,

$$\sup_{P\in\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}} \mathbb{E}_Q \mathrm{d_H}(M,\hat{M}_{\texttt{TDC+}}) \leq C'_{d,f_{min},f_{max},\rho}\left(\frac{\log n}{\beta n}\right)^{2/(d+1)}.$$

$\hat{M}_{\texttt{TDC+}}$ converges at the rate at least $\left(\frac{\log n}{n}\right)^{2/(d+1)}$, which is not the minimax optimal rate according to Theorem 5. However, to our knowledge, $\hat{M}_{\texttt{TDC+}}$ is the first explicit estimator to provably achieve such a rate in the presence of outliers. Again, it is worth noting that the constants involved in Theorem 7 do not depend on the ambient dimension D.

The statistical analysis of the problem is postponed to Section 5. Beforehand, let us describe the tangential Delaunay complex in a deterministic and idealized framework where the tangent spaces are known and no outliers are present.

3 Tangential Delaunay Complex

Let $\mathcal{P}=\{p_1,\dots,p_n\}\subset M$. In this section, we denote the point cloud $\mathcal{P}$ to emphasize the fact that it is considered non random. For $\varepsilon,\delta>0$, $\mathcal{P}$ is said to be *ε-dense* in M if $\mathrm{d_H}(M,\mathcal{P})\leq\varepsilon$, and *$\delta$-sparse* if $\mathrm{d}(p,\mathcal{P}\setminus\{p\})\geq\delta$ for all $p\in\mathcal{P}$. A *(δ,ε)-net* (of M) is a δ-sparse and ε-dense point cloud.

3.1 Restricted Weighted Delaunay Triangulations

A weight assignment to $\mathcal{P}$ is a function $\omega:\mathcal{P}\longrightarrow[0,\infty)$. The *weighted Voronoi diagram* is defined to be the Voronoi diagram associated to the weighted distance $\mathrm{d}(x,p^\omega)^2=\|x-p\|^2-\omega(p)^2$. Every $p\in\mathcal{P}$ is associated to its weighted Voronoi cell $\mathrm{Vor}^\omega(p)$. For $\tau\subset\mathcal{P}$, let

$$\mathrm{Vor}^\omega(\tau)=\bigcap_{p\in\tau}\mathrm{Vor}^\omega(p)$$

be the common face of the Voronoi cells of the points of τ. The *weighted Delaunay triangulation* $\mathrm{Del}^\omega(\mathcal{P})$ is the dual triangulation to the decomposition given by the Voronoi diagram. In other words, for $\tau\subset\mathcal{P}$, the simplex with vertices τ, denoted by $[\tau]$, satisfies

$$[\tau]\in\mathrm{Del}^\omega(\mathcal{P})\Leftrightarrow\mathrm{Vor}^\omega(\tau)\neq\emptyset.$$

Note that for a constant weight assignment $\omega(p) = \omega_0$, $\mathrm{Del}^\omega(\mathcal{P})$ is the usual Delaunay triangulation of $\mathcal{P}$. Under mild assumptions on $\mathcal{P}$ and ω, $\mathrm{Del}^\omega(\mathcal{P})$ is an embedded triangulation with vertex set $\mathcal{P}$ [3]. The proposed reconstruction is based on $\mathrm{Del}^\omega(\mathcal{P})$ for some weights ω to be chosen later. As it is a triangulation of the whole convex hull of $\mathcal{P}$ and may fail to recover the geometric structure of M, we take restrictions of it.

Given a family $R = \{R_p\}_{p\in\mathcal{P}}$ of subsets $R_p \subset \mathbb{R}^D$ indexed by $\mathcal{P}$, the weighted Delaunay complex restricted to R is the sub-complex of $\mathrm{Del}^\omega(\mathcal{P})$ defined by

$$[\tau] \in \mathrm{Del}^\omega(\mathcal{P}, R) \Leftrightarrow \mathrm{Vor}^\omega(\tau) \cap \left(\bigcup_{p\in\tau} R_p \right) \neq \emptyset.$$

In particular, we define the *Tangential Delaunay Complex* $\mathrm{Del}^\omega(\mathcal{P}, T)$ by taking $R = T = \{T_pM\}_{p\in\mathcal{P}}$, the family of tangent spaces taken at the points of $\mathcal{P} \subset M$ [3]. $\mathrm{Del}^\omega(\mathcal{P}, T)$ is a pruned version of $\mathrm{Del}^\omega(\mathcal{P})$ where only the simplices with directions close to the tangent spaces are kept. Indeed, T_pM being the best linear approximation of M at p, it is very unlikely for a reconstruction of M to have components in directions normal to T_pM - see Figure 2. As pointed out in [3], computing $\mathrm{Del}^\omega(\mathcal{P}, T)$ only requires to compute Delaunay triangulations in the tangent spaces of dimension d. This reduces the computational complexity dependency on the ambient dimension $D > d$. The

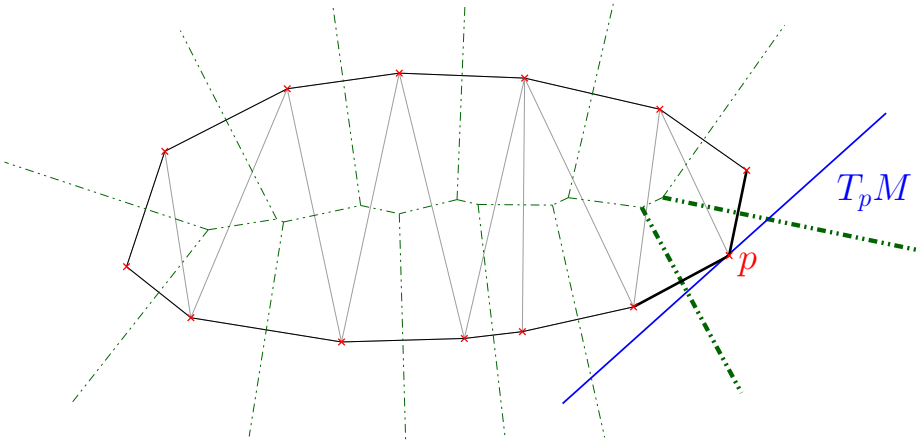


Figure 2: Construction of $\mathrm{Del}^\omega(\mathcal{P}, T)$ at p for $\omega \equiv 0$.

weight assignment ω gives degrees of freedom for the reconstruction. It allows to stabilize the triangulation and enables to remove the so-called *inconsistencies*, letting the points fixed. For further details, see *e.g.* [4, 3].

3.2 Guarantees

The following result sums up the reconstruction properties of the tangential Delaunay complex that we will use. For more details about it, the reader is referred to [3].

Theorem 8 (Adapted from Theorem 5.3 in [3])**.** *There exists ε_0 such that for all $\varepsilon \leq \varepsilon_0$ and all $M \in \mathcal{M}_{D,d,\rho}$, if $\mathcal{P} \subset M$ is an $(\varepsilon, 2\varepsilon)$-net, there exists a weight assignment $\omega_* = \omega_{*\mathcal{P},T}$, depending on $\mathcal{P}$ and $T = \{T_pM\}_{p\in\mathcal{P}}$ such that*

- $\mathrm{d_H}\left(M, \mathrm{Del}^{\omega_*}(\mathcal{P}, T)\right) \leq C_{d,\rho}\varepsilon^2$,
- *M and $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ are ambient isotopic.*

Moreover, a greedy algorithm to compute ω_* is given in [3].

Given an $(\varepsilon, 2\varepsilon)$-net $\mathcal{P}$ for ε small enough, $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ recovers M up to ambient isotopy and approximates it at the scale ε^2. The order of magnitude ε^2 with an input $\mathcal{P}$ of scale ε is remarkable. Another instance of this phenomenon is present in [9] in codimension 1. We will show that this ε^2 provides the minimax rate of approximation when dealing with random samples. Therefore, it can be thought of as optimal.

Theorem 8 suffers two major imperfections. First, it requires the knowledge of the tangent spaces at each sample point - since $\omega_* = \omega_{*\mathcal{P},T}$ - and no guarantee remains if only approximate tangent spaces are known. Second, the points are assumed to lie exactly on the manifold M, and no noise is allowed. The analysis of $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ is sophisticated [3]. Rather than redo the whole study with milder assumptions, we tackle this question with an approximation theory approach (Theorem 9). Instead of studying if $\mathrm{Del}^{\omega_*}(\mathcal{P}', T')$ is stable when $\mathcal{P}'$ lies close to M and T' close to T, we examine what $\mathrm{Del}^{\omega_*}(\mathcal{P}', T')$ actually reconstructs, as detailed in Section 4.

3.3 On the Sparsity Assumption

In Theorem 8, $\mathcal{P}$ is assumed to be dense enough so that it covers all the areas of M. It is also supposed to be sparse at the same scale as the density parameter ε. Indeed, arbitrarily accumulated points would generate non-uniformity and instability for $\mathrm{Del}^{\omega_*}(\mathcal{P}, T)$ [4, 3]. At this stage, an important remark consists in the fact that the construction of a $(\varepsilon, 2\varepsilon)$-net can be carried out with an ε-dense sample with the following procedure. Given an ε-dense sample $\mathcal{P}$, the *farthest point sampling* algorithm prunes $\mathcal{P}$ and outputs an $(\varepsilon, 2\varepsilon)$-net $\mathcal{Q} \subset \mathcal{P}$ of M as follows. Initialize at $\mathcal{Q} = \{p_1\} \subset \mathcal{P}$, and while $\max_{p \in \mathcal{P}} \mathrm{d}(p, \mathcal{Q}) > \varepsilon$, add to $\mathcal{Q}$ the farthest point to $\mathcal{Q}$ in $\mathcal{P}$, *i.e.* $\mathcal{Q} \leftarrow \mathcal{Q} \cup \{\operatorname{argmax}_{p \in \mathcal{P}} \mathrm{d}(p, \mathcal{Q})\}$. The output $\mathcal{Q}$ is ε-sparse and satisfies $\mathrm{d_H}(\mathcal{P}, \mathcal{Q}) \leq \varepsilon$, so it is a $(\varepsilon, 2\varepsilon)$-net of M. Therefore, up to the multiplicative constant 2, sparsifying $\mathcal{P}$ at scale ε will not deteriorate its density property. Then, we can run the farthest point sampling algorithm to preprocess the data, so that the obtained point cloud is a net.

4 Stability Result

4.1 Interpolation Theorem

As mentioned above, if the data do not lie exactly on M and if we do not have the exact knowledge of the tangent spaces, Theorem 8 does not apply. To bypass this issue, we interpolate the data with another manifold M' satisfying good properties, as stated in the following result.

Theorem 9 (Interpolation)**.** *Let $M \in \mathcal{M}_{D,d,\rho}$. Let $\mathcal{P} = \{p_1, \ldots, p_q\} \subset \mathbb{R}^D$ be a finite point cloud and $\tilde{T} = \left\{\tilde{T}_1, \ldots, \tilde{T}_q\right\}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$. For $\theta \leq \pi/64$ and $18\eta < \delta \leq \rho$, assume that*

- *$\mathcal{P}$ is δ-sparse:* $\min_{i \neq j} \|p_j - p_i\| \geq \delta$,

- *the p_j's are η-close to M:* $\max_{1\leq j\leq q} \mathrm{d}(p_j, M) \leq \eta$,
- $\max_{1\leq j\leq q} \angle(T_{\pi(p_j)}M, \tilde{T}_j) \leq \theta$.

Then, there exist universal constants $c_1 \leq 11, c_2 \leq 252$ and a smooth submanifold $M' \subset \mathbb{R}^D$ such that

1. $\mathcal{P} \subset M'$,
2. $\mathrm{reach}(M') \geq \dfrac{\rho \cdot \left(1 - c_1\left(\frac{\eta}{\delta} + \theta\right)\right)^2}{1 + c_1\left(\frac{\eta}{\delta} + \theta\right) + c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\rho}$,
3. $T_{p_j}M' = \tilde{T}_j$ *for all* $1 \leq j \leq q$,
4. $\mathrm{d_H}(M, M') \leq \delta\theta + \eta$,
5. M *and* M' *are ambient isotopic.*

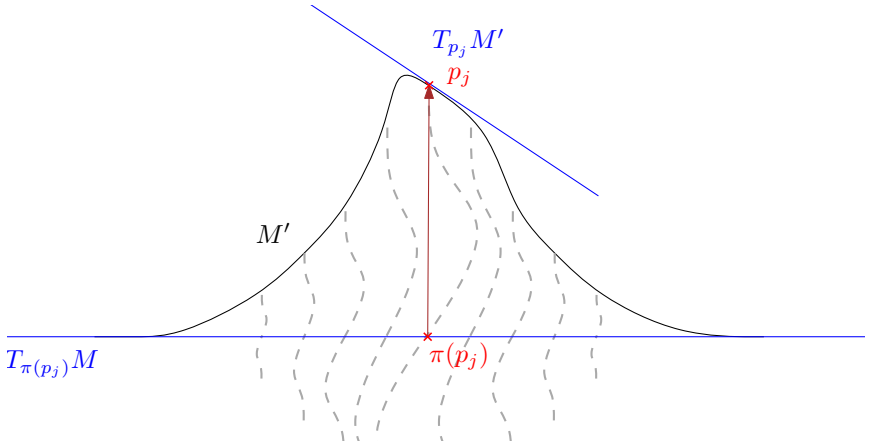


Figure 3: An instance of the interpolating manifold M'. Dashed lines correspond to the image of vertical lines by the ambient diffeomorphism Φ defining $M' = \Phi(M)$.

Theorem 9 fits a manifold M' to noisy points and perturbed tangent spaces with no change of topology and a controlled reach loss. We will use M' as a proxy of M. Indeed, if $\tilde{T}_1, \ldots, \tilde{T}_q$ are estimated tangent spaces at the noisy base points $p_1, \ldots, p_q$, M' has the major advantage to be reconstructed by $\mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$ from Theorem 8. Since M' is topologically and geometrically close to M, we conclude that M is reconstructed as well by transitivity. In other words, Theorem 9 allows to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces. M' is built pushing and rotating M towards the p_j's locally along the vector $(p_j - \pi(p_j))$, as illustrated in Figure 3. Since the construction is quite general and may be applied in various settings, we provide an outline of the construction.

Let $\phi(x) = \exp\left(\frac{\|x\|^2}{\|x\|^2-1}\right)\mathbb{1}_{\|x\|^2<1}$. ϕ is smooth and satisfies $\phi(0) = 1$, $\|\phi\|_\infty \leq 1$ and $\mathrm{d}_0\phi = 0$. For $j = 1, \ldots, q$, it follows easily from the definition of $\angle(T_{\pi(p_j)}M, \tilde{T}_j)$ - *e.g.* by induction on the dimension - that there exists a rotation R_j of $\mathbb{R}^D$ mapping $T_{\pi(p_j)}M$ onto $\tilde{T}_j$ that satisfies

$\|R_j - I_D\|_{op} \leq \theta$. For $\ell > 0$ to be chosen later, and all $a \in \mathbb{R}^D$, let us define $\Phi : \mathbb{R}^D \to \mathbb{R}^D$ by

$$\Phi(a) = a + \sum_{j=1}^{q} \phi\left(\frac{a - \pi(p_j)}{\ell}\right) \underbrace{\left[(R_j - I_D)(a - \pi(p_j)) + (p_j - \pi(p_j))\right]}_{\psi_j(a)}.$$

Φ is designed to map $\pi(p_j)$ onto p_j with $\mathrm{d}_{\pi(p_j)}\Phi = R_j$. Roughly speaking, in balls of radii ℓ around each $\pi(p_j)$, Φ shifts the points in the direction $p_j - \pi(p_j)$ and rotates it around $\pi(p_j)$. Off these balls, Φ is the identity map. To guarantee smoothness, the shifting and the rotation are modulated by the kernel ϕ, as $\|a - \pi(p_j)\|$ increases. Notice that $\mathrm{d}_a\psi_j = (R_j - I_D)$ and $\|\psi_j(a)\| \leq \ell\theta + \eta$ whenever $\phi\left(\frac{a-\pi(p_j)}{\ell}\right) \neq 0$. Defining $M' = \Phi(M)$, the facts that M' fits to $\mathcal{P}$ and $\tilde{T}$ and is Hausdorff-close to M follow by construction. Moreover, Theorem 4.19 of [16] states that the reach is stable with respect to $\mathcal{C}^2$-diffeomorphisms of the ambient space. The estimate on $\mathrm{reach}(M')$ relies on the following lemma stating differentials estimates on Φ.

Lemma 10. *There exist universal constants $C_1 \leq 7/2$ and $C_2 \leq 28$ such that if $6\eta < \ell \leq \delta/3$ and $\theta \leq \pi/64$, $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a global $\mathcal{C}^\infty$-diffeomorphism. In addition, Φ satisfies for all a in $\mathbb{R}^D$,*

$$\|\mathrm{d}_a\Phi\|_{\mathrm{op}} \leq 1 + C_1\left(\frac{\eta}{\ell} + \theta\right), \quad \left\|\mathrm{d}_a\Phi^{-1}\right\|_{\mathrm{op}} \leq \frac{1}{1 - C_1\left(\frac{\eta}{\ell} + \theta\right)}, \quad \left\|\mathrm{d}_a^2\Phi\right\|_{\mathrm{op}} \leq C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right).$$

The ambient isotopy follows easily by considering the weighted version $\Phi_{(t)}(a) = a + t\,(\Phi(a) - a)$ for $0 \leq t \leq 1$ and the same differential estimates. We then take the maximum possible value $\ell = \delta/3$ and $M' = \Phi(M)$.

Remark 11. *Changing slightly the construction of M', one can also build it such that the curvature tensor at each p_j corresponds to that of M at $\pi(p_j)$. For this purpose it suffices to take a localizing function ϕ identically equal to 1 in a neighborhood of 0. This additional condition would impact the universal constants appearing in Theorem 9.*

4.2 Stability of the Tangential Delaunay Complex

Theorem 9 shows that even in the presence of outliers at distance η from M, and with the knowledge of the tangent spaces up to some angle θ, it is still possible to apply Theorem 8 to some virtual manifold M'. Denoting $\tilde{M} = \mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$, since $\mathrm{d_H}(M, \tilde{M}) \leq \mathrm{d_H}(M, M') + \mathrm{d_H}(M', \tilde{M})$ and since the ambient isotopy relation is transitive $M \cong M' \cong \tilde{M}$, we get the following result as a straightforward combination of Theorem 8 and Theorem 9.

Theorem 12 (Stability of the Tangential Delaunay Complex)**.** *There exists $\varepsilon_{d,\rho,\tau_1,\tau_2}$ such that for all $\varepsilon \leq \varepsilon_{d,\rho,\tau_1,\tau_2}$ and all $M \in \mathcal{M}_{D,d,\rho}$, the following holds. Given any finite point cloud $\mathcal{P} \subset \mathbb{R}^D$ and a family $\tilde{T} = \left\{\tilde{T}_p\right\}_{p\in\mathcal{P}}$ of d-dimensional linear subspaces of $\mathbb{R}^D$ such that*

- $\max_{p\in\mathcal{P}} \mathrm{d}(p, M) \leq \tau_1\varepsilon^2$,
- $\mathcal{P}$ *is ε-sparse,*
- $\max_{p\in\mathcal{P}} \angle(T_{\pi(p)}M, \tilde{T}_p) \leq \tau_2\varepsilon$,
- $\max_{x\in M} \mathrm{d}(x, \mathcal{P}) \leq 2\varepsilon$,

then,

- $\mathrm{d_H}\left(M, \mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})\right) \leq \left(C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_1 + \tau_2\right)\varepsilon^2$,
- *M and $\mathrm{Del}^{\omega_*}(\mathcal{P}, \tilde{T})$ are ambient isotopic.*

Indeed, applying the reconstruction algorithm of Theorem 8 even in the presence of noise and uncertainty on the tangent spaces actually recovers the manifold M' built in Theorem 9. M' is isotopic to M and the quality of the approximation of M is at most impacted by the term $\mathrm{d_H}(M, M') \leq (\tau_1 + \tau_2)\varepsilon^2$. The control on $\mathrm{reach}(M')$ from below is crucial, as constants appearing in Theorem 8 are not bounded for arbitrarily small reach ρ'. It is worth noting that no extra analysis of the tangential Delaunay complex was needed to derive its stability. The argument is global, constructive, and may be applied to other reconstruction methods taking tangent spaces as input.

5 Tangent Space Estimation and Denoising Procedure

5.1 Noise-Free Case

We now focus on the estimation of tangent spaces in the model without outliers, namely $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$. The proposed method is similar to that of [2]. A point $p \in M$ being fixed, T_pM is the best local d-dimensional linear approximation of M at p. Performing a Local Principal Component Analysis (PCA) in a neighborhood of p would recover the main directions spanned by M at p, and therefore yield a good approximation of T_pM. For $j = 1, \ldots, n$ and $h > 0$ to be chosen later, define the local covariance matrix at X_j by

$$\hat{\Sigma}_j = \frac{1}{N_j} \sum_{i \neq j} \left(X_i - \bar{X}_j\right)\left(X_i - \bar{X}_j\right)^t \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i),$$

where $N_j = |\mathcal{B}(X_j, h) \cap \mathbb{X}_n|$ is the number of sample points contained in $\mathcal{B}(X_j, h)$, and $\bar{X}_j = \frac{1}{N_j}\sum_{i\neq j} X_i \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i)$ is their barycenter. Set $\hat{T}_j$ to be the linear space spanned by the d eigenvectors associated with the d largest eigenvalues of $\hat{\Sigma}_j$. Computing a basis of $\hat{T}_j$ can be performed naively using a singular value decomposition of the full matrix $\hat{\Sigma}_j$, although fast PCA algorithms [25] may lessen the computational dependence on the ambient dimension.

Proposition 13. *Set $h = \left(c_{d,f_{min},f_{max}} \frac{\log n}{n-1}\right)^{1/d}$. Then, for all $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$, and n large enough, we have*

$$\max_{1\leq j\leq n} \angle(T_{X_j}M, \hat{T}_j) \leq C_{d,f_{min},f_{max}} \frac{h}{\rho},$$

with probability larger than $1 - \left(\frac{1}{n}\right)^{\frac{2}{d}}$.

Furthermore, it is shown in [6] that for $c_{d,f_{min},f_{max}}$ large enough, $\mathbb{X}_n$ is $c_{d,f_{min},f_{max}}h$-dense in M with probability larger than $1 - \left(\frac{1}{n}\right)^{2/d}$. Since $\mathbb{X}_n$ may not be sparse at the scale $c_{d,f_{min},f_{max}}h$, and for the stability reasons described Section 3, we sparsify it with the farthest point sampling algorithm (Section 3.3) with scale parameter $\varepsilon = c_{d,f_{min},f_{max}}h$. Let $\mathbb{Y}_n$ denote the output of the algorithm.

Corollary 14. *With the above notation, for n large enough, with probability at least $1 - 2\left(\frac{1}{n}\right)^{2/d}$,*

- $\max_{X_j \in \mathbb{Y}_n} \angle(T_{X_j} M, \hat{T}_j) \leq C_{d, f_{min}, f_{max}} \frac{\varepsilon}{\rho}$,
- $\max_{x \in M} \mathrm{d}(x, \mathbb{Y}_n) \leq 2\varepsilon$.
- $\mathbb{Y}_n$ *is ε-sparse,*

In other words, the previous result shows that $\mathbb{Y}_n$ satisfies the assumptions of Theorem 12. As a consequence, taking the notation $\hat{T}_{\mathbb{Y}_n} = \left\{\hat{T}_j\right\}_{X_j \in \mathbb{Y}_n}$, we define $\hat{M}_{\texttt{TDC}}$ to be the tangential Delaunay complex computed on $\mathbb{Y}_n$ and the collection of estimated tangent spaces $\hat{T}_{\mathbb{Y}_n}$.

Definition 15. *With the notation above, define* $\hat{M}_{\texttt{TDC}} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n, \hat{T}_{\mathbb{Y}_n}\right)$.

Combining Theorem 12 and Corollary 14, it is clear that $\hat{M}_{\texttt{TDC}}$ satisfies Theorem 6.

5.2 Clutter Noise Case

We now address the denoising problem, following ideas from [18]. To distinguish whether X_j is an outlier or belongs to M, we notice again that points drawn from M approximately lie on a low dimensional structure. On the other hand, the neighborhood points of an outlier drawn far away from M should typically be distributed in an isotropic way. Let $k_1, k_2, h > 0$, $x \in \mathbb{R}^D$ and $T \subset \mathbb{R}^D$ a d-dimensional linear subspace. The *slab* at x in the direction T is the set $S(x,T) = \{x\} \oplus \mathcal{B}_T(0, k_1 h) \oplus \mathcal{B}_{T^\perp}(0, k_2 h^2) \subset \mathbb{R}^D$, where $\oplus$ denotes the Minkovski sum, and $\mathcal{B}_T, \mathcal{B}_{T^\perp}$ are the Euclidean balls in T and $T^\perp$ respectively.

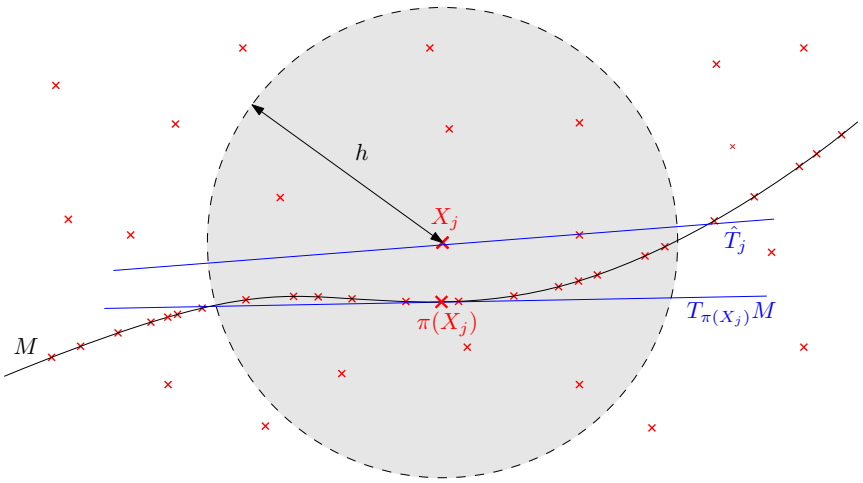


Figure 4: Local PCA at an outlier point $X_j \in \mathbb{X}_n$.

Using the notation of Section 2.1, we roughly argue that for all $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, Figure 5 indicates that

$$\begin{array}{lll} P\left(S(x, T_\pi(x)M)\right) \asymp h^d & & \text{if} \quad \mathrm{d}(x,M) \leq h^2, \\ P\left(S(x,T)\right) \asymp h^{2D-d} & \text{for all } T, & \text{if} \quad \mathrm{d}(x,M) > h^2, \end{array}$$

as h goes to 0, for k_1 and k_2 small enough. Since $h^{2D-d} \ll h^d$, the measure of the slabs $P(S(x,T))$ clearly is discriminative for denoising. The empirical counterpart of the above observation relies on counting how many sample points lie in the slabs of direction $\hat{T}_j$. For all $j = 1, \ldots, n$ define the empirical slab $S_j = S(X_j, \hat{T}_j)$ (see Figure 5) and let $|S_j \cap \mathbb{X}_n|$ denote the number of sample points it contains.

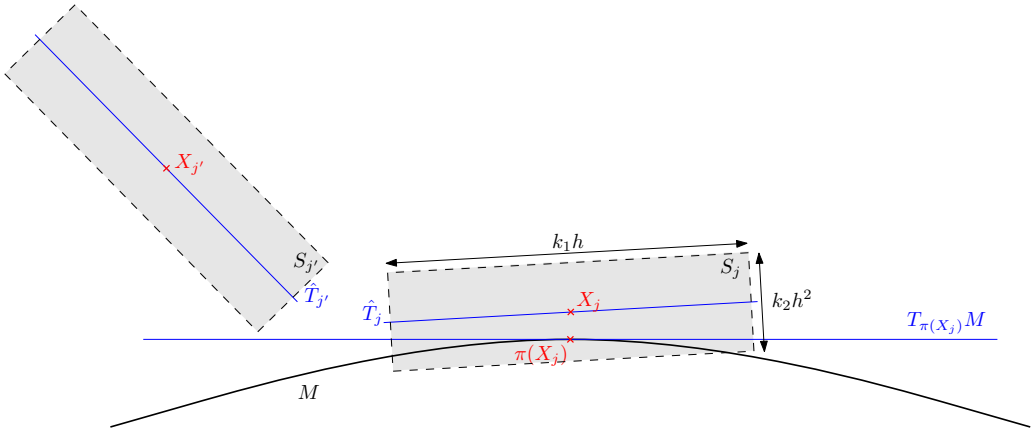


Figure 5: The slab S_j is centered at X_j and has size $k_1 h$ in the d directions spanned by $\hat{T}_j$, and size $k_2 h^2$ in the $D-d$ directions normal to $\hat{T}_j$.

Proposition 16. *In the clutter noise model, fix $m > 1$, k_1 and k_2 small enough, and set $h = \left(c_{d,f_{min},f_{max}} \frac{\log n}{\beta(n-1)}\right)^{\frac{1}{d+1}}$. Then, for n large enough, with probability larger than $1 - 8\left(\frac{1}{n}\right)^{\frac{2}{d}} - \left(\frac{1}{n}\right)^{2D}$, we have*

- $\sqrt{h}\frac{n}{\log(n)}|S_j \cap \mathbb{X}_n| \geq m$ *if* $X_j \in M$,
- $\sqrt{h}\frac{n}{\log(n)}|S_j \cap \mathbb{X}_n| \leq 1/m$ *if* $\mathrm{d}(X_j, M) > h^2$,
- *for all X_j such that* $\mathrm{d}(X_j, M) \leq h/\sqrt{2}$, $\angle(\hat{T}_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max},\rho} h$.

This result is twofold. Not only can we distinguish data and outliers within an offset of size h^2 around M, but we also provide a rate of convergence of the estimated tangent spaces for these points. Filtering the data set $\mathbb{X}_n$ by only keeping points X_j such that $\sqrt{h}\frac{n}{\log(n)}|S_j \cap \mathbb{X}_n| \geq m$, Proposition 16 states that: (i) no data point $X_j \in M$ is removed, (ii) any undetected outlier satisfies $\mathrm{d}(X_j, M) < h^2$, (iii) all the kept points have good estimated tangent spaces. In what follows, we will denote the filtered point cloud by

$$\mathbb{Z}_n = \left\{ X_j \in \mathbb{X}_n | \sqrt{h}\frac{n}{\log(n)}|S_j \cap \mathbb{X}_n| \geq m \right\}.$$

According to Proposition 16, the denoising procedure removes no data point on M with high probability. In other words, $\mathbb{X}_n \cap M = \mathbb{Z}_n \cap M$, and as a consequence, $\max_{x \in M} \mathrm{d}(x, \mathbb{Z}_n) \leq c_{d,f_{min},\beta} \left(\frac{\log n}{n}\right)^{1/d} \ll$

h with high probability, using the same result of [6] as in the noise-free case. As in the previous section, $\mathbb{Z}_n$ is itself pruned to ensure a good density-sparsity ratio using the farthest point sampling algorithm (Section 3.3), with scale parameter $\varepsilon = h$. Let $\mathbb{Y}_n^+$ denote the output of the algorithm.

Corollary 17. *With the above notation, for n large enough, with probability larger than $1 - 10\left(\frac{1}{n}\right)^{2/d}$,*

- $\max_{X_j \in \mathbb{Y}_n^+} \mathrm{d}(X_j, M) \leq \varepsilon^2$,
- $\mathbb{Y}_n^+$ *is ε-sparse,*
- $\max_{X_j \in \mathbb{Y}_n^+} \angle(\hat{T}_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max},\rho}\varepsilon$,
- $\max_{x \in M} \mathrm{d}(x, \mathbb{Y}_n^+) \leq 2\varepsilon$.

Taking the notation $\hat{T}_{\mathbb{Y}_n^+} = \left\{\hat{T}_j\right\}_{X_j \in \mathbb{Y}_n^+}$, we are now able to define the estimator $\hat{M}_{\texttt{TDC+}}$.

Definition 18. *With the notation above, define* $\hat{M}_{\texttt{TDC+}} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n^+, \hat{T}_{\mathbb{Y}_n^+}\right)$.

Combining Theorem 12 and Corollary 17 it is clear that $\hat{M}_{\texttt{TDC+}}$ satisfies Theorem 7.

6 Conclusion

In this work, we gave results on explicit manifold reconstruction with simplicial complexes. We built estimators $\hat{M}_{\texttt{TDC}}$ and $\hat{M}_{\texttt{TDC+}}$ in two statistical models. We proved minimax rates of convergence for the Hausdorff distance and consistency results for ambient isotopic reconstruction. Since $\hat{M}_{\texttt{TDC}}$ is minimax optimal and uses the tangential Delaunay complex of [3], the latter is proved to be optimal. Moreover, rates of [18] are proved to be achievable with simplicial complexes that are computable using existing algorithms. To prove the stability of the tangential Delaunay complex, a generic interpolation result was derived. In the process, a tangent space estimation procedure and a denoising method both based on local PCA were studied.

In the model with outliers, the proposed method does not provably achieve the minimax rate of convergence (Theorem 7). This sub-optimality comes from a deficiency in the denoising part. Indeed, we can prove that the procedure detects each outlier point that is at distance $\left(\frac{\log n}{n}\right)^{2/d}$ with high probability, but not all of them simultaneously. We wonder whether this is only a artefact in the proof, or that it is impossible under computational constraints. Extensions of this method to other models remain to be studied, especially the additive noise model [18] where a challenging question consists in studying how deconvolution behaves with tangent spaces estimation.

Even though Theorem 9 is applied to manifold estimation, the authors believe it may be applied in various settings. Beyond its statement, the way that it is used is quite general. When intermediate objects (here, tangent spaces) are used in a procedure, this kind of proxy method can provide extensions of existing results to the case where these objects are only approximated.

As local PCA is performed throughout the paper, the knowledge of the bandwidth h is needed for actual implementation. In practice its choice is a difficult question and adaptive selection of h remains to be considered.

In the process, we derived rates of convergence for tangent space estimation. The optimality of the method will be the object of a further paper.

Acknowledgements

The authors would like to thank Jean-Daniel Boissonnat , Frédéric Chazal, Pascal Massart, and Steve Oudot for their insight and the interest they brought to this work.

References

[1] Stephanie B. Alexander and Richard L. Bishop. Gauss equation and injectivity radii for subspaces in spaces of curvature bounded above. Geom. Dedicata, 117:65–84, 2006.

[2] E. Arias-Castro, G. Lerman, and T. Zhang. Spectral Clustering Based on Local PCA. ArXiv e-prints, January 2013.

[3] Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. Discrete Comput. Geom., 51(1):221–267, 2014.

[4] Jean-Daniel Boissonnat, Leonidas J. Guibas, and Steve Y. Oudot. Manifold reconstruction in arbitrary dimensions using witness complexes. Discrete Comput. Geom., 42(1):37–70, 2009.

[5] Olivier Bousquet. A Bennett concentration inequality and its application to suprema of empirical processes. C. R. Math. Acad. Sci. Paris, 334(6):495–500, 2002.

[6] F. Chazal, M. Glisse, C. Labruère, and B. Michel. Optimal rates of convergence for persistence diagrams in Topological Data Analysis. ArXiv e-prints, May 2013.

[7] Frédéric Chazal, David Cohen-Steiner, and André Lieutier. A sampling theory for compact sets in Euclidean space. In Computational geometry (SCG'06), pages 319–326. ACM, New York, 2006.

[8] Siu-Wing Cheng, Tamal K. Dey, and Edgar A. Ramos. Manifold reconstruction from point samples. In Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pages 1018–1027. ACM, New York, 2005.

[9] Kenneth L Clarkson. Building triangulations using ε-nets. In Proceedings of the thirty-eighth annual ACM symposium on Theory of computing, pages 326–335. ACM, 2006.

[10] Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. Adv. in Appl. Probab., 36(2):340–354, 2004.

[11] Giuseppe De Marco, Gianluca Gorni, and Gaetano Zampieri. Global inversion of functions: an introduction. NoDEA Nonlinear Differential Equations Appl., 1(3):229–248, 1994.

[12] Tamal K. Dey. Curve and surface reconstruction: algorithms with mathematical analysis, volume 23 of Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2007.

[13] Tamal K. Dey and Kuiyu Li. Topology from data via geodesic complexes. Technical report OSU-CISRC-3/09-TR05.

[14] Lutz Dümbgen and Günther Walther. Rates of convergence for random approximations of convex sets. Adv. in Appl. Probab., 28(2):384–393, 1996.

[15] Ramsay Dyer, Gert Vegter, and Mathijs Wintraecken. Riemannian Simplices and Triangulations. In Lars Arge and János Pach, editors, 31st International Symposium on Computational Geometry (SoCG 2015), volume 34 of Leibniz International Proceedings in Informatics (LIPIcs), pages 255–269, Dagstuhl, Germany, 2015. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

[16] Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93:418–491, 1959.

[17] Herbert Federer. Geometric measure theory. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.

[18] Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdorff loss. Ann. Statist., 40(2):941–963, 2012.

[19] Gene H. Golub and Charles F. Van Loan. Matrix computations. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[20] Arlene K. H. Kim and Harrison H. Zhou. Tight minimax rates for manifold estimation under Hausdorff loss. Electron. J. Stat., 9(1):1562–1582, 2015.

[21] L. LeCam. Convergence of estimates under dimensionality restrictions. Ann. Statist., 1:38–53, 1973.

[22] E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. Ann. Statist., 23(2):502–524, 1995.

[23] Pascal Massart. Concentration inequalities and model selection, volume 1896 of Lecture Notes in Mathematics. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[24] Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. Discrete Comput. Geom., 39(1-3):419–441, 2008.

[25] Alok Sharma and Kuldip K Paliwal. Fast principal component analysis using fixed-point algorithm. Pattern Recognition Letters, 28(10):1151–1155, 2007.

A Interpolation Theorem

This section is devoted to prove the interpolation results of Section 4.1. For sake of completeness, let us state a stability result for the reach with respect to $\mathcal{C}^2$ diffeomorphisms.

Lemma 19 (Theorem 4.19 in [16])**.** *Let $A \subset \mathbb{R}^D$ with $\mathrm{reach}(A) \geq \rho > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathcal{C}^1$-diffeomorphism such that Φ,Φ^{-1}, and $\mathrm{d}\Phi$ are Lipschitz with Lipschitz constants K,N and R respectively, then*

$$\mathrm{reach}(\Phi(A)) \geq \frac{1}{(K\rho^{-1} + R)N^2}.$$

Writing $\phi_\ell(\cdot) = \phi(\cdot/\ell)$, we recall that $\psi_j(a) = (R_j - I_D)(a - \pi(p_j)) + (p_j - \pi(p_j))$ and

$$\Phi(a) = a + \sum_{j=1}^{q} \phi_\ell\left(a - \pi(p_j)\right) \psi_j(a). \tag{1}$$

Let us denote $b_1 = \sup_x \|\mathrm{d}_x\phi\|$, $b_2 = \sup_x \left\|\mathrm{d}^2_x\phi\right\|_{\mathrm{op}}$, and write $C_1 = 1 + b_1$, $C_2 = b_2 + 2b_1$. Straightforward computation yields $C_1 \leq 7/2$ and $C_2 \leq 28$.

Proof of Lemma 10. First notice that the sum appearing in (1) consists of at most one term. Indeed, since $\phi \equiv 0$ outside $\mathcal{B}(0,1)$, if $\phi_\ell\left(a - \pi(p_j)\right) \neq 0$ for some $j \in \{1, \ldots, q\}$, then $\|a - \pi(p_j)\| \leq \ell$. Consequently, for all $i \neq j$,

$$\begin{aligned}
\|a - \pi(p_i)\| &\geq \|p_j - p_i\| - \|p_j - \pi(p_j)\| - \|\pi(p_j) - a\| - \|\pi(p_i) - p_i\| \\
&\geq \delta - \eta - \ell - \eta \\
&\geq \delta - 2\ell \geq \ell,
\end{aligned}$$

where we used that $6\eta \leq \ell \leq \delta/3$. Therefore, $\phi_\ell\left(a - \pi(p_i)\right) = 0$ for all $i \neq j$. In other words, if a p_j actually appears in $\Phi(a)$ then the others do not.

Global diffeomorphism: As the sum in (1) is at most composed of one term, chain rule yields

$$\begin{aligned}
\left\|\mathrm{d}_a\Phi - \mathrm{Id}\right\|_{\mathrm{op}} &= \max_{1\leq j\leq q} \left\|\mathrm{d}_a\left[\phi_\ell\left(a - \pi(p_j)\right)\psi_j(a)\right]\right\|_{\mathrm{op}} \\
&= \max_{1\leq j\leq q} \left\| \psi_j(a)\, \frac{\mathrm{d}_b\phi}{\ell}\Big|_{b=\frac{a-\pi(p_j)}{\ell}} + \phi_\ell\left(a - \pi(p_j)\right)\left(R_j - I_D\right) \right\|_{\mathrm{op}} \\
&\leq \big(b_1 + 1\big)\theta + b_1\frac{\eta}{\ell} < 1,
\end{aligned}$$

where the last line follows from $b_1 \leq 5/2$, $6\eta \leq \ell$ and $\theta \leq \pi/64$.

Therefore, $\mathrm{d}_a\Phi$ is invertible for all $a \in \mathbb{R}^D$, and $\left(\mathrm{d}_a\Phi\right)^{-1} = \sum_{i=0}^{\infty}\left(I_D - \mathrm{d}_a\Phi\right)^i$. Φ is a local diffeomorphism according to the local inverse function theorem. Moreover, $\|\Phi(a)\| \to \infty$ as $\|a\| \to \infty$, so that Φ is a global $\mathcal{C}^\infty$-diffeomorphism by Hadamard-Cacciopoli theorem [11].

Differentials estimates: (i) First order: From the estimates above,

$$\left\|\mathrm{d}_a\Phi\right\|_{\mathrm{op}} \leq \left\|I_D\right\|_{op} + \left\|\mathrm{d}_a\Phi - I_D\right\|_{op} \leq 1 + \big(b_1 + 1\big)\theta + b_1\frac{\eta}{\ell}.$$

(ii) Inverse: Write for all $a \in \mathbb{R}^D$,

$$\begin{aligned}\left\|\mathrm{d}_{\Phi(a)}\Phi^{-1}\right\|_{\mathrm{op}} &= \left\|(\mathrm{d}_a\Phi)^{-1}\right\|_{\mathrm{op}} = \left\|\sum_{i=0}^{\infty}\left(I_D - \mathrm{d}_a\Phi\right)^i\right\|_{\mathrm{op}} \\ &\leq \frac{1}{1-\left\|\mathrm{Id}-\mathrm{d}_a\Phi\right\|_{\mathrm{op}}} \leq \frac{1}{1-\big(b_1+1\big)\theta - b_1\frac{\eta}{\ell}},\end{aligned}$$

where the third line holds since $\left\|\mathrm{d}_a\Phi - \mathrm{Id}\right\|_{\mathrm{op}} < 1$, and $\|\cdot\|_{\mathrm{op}}$ is sub-multiplicative.
(iii) Second order: Again, since the sum (1) includes at most one term,

$$\begin{aligned}\left\|\mathrm{d}_a^2\Phi\right\|_{\mathrm{op}} &= \max_{1\leq j\leq q}\left\|\mathrm{d}_a^2\left[\phi_\ell\left(a-\pi(p_j)\right)\psi_j(a)\right]\right\|_{\mathrm{op}} \\ &\leq \max_{1\leq j\leq q}\left\{\frac{\left\|\mathrm{d}^2\phi\right\|_{\mathrm{op}}}{\ell^2}\left\|\psi_j(a)\right\| + 2\frac{\left\|\mathrm{d}\phi\right\|_{\mathrm{op}}}{\ell}\left\|R_j - I_D\right\|_{\mathrm{op}}\right\} \\ &\leq b_2\frac{\eta}{\ell^2} + \left(b_2+2b_1\right)\frac{\theta}{\ell}.\end{aligned}$$

□

Proof of Theorem 9. Set $\ell = \delta/3$ and $M' = \Phi(M)$.

- *Interpolation:* For all j, $p_j = \Phi(\pi(p_j)) \in M'$ by construction since $\phi_\ell(0) = 1$.
- *Tangent spaces:* Since $\mathrm{d}_x\phi_l|_{x=0} = 0$, for all $j \in \big\{1,\ldots,q\big\}$, $\mathrm{d}_a\Phi|_{a=\pi(p_j)} = R_j$. Thus,

$$\begin{aligned}T_{p_j}M' &= T_{\Phi(\pi(p_j))}\Phi(M) \\ &= \mathrm{d}_a\Phi|_{a=\pi(p_j)}\left(T_{\pi(p_j)}M\right) \\ &= R_j\left(T_{\pi(p_j)}M\right) = T_j,\end{aligned}$$

  by definition of R_j.

- *Proximity to M:* It follows from the correspondence

$$\begin{aligned}\left\|\Phi(a)-a\right\| &\leq \sup_{a\in\mathbb{R}^D}\max_{1\leq j\leq q}\phi_\ell\left(a-\pi(p_j)\right)\left\|\psi_j(a)\right\| \\ &\leq \ell\theta + \eta \leq \delta\theta + \eta.\end{aligned}$$

- *Isotopy:* Consider the continuous family of maps

$$\Phi_{(t)}(a) = a + t\left(\sum_{j=1}^{q}\phi_\ell\left(a-\pi(p_j)\right)\psi_j(a)\right),$$

  for $0 \leq t \leq 1$. Since $\Phi_{(t)} - \mathrm{Id} = t\big(\Phi - \mathrm{Id}\big)$, the arguments above show that $\Phi_{(t)}$ is a global diffeomorphism of $\mathbb{R}^D$ for all $t \in [0,1]$. Moreover $\Phi_{(0)} = \mathrm{Id}$, and $\Phi_{(1)} = \Phi$. Thus, $M = \Phi_{(0)}(M)$ and $M' = \Phi_{(1)}(M)$ are ambient isotopic.

- *Reach lower bound:* The differentials estimates of order 1 and 2 of Φ translate into estimates on Lipschitz constants of Φ,Φ^{-1} and $\mathrm{d}\Phi$. Applying Lemma 19 leads to

$$\mathrm{reach}\,(M') \geq \frac{\left(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\right)^2}{\frac{1 + C_1\left(\frac{\eta}{\ell} + \theta\right)}{\rho} + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)} = \rho \cdot \frac{\left(1 - C_1\left(\frac{\eta}{\ell} + \theta\right)\right)^2}{1 + C_1\left(\frac{\eta}{\ell} + \theta\right) + C_2\left(\frac{\eta}{\ell^2} + \frac{\theta}{\ell}\right)\rho}.$$

  The desired lower bound follows by replacing ℓ by its value $\delta/3$, and setting $c_1 = 3C_1 \leq 21/2 \leq 11$ and $c_2 = 9C_2 \leq 252$.

□

B Some Geometric Properties Under Reach Condition

In this section we state intermediate results that connect Euclidean an geodesic quantities under reach condition. We begin with a result connecting reach and principal curvatures.

Proposition 20 (Proposition 2.1 in [13])**.** *For all $x \in M$, writing II_x for the second fundamental form of M at x, for all unitary $w \in T_xM$, we have $\|II_x(w,w)\| \leq 1/\rho$.*

For all $x \in M$ and $v \in T_xM$, let us denote by $\exp_x(v)$ the exponential map at x of direction v. According to the following Proposition, this exponential map turns out to be a diffeomorphism on balls of radius at most $\pi\rho$.

Proposition 21 (Corollary 1.4 in [1])**.** *The injectivity radius of M is at least $\pi\rho$.*

Denoting by $\mathrm{d}_M(\cdot,\cdot)$ the geodesic distance on M, we are in position to connect geodesic and Euclidean distance. In what follows, we fix the constant $\alpha = 1 + \frac{1}{4\sqrt{2}}$.

Proposition 22. *For all $x, y \in M$ such that $\|x - y\| \leq \rho/4$,*

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq \alpha\,\|x - y\|\,.$$

Moreover, writing $y = \exp_x(rv)$ for $v \in T_xM$ with $\|v\| = 1$ and $r \leq \rho/4$,

$$y = x + rv + R(r,v)$$

with $\|R(r,v\| \leq \frac{r^2}{2\rho}$. As a consequence,

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq \|x - y\| + \frac{\alpha^2\,\|x - y\|^2}{2\rho}.$$

Proof of Proposition 22. The first statement is a direct consequence of Proposition 6.3 in [24]. Let us define $u(t) = \exp_x(tv) - \exp_x(0) - tv$ and $w(t) = \exp_x(tv)$ for all $0 \leq t \leq r$. It is clear that $u(0) = 0$ and $u'(0) = 0$. Moreover, $\|u''(t)\| = \left\|II_{w(t)}\left(w'(t), w'(t)\right)\right\| \leq 1/\rho$. Therefore, a Taylor expansion at order two gives $\|R(r,v)\| = u(r) \leq r^2/(2\rho)$. Applying the first statement of the proposition gives $r \leq \alpha\,\|x - y\|$. Therefore,

$$\|x - y\| \leq \mathrm{d}_M(x,y) \leq r \leq \|x - y\| + \|R(r,v)\| \leq \|x - y\| + \frac{\alpha^2\,\|x - y\|^2}{2\rho}.$$

□

The next proposition gives bounds on the volume form expressed in polar coordinates in a neighborhood of points of M.

Proposition 23. *Let $x \in M$ be fixed. Denote by $J(r,v)$ the Jacobian of the volume form expressed in polar coordinates around x, for $r \leq \frac{\rho}{4}$ and v a unit vector in T_xM. In other words, if $y = \exp_x(rv)$, $dV_y = J(r,v)drdv$. Then*

$$c_d r^{d-1} \leq J(r,v) \leq C_d r^{d-1},$$

where $c_d = 2^{-d}$ and $C_d = 2^d$.

Proof of Proposition 23. Denoting $A_{r,v} = \mathrm{d}_{rv}\exp_x$, the Area Formula (see, *e.g.*, Section 3.2.5 in [17]) gives $J(r,v) = r^{d-1}\sqrt{\det\left(A_{r,v}^t A_{r,v}\right)}$. Note that from Proposition 2.1 in [13], the sectional curvatures in M are bounded by $|\kappa| \leq 2/\rho^2$. Therefore, the Rauch theorem (see, *e.g.*, Lemma 5 of [15]) states that

$$\left(1 - \frac{r^2}{3\rho^2}\right)\|w\| \leq \|A_{r,v}w\| \leq \left(1 + \frac{r^2}{\rho^2}\right)\|w\|,$$

for all $w \in T_xM$. As a consequence,

$$2^{-d} \leq \left(1 - \frac{r^2}{3\rho^2}\right)^d \leq \sqrt{\det\left(A_{r,v}^t A_{r,v}\right)} \leq \left(1 + \frac{r^2}{\rho^2}\right)^d \leq 2^d.$$

□

We will also need the following result giving an estimate of the size of projections of Euclidean balls onto M, when the center is close to M.

Proposition 24. *Let $x \in \mathbb{R}^D$ be such that $\mathrm{d}(x,\mathcal{M}) = \Delta \leq h \leq \frac{\rho}{8}$, and let y denote $\pi(x)$. Then,*

$$\mathcal{B}\left(y, \left(1 - \frac{\alpha^2\Delta}{\rho}\right)r_h\right) \cap M \subset \mathcal{B}(x,h) \cap M \subset \mathcal{B}\left(y, \left(1 + \frac{4\alpha^2\Delta}{\rho}\right)r_h\right) \cap M,$$

where $r_h^2 + \Delta^2 = h^2$.

Proof of Proposition 24. Let z be in $M \cap \mathcal{B}(x,h)$, and denote by δ the quantity $\|z - y\|$. Since $\delta \leq 2h$, according to proposition 22, we may write

$$\begin{aligned}\delta^2 &= \|z - x\|^2 - \|y - x\|^2 - 2\langle z - y, y - x\rangle \\ &\leq r_h^2 + \frac{\Delta\alpha^2}{\rho}\delta^2,\end{aligned}$$

hence $\delta \leq \left(1 + \frac{4\Delta\alpha^2}{\rho}\right)r_h$.

On the other hand, the same inequality ensures that, for any $z \in \mathcal{B}(y,\rho/4) \cap M$,

$$\|z - x\|^2 \leq \delta^2 + \Delta^2 + \frac{\alpha^2\Delta}{\rho}\delta^2.$$

Hence $\delta \leq \left(1 - \frac{\alpha^2\Delta}{\rho}\right)r_h$ ensures that $\|z - x\| \leq h$.

□

C Some Technical Properties of the Statistical Model

Lemma 25. *Let $P \in \mathcal{U}_M(f_{min}, f_{max})$. Then for all $p \in M$ and $r \leq \rho/4$,*

$$P(\mathcal{B}(p,r)) \geq a_d f_{min} r^d.$$

As a consequence, for n large enough and for all $Q \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$, with probability larger that $1 - \left(\frac{1}{n}\right)^{2/d}$,

$$\mathrm{d_H}(M, \mathbb{X}_n \cap M) \leq C_{d,f_{min},\beta} \left(\frac{\log n}{n}\right)^{1/d}.$$

Proof of Lemma 25. The first statement is a direct corollary of Proposition 23. Denoting $a = a_d f_{min}$, it corresponds to the (a,d)-standardness condition of [6] for small radii r. For all $k = 0, \ldots, n$, conditionally on the event $\{|\mathbb{X}_n \cap M| = k\}$, $\mathbb{X}_n \cap M$ has the distribution of a k-sample of P. But from the previous point, P fulfils the (a,d)-standard assumption of [6] for $r \leq \rho/4$. Looking carefully at the proof of Theorem 3.3 in [6] shows that its conclusion still holds for measures satisfying the (a,d)-standard assumption for small radii only. Therefore, for $r \leq \rho/8$,

$$\mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r \middle| |\mathbb{X}_n \cap M| = k\right) \leq \frac{4^d}{ar^d} \exp\left(-k\frac{a}{2^d} r^d\right).$$

Hence,

$$\begin{aligned}
&\mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r\right) \\
&= \sum_{k=0}^{n} \mathbb{P}\left(\mathrm{d_H}(M, \mathbb{X}_n \cap M) > r \middle| |\mathbb{X}_n \cap M| = k\right) \mathbb{P}(\mathbb{X}_n \cap M| = k) \\
&\leq \sum_{k=0}^{n} \frac{4^d}{ar^d} \exp\left(-k\frac{a}{2^d} r^d\right) \binom{n}{k} \beta^k (1-\beta)^{n-k} \\
&= \frac{4^d}{ar^d} \left[1 - \beta\left(1 - \exp\left(-\frac{a}{2^d} r^d\right)\right)\right]^n \\
&\leq \frac{4^d}{ar^d} \exp\left[-n\beta\left(1 - \exp\left(-\frac{a}{2^d} r^d\right)\right)\right] \\
&\leq \frac{4^d}{ar^d} \exp\left[-n\beta \frac{a}{2^{d+1}} r^d\right],
\end{aligned}$$

whenever $r \leq \rho/8$ and $ar^d \leq 2^d$. Taking $r = C_{d,f_{min},\beta} \left(\frac{\log n}{n}\right)^{1/d}$ with $C^d_{d,f_{min},\beta} \frac{\beta a}{2^{d+1}} \geq 1 + 2/d$ yields the result. □

We are now able to prove Lemma 2.

Proof of Lemma 2. For $\varepsilon \leq \rho/4$, let $\mathrm{cv}_M(\varepsilon)$ be the geodesic covering number of M. $\mathrm{cv}_M(\varepsilon)$ is the minimal number k of geodesic balls $\mathcal{B}_M(x_1, \varepsilon), \ldots, \mathcal{B}_M(x_k, \varepsilon)$ of radius ε needed to cover M. Notice that according to Lemma 25 and Proposition 22,

$$P\left(\mathcal{B}_M(x_k, \varepsilon)\right) \geq \frac{f_{min} \varepsilon^d}{c_d}.$$

Therefore, a packing argument (see Section B.1 of [6]) yields

$$\mathrm{cv}_M(\varepsilon) \leq \frac{2^d c_d}{f_{min}\varepsilon^d}.$$

Now, take $p, q \in M$ such that $\|p-q\| = \mathrm{diam}(M)$. Let γ be a minimal geodesic joining p and q. Since γ is minimal, it intersects each $\mathcal{B}_M(x_k, \varepsilon)$ at most once. Reorder the centers of the balls of the covering that intersect γ successively: $x_{(1)}, \ldots, x_{(N)}$. Taking $\varepsilon = \rho/4$ and using triangle inequality, we may write

$$\begin{aligned}
\mathrm{diam}(M) &= \|p-q\| \\
&\leq \left\|p - x_{(1)}\right\| + \sum_{k=1}^{N-1} \left\|x_{(k)} - x_{(k+1)}\right\| + \left\|x_{(K)} - q\right\| \\
&\leq 2\varepsilon N \leq 2\varepsilon \mathrm{cv}_M(\varepsilon) \\
&\leq 2\frac{\rho}{4}\frac{2^d c_d}{f_{min}\left(\frac{\rho}{4}\right)^d} := \frac{C_d}{f_{min}\rho^{d-1}}.
\end{aligned}$$

□

D Matrix Decomposition and Principal Angles

This section is devoted to show the following proposition linking matrix decomposition and principal angles between linear spans of main eigenvectors. For symmetric matrices, we let $\lambda_i(\cdot)$ denote their i-th largest eigenvalue and $\lambda_{min}(\cdot)$ the smallest one.

Proposition 26. *Let $O \in \mathbb{R}^{D\times D}$, $B \in \mathbb{R}^{d\times d}$ be symmetric matrices such that*

$$O = \left(\begin{array}{c|c} B & 0 \\ \hline 0 & 0 \end{array}\right) + E,$$

$\lambda_{min}(B) \geq 1 - e_1$ and $\|E\|_{\mathcal{F}} \leq e_2$. Let T_0 (resp. T) be the vector space spanned by the first d vectors of the canonical basis, (resp. by the first d eigenvectors of O). If $e_1 + e_2 \leq 1/2$, then $\angle(T_0, T) \leq 2de_2$.

The proof of Proposition 26 relies on the application of Wielandt-Hoffmann Theorem, which is recalled below.

Theorem 27 (Wielandt-Hoffmann, Theorem 8.1.4 in [19])**.** *Let $A, E \in \mathbb{R}^{D\times D}$ be symmetric matrices. Then,*

$$\sum_{i=1}^{D} \left(\lambda_i(A+E) - \lambda_i(A)\right)^2 \leq \|E\|_{\mathcal{F}}^2.$$

Proof of Proposition 26. Denoting $\lambda_i = \lambda_i(O)$ and $b_i = \lambda_i(B)$, Theorem 27 yields

$$\sum_{i=1}^{d}(\lambda_i - b_i)^2 + \sum_{i=d+1}^{D} \lambda_i^2 \leq e_2^2.$$

Let $u_1, \dots, u_d$ denote the first d eigenvectors of O (that span T). Moreover, denote by π_1 the orthogonal projection onto T_0 and by π_2 the orthogonal projection onto its orthogonal complement $T_0^\perp$. Then O takes the form $\tilde{\pi}_1 + E$, where $\tilde{\pi}_1$ has the same range as π_1. Let k be in $\{1, \dots, d\}$, then, by definition

$$(\tilde{\pi}_1 + E)\, u_k = \lambda_k u_k.$$

Therefore

$$\pi_2 E u_k = \lambda_k \pi_2 u_k.$$

According to Theorem 27, $|\lambda_k - b_k| \leq e_1$, hence $\lambda_k \geq 1 - e_1 - e_2$. We deduce that

$$\|\pi_2 u_k\| \leq \frac{e_2}{1 - e_1 - e_2} \leq 2e_2.$$

As a consequence, for all $u \in T$, $\|\pi_2 u\| \leq 2e_2 d\, \|u\|$. Hence,

$$\begin{aligned} \angle(T_0, T) &= \max_{u \in T} \max_{v' \in T_0^\perp} \frac{|\langle u, v' \rangle|}{\|u\| \|v'\|} \\ &\leq 2e_2 d. \end{aligned}$$

□

E Local PCA for Tangent Space Estimation

This section is dedicated to the proofs of Section 5. The models with and without are considered jointly as often as possible. When separate analyses are needed, we refer to the model without outliers $\mathcal{G}_{D,d,f_{min},f_{max},\rho}$ as "$\beta = 1$", and to the one with outliers $\mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$ as "$\beta < 1$". $\wedge$ and $\vee$ denote respectively the minimum and the maximum of real numbers. We recall that

$$\hat{\Sigma}_j = \frac{1}{N_j} \sum_{i \neq j} \left(X_i - \bar{X}_j\right) \left(X_i - \bar{X}_j\right)^t \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i),$$

$$\bar{X}_j = \frac{1}{N_j} \sum_{i \neq j} X_i \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i),$$

$$N_j = \sum_{i \neq j} \mathbb{1}_{\mathcal{B}(X_j,h)}(X_i).$$

When a single point X_j is considered, we set $j = 1$ without loss of generality. Let $U_{2:n}$ denote $\left(\mathbb{1}_{\mathcal{B}(X_1,h)}(X_i)\right)_{i=1,\dots,n}$, and let $V_{2:n}$ denote the vector such that $V_i = 1$ if X_i is drawn from the noise distribution. It is immediate that the (U_i, V_i)'s are independent and identically distributed, with common law (U, V). With a slight abuse of notation, we will denote by $\mathbb{P}^A$ (respectively $\mathbb{E}^A$) conditional probabilities and expectations, A being an event or a random variable.

We will restrict our attention to points X_j that are close enough to M so that $B(X_j, h) \cap M$ has enough probability mass. To this aim, we adopt the following notation. Let $p(x)$ denote $P(\mathcal{B}(x, h))$, and set $p(x) = \beta q(x) + (1 - \beta) q'$, where $q(x) = Q(\mathcal{B}(x, h))$, and $q' = (h/K_0)^D$.

Lemma 28. *There exists $h_0(\rho, \beta, f_{min}, f_{max})$ such that, if $h \leq h_0$, for every x satisfying $\mathrm{d}(x, M) \leq h/\sqrt{2}$, we have*

- $\mathcal{B}(\pi_M(x), h/(2\sqrt{2})) \cap M \subset \mathcal{B}(x, h) \cap M \subset \mathcal{B}(\pi_M(x), 2h) \cap M$,
- $c_d f_{min} h^d \leq q(x) \leq C_d f_{max} h^d$,
- $p(x) \leq 2\beta q(x)$.

Proof of Lemma 28. Set $h_1(\rho) = \rho/8$, and let x be such that $\mathrm{d}(x, M) \leq h/\sqrt{2}$, and $h \leq h_0$. According to Proposition 22 and Proposition 24, then $\mathcal{B}(\pi_M(x), r_h^-) \cap M \subset \mathcal{B}(x, h) \cap M \subset \mathcal{B}(\pi_M(x), r_h^+) \cap M$, with $r_h^+ = (1 + 4\alpha^2\Delta/\rho) r_h \leq 2r_h$ and $r_h^- = (1 - \alpha^2\Delta/\rho) r_h \geq r_h/2 \geq h/(2\sqrt{2})$.

The two remaining items of Lemma 28 proceed from what follows. According to Proposition 23, if $h \leq h_1$, then for any x such that $\mathrm{d}(x, M) \leq h/\sqrt{2}$, we have $c_d f_{min} h^d \leq q(x) \leq C_d f_{max} h^d$, for some constants c_d and C_d. Applying Proposition 23 again , there exists $h_2(f_{min}, d, D, \beta, \rho)$ such that if $h \leq h_1 \wedge h_0$, then for any x such that $\mathrm{d}(x, M) \leq h/\sqrt{2}$ we have $q' \leq C_{d, f_{min}, \rho} h$, along with $p(x) \leq 2\beta q(x)$. Taking $h_0 = h_1 \wedge h_2$ leads to the result. □

Assume that $h \leq h_0$. For convenience assume that $\pi_M X_1 = 0$ and that $T_0 M$ is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. Let m_1 denote $\mathbb{E}^{U=1}(X)$. Then m_1 may be decomposed as $m_1 = m_1^\top + m_1^\perp$, where $m_1^\top$ denotes the orthogonal projection of m_1 onto $T_0 M$.

Lemma 29. *Assume that $\mathrm{d}(X_1, M) \leq h/\sqrt{2}$ and $h \leq h_0$. Then*

$$\begin{array}{rcll} \|m_1^\perp\| & \leq & C_{d, f_{min}} \frac{h^2}{\rho} & \text{if } \beta = 1, \\ \|m_1^\perp\| & \leq & C_{d, f_{min}, \rho} h^2 & \text{if } \beta < 1. \end{array}$$

Proof of Lemma 29.

$$\begin{aligned} \|m_1^\perp\| &\leq \mathbb{E}^{U=1} \|X^\perp\| \\ &\leq \frac{\beta q(X_1)}{p(X_1)} \frac{(\alpha r_h^+)^2}{2\rho} + \frac{(1-\beta) q'}{p(X_1)} 2h, \end{aligned}$$

according to Proposition 22. The case $\beta = 1$ follows straightforwardly. Noting that $q' \leq C_{d, f_{min}, \rho} h q(X_1)$ in the case $\beta < 1$ gives the second inequality. □

The local covariance matrix may be decomposed as follows.

$$\begin{aligned} \hat{\Sigma}_1 &= \frac{1}{N_1} \sum_{i=2}^{n} (X_i - m_1)(X_i - m_1)^t U_i - \left(\bar{X}_1 - m_1\right) \left(\bar{X}_1 - m_1\right)^t \\ &= \hat{O}_1 + \hat{R}_1. \end{aligned} \tag{2}$$

The following Lemma gives the shape of the expectation of $\hat{O}_1$.

Lemma 30.

$$\mathbb{E}^{X_1, U_{2:n}} \hat{O}_1 = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & 0 \end{array}\right) + Q(X_1, h),$$

with

$$\begin{array}{rcll} \|Q(X_1, h)\|_{\mathcal{F}} & \leq & C_d \frac{h^3}{\rho} & \text{if } \beta = 1, \\ \|Q(X_1, h)\|_{\mathcal{F}} & \leq & C_{d, f_{min}, \rho} h^3 & \text{if } \beta < 1. \end{array}$$

and

$$\lambda_{min}(A) \geq c_{d,f_{min},f_{max}} h^2.$$

Proof of Lemma 30. First we note that, conditionnaly on the U_i's and X_1 the X_i's are independent for $i \geq 2$. Then, according to Proposition 21 combined with Proposition 22 and Proposition 23, we may write, for $h \leq h_0$ and y in $\mathcal{B}(X_1, h) \cap M$,

$$y = rv + R(r, v),$$

in local polar coordinates. According to Lemma 28, we have $\mathcal{B}(\pi_M(x), h/(2\sqrt{2})) \cap M \subset \mathcal{B}(x, h) \cap M \subset \mathcal{B}(\pi_M(x), 2h) \cap M$. It follows that

$$\begin{aligned}
&\mathbb{E}^{X_1, U_{2:n}} \hat{O}_1 \\
&\qquad = \beta \int_{\mathcal{B}(X_1,h)\cap M} (rv + R(r,v) - m_1)(rv + R(r,v) - m_1)^t \frac{f(r,v)J(r,v)}{p(X_1)} dr dv \\
&\qquad\qquad + \frac{1-\beta}{p(X_1)} \int_{\mathcal{B}(X_1,h)} (x - m_1)(x - m_1)^t \frac{d\lambda_D(x)}{\lambda_D(\mathcal{B}(0, K_0))} \\
&\qquad = \beta \int_{\mathcal{B}(X_1,h)\cap M} (rv - m_1^\top)(rv - m_1^\top)^t \frac{f(r,v)J(r,v)}{p(X_1)} dr dv + Q(X_1, h),
\end{aligned}$$

where

$$\begin{array}{rcll}
\|Q(X_1,h)\|_{\mathcal{F}} & \leq & C_d \frac{h^3}{\rho} & \text{if } \beta = 1, \\
\|Q(X_1,h)\|_{\mathcal{F}} & \leq & C_{d,f_{min},\rho} h^3 & \text{if } \beta < 1.
\end{array}$$

On the other hand, it is easy to see that

$$\beta \int_{\mathcal{B}(X_1,h)\cap M} (rv - m_1^\top)(rv - m_1^\top)^t J(r,v) \frac{f(r,v)}{p(X_1)} dr dv = \left(\begin{array}{c|c} A & 0 \\ \hline 0 & 0 \end{array}\right).$$

Let u be a unit vector in $T_0 M$. Then

$$\begin{aligned}
\langle Au, u \rangle &= \beta \int_{\mathcal{B}(X_1,h)\cap M} r^2 \langle u, rv - m_1 \rangle^2 J(r,v) \frac{f(r,v)}{p(X_1)} dr dv \\
&\geq \beta \frac{\sigma_{d-1} f_{min} c_{vol}^-(d)}{p(X_1)} \int_{r=0}^{r_h^-} \langle u, rv - m_1 \rangle^2 r^{d-1} dr dv \\
&\geq \beta \frac{f_{min} c_{vol}^-(d) \sigma_{d-1}}{d(d+2)p(X_1)} r_h^{-d+2} \\
&\geq c_{d,f_{min}} \frac{\beta}{p(X_1)} h^{d+2} \\
&\geq \frac{c_{d,f_{min}}}{q(X_1)} h^{d+2} \\
&\geq c_{d,f_{min},f_{max}} h^2.
\end{aligned}$$

where $p(X_1) \leq 2\beta q(X_1)$ has been used to derive the fourth inequality. □

Now we bound the deviations of $\hat{O}_1$ from its expectations. For any $D \times D$ matrix A we denote by A_{11}, A_{12}, A_{21} and A_{22} the $d \times d$, $(D-d) \times d$, $d \times (D-d)$ and $(D-d) \times (D-d)$ submatrices such that

$$A = \left(\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array} \right)$$

For r, s in $\{1,2\}$, denote by $Z_{1,rs}$ the quantity $\|\hat{O}_{1,rs} - \mathbb{E}^{X_1,U_{2:n}}\hat{O}_{1,rs}\|$.

Lemma 31. *Suppose that* $\mathrm{d}(X_1, M) \leq h_0$. *Then, for every* $x > 0$, *we have*

$$\mathbb{P}^{X_1}\left(Z_{r,s} \geq \frac{6\sigma_{r,s}}{\sqrt{(n-1)p(X_1)}}(1+\sqrt{x}) + \frac{2b_{rs}x}{(n-1)p(X_1)} \right) \leq e^{-\frac{p(X_1)(n-1)}{8}} + e^{-x},$$

with

$$\begin{array}{rcll} b_{r,s} & \leq & 16h^2\left(\frac{h}{2\rho}\right)^{r\vee s-1} & \text{if } \beta = 1 \\ b_{r,s} & \leq & 4h^2 & \text{if } \beta < 1, \end{array}$$

and

$$\begin{array}{rcll} \sigma_{r,s}^2 & \leq & b_{r,s}^2 & \text{if } \beta = 1 \\ \sigma_{r,s}^2 & \leq & C_{d,f_{min},\rho}h^{3+r\vee s} & \text{if } \beta < 1. \end{array}$$

Proof of Lemma 31. To control the deviations of $\hat{O}_1$ from its expectation, we will make use of the version of Talagrand's inequality that can be found in [5], and is recalled below.

Theorem 32 (Talagrand-Bousquet's inequality)**.** *Let* $Y_1, \ldots, Y_n$ *be independent copies of* Y, *and let* $\mathcal{F}$ *be a set of real-valued functions, such that, for all* f *in* $\mathcal{F}$, $\|f\|_\infty \leq b$ *and* $Var(f(Y)) \leq \sigma^2$. *Denote by* Z *the quantity*

$$Z = \sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}(f(Y)) \right|.$$

Then, for every $x > 0$, *we have, with probability larger than* $1 - e^{-x}$

$$Z \leq \mathbb{E}(Z) + \sqrt{2x\left(\frac{\sigma^2}{n} + \frac{4b\mathbb{E}(Z)}{n}\right)} + \frac{2bx}{3n}.$$

Denote by $b_{r,s}$ the quantity

$$b_{r,s} = \left\| \left((X - m_1)(X - m_1)^t U \right)_{rs} \right\|_{\mathcal{F},\infty}$$

We note that, if $\beta = 1$,

$$b_{r,s} \leq \frac{(\alpha r_h^+)^{r+s}}{(2\rho)^{r+s-2}} \leq 16h^2\left(\frac{h}{2\rho}\right)^{r\vee s-1},$$

according to Proposition 22 and Lemma 28. On the other hand, if $\beta < 1$, we only have

$$b_{r,s} \leq 4h^2.$$

Denote by $\sigma_{r,s}^2$ the quantity

$$\mathbb{E}^{U=1}\| \left((X - m_1)(X - m_1)_{r,s}^t \right) \|^2.$$

Then we get, if $\beta = 1$,

$$\sigma_{r,s}^2 \leq b_{r,s}^2 = 256h^4 \left(\frac{h}{2\rho}\right)^{2(r\vee s-1)}.$$

If $\beta < 1$, we may write

$$\sigma_{r,s}^2 \leq C_{d,f_{min},\rho} h^{3+r\vee s}.$$

Remarking that

$$Z_{1,rs} = \sup_{T/\|T\|_{\mathcal{F}}\leq 1} \left\langle T, \frac{1}{N_1}\sum_{i|U_i=1}^{n} (A_{rs}(X_i) - \mathbb{E}^{U=1}(A_{rs}(X)) \right\rangle,$$

with $A_{r,s}(x) = \left((x-m_1)(x-m_1)_{r,s}^t\right)$, and $\mathbb{E}^{X_1,U_{2:n}} Z_{r,s} \leq \sqrt{\frac{\sigma_{r,s}^2}{N_1}}$, a direct application of Theorem 32 yields

$$\mathbb{P}^{X_1,U_{2:n}}\left(Z_{r,s} \geq \frac{3\sigma_{r,s}}{\sqrt{N_1}}(1+\sqrt{x}) + \frac{2b_{rs}x}{N_1}\right) \leq e^{-x}.$$

According to Bernstein's inequality, (see, e.g., Proposition 2.9 in [23]), we may write

$$\mathbb{P}^{X_1}\left(N_1 \leq \frac{(n-1)p(X_1)}{4}\right) \leq e^{-\frac{p(X_1)(n-1)}{8}}.$$

Hence we have

$$\mathbb{P}^{X_1}\left(Z_{r,s} \geq \frac{6\sigma_{r,s}}{\sqrt{(n-1)p(X_1)}}(1+\sqrt{x}) + \frac{2b_{rs}x}{(n-1)p(X_1)}\right) \leq e^{-\frac{p(X_1)(n-1)}{8}} + e^{-x}.$$

□

Now we turn to the second term in (2). The following Lemma gives a bound on $\|\hat{R}_{1,rs}\|_{\mathcal{F}}$.

Lemma 33. *If* $\mathrm{d}(X_1, M) \leq h_0$, *we have*

$$\mathbb{P}^{X_1}\left(\left\|(\bar{X}_1 - m_1)^{\top}\right\|_{\mathcal{F}} \geq \frac{6\sigma^{\top}}{\sqrt{(n-1)p(X_1)}}(1+\sqrt{x}) + \frac{2b^{\top}x}{(n-1)p(X_1)}\right) \leq e^{-\frac{p(X_1)(n-1)}{8}} + e^{-x},$$

with

$$\begin{array}{rcl} b^{\top} & \leq & 2h, \\ \sigma^{\top} & \leq & 2h. \end{array}$$

Moreover

$$\mathbb{P}^{X_1}\left(\left\|(\bar{X}_1 - m_1)^{\perp}\right\|_{\mathcal{F}} \geq \frac{6\sigma^{\perp}}{\sqrt{(n-1)p(X_1)}}(1+\sqrt{x}) + \frac{2b^{\perp}x}{(n-1)p(X_1)}\right) \leq e^{-\frac{p(X_1)(n-1)}{8}} + e^{-x},$$

with

$$\begin{array}{rcll} b^{\perp} & \leq & \frac{4h^2}{\rho} & \text{if } \beta = 1, \\ b^{\perp} & \leq & 2h & \text{if } \beta < 1, \end{array}$$

and

$$\begin{array}{rcll} \sigma^{\perp} & \leq & b^{\perp} & \text{if } \beta = 1, \\ \sigma^{2,\perp} & \leq & C_{d,f_{min},\rho} h^3 & \text{if } \beta < 1. \end{array}$$

Proof of Lemma 33. Lemma 33 follows from Theorem 32 combined with

$$\mathbb{P}^{X_1}\left(N_1 \leq \frac{(n-1)p(X_1)}{4}\right) \leq e^{-\frac{p(X_1)(n-1)}{8}},$$

as in the proof of Lemma 31. □

Assume that $\beta = 1$, set $x = (1 + 2/d)\log(n)$ in the previous results, and set the bandwidth $h = \left(K_{f_{min},f_{max},d}\frac{\log(n)}{n}\right)^{\frac{1}{d}}$, for K large enough so that

$$\mathbb{P}^{X_1}(\lambda_{min}(\hat{\Sigma}_{1,1}) \leq c_{d,f_{min},f_{max}}h^2) \leq 4\left(\frac{1}{n}\right)^{1+\frac{2}{d}},$$

and

$$\mathbb{P}^{X_1}\left(\|\hat{\Sigma}_{r,s}\|_{\mathcal{F}} \geq c_{d,f_{min},f_{max}}h^2\frac{h}{\rho}\right) \leq 4\left(\frac{1}{n}\right)^{1+\frac{2}{d}},$$

if $r \vee s = 2$. Then, we get

$$\mathbb{P}^{X_1}\left(\angle(\hat{T}_{X_1}\mathcal{M}, T_{X_1}\mathcal{M}) \geq \pi d\frac{h}{\rho}\right) \leq 8\left(\frac{1}{n}\right)^{1+\frac{2}{d}}.$$

Since, in the noiseless case, $\mathrm{d}(X_1, \mathcal{M}) \leq h/\sqrt{2}$ almost surely, we may write, taking expectation with respect to the law of X_1,

$$\mathbb{P}\left(\angle(\hat{T}_{X_1}\mathcal{M}, T_{X_1}\mathcal{M}) \geq \pi d\frac{h}{\rho}\right) \leq 8\left(\frac{1}{n}\right)^{1+\frac{2}{d}}.$$

Then, a union bound yields

$$\mathbb{P}\left(\sup_{j=1,\dots,n}\angle(\hat{T}_{X_j}\mathcal{M}, T_{X_j}\mathcal{M}) \geq \pi d\frac{h}{\rho}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d}},$$

hence Proposition 13 is proved.

F Convergence Rate for the Denoising Procedure

We are now able to prove the efficiency of the denoising procedure, namely Proposition 16.

Proof of Proposition 16. We recall that the slab $S_T(x)$ is the set of points y such that $\|\pi_T(y-x)\| \leq k_1 h$ and $\|\pi_{T^\perp}(y-x)\| \leq k_2 h^2$, where π_T denotes the orthogonal projection onto T. The following Lemma ensures that slabs that are centered on points far away from M of at least h^2 would catch only noise distribution, provided that their tangent space is well-approximated.

Lemma 34. *Let $K > 0$ be fixed. There exists constants $k_1(\rho, K, d, D)$ and $k_2(d, D)$ such that, for every $h \leq h_0 \wedge 1$ and x such that $\mathrm{d}(x, M) \geq h^2$, we have*

$$\angle\left(T_{\pi(x)}M, T\right) \leq Kh \quad \Rightarrow \quad S_T(x) \cap \mathcal{M} = \emptyset.$$

Furthermore, if x and y are in M, then there exists a constant k_3 such that

$$\|x-y\| \leq k_3 h \quad \Rightarrow \quad y \in S_{T'}(x).$$

Possible values for k_1 and k_2 are, respectively, $\frac{7\rho}{16d} \wedge \frac{7}{32K\sqrt{d}} \wedge 1$ and $\frac{7}{18(D-d)}$. Then k_3 may be chosen as $\frac{k_2}{2K} \wedge \frac{k_1}{2} \wedge \sqrt{\rho k_1} \wedge \sqrt{\rho k_2}$.

The proof of Lemma 34, mentioned in [18], follows from elementary geometry, combined with the definition of the reach and Proposition 22.

According to the results of Section E, choosing $h = \left(C_{f_{min},f_{max},d} \frac{\log(n)}{\beta(n-1)}\right)^{\frac{1}{d+1}}$ yields, for n large enough,

$$\mathbb{P}\left(\left\{\exists j | \mathrm{d}(X_j, M) \leq h/\sqrt{2} \quad \text{and} \quad \angle(\hat{T}_{X_j}\mathcal{M}, T_{X_j}\mathcal{M}) \geq c_{d,f_{min},f_{max},\rho} h\right\}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d+1}}.$$

Hence, choosing $K = c_{d,f_{min},f_{max},\rho}$ in Lemma 34, and $k_1 = k_1(\rho, K, d, D) \wedge 1/(2\sqrt{2d})$, ensures that, for n large enough,

$$\mathbb{P}\left(\left\{\exists j | \mathrm{d}(X_j, M) \geq h^2 \quad \text{and} \quad S_{\hat{T}_j}(X_j) \neq \emptyset\right\}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d+1}},$$

along with

$$\mathbb{P}\left(\left\{\exists j | X_j \in M \quad \text{and} \quad \mathcal{B}(X_j, k_3 h) \cap M \subsetneq S_{T_j}(X_j)\right\}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d+1}}.$$

It remains to give bounds on the number of points that are falling into the slabs. To this aim, we use the same Lemma as in [18], stated below.

Lemma 35 (Lemma 7 in [18])**.** *Let $\mathcal{A}$ denote the set of hyperrectangles in $\mathbb{R}^D$. Then, with probability larger than $1 - \left(\frac{1}{n}\right)^{2D}$, for every A in $\mathcal{A}$, we have*

$$\begin{cases} P_n(A) & \leq & P(A) + \frac{16D\log(2n)}{n} + \sqrt{\frac{16D\log(2n)}{n}}\sqrt{P(A)} \\ P_n(A) & \geq & P(A) - \sqrt{\frac{16D\log(2n)}{n}}\sqrt{P(A)}. \end{cases}$$

If $S_T(x) \cap M = \emptyset$, then $P(S_T(x)) \leq \left(\frac{h}{K_0}\right)^{d+1}$. On the other hand, if $\mathcal{B}(x, k_3 h) \cap M \subset S_T(x)$, then $P(S_T(x)) \geq C_{f_{min},f_{max},d} \kappa_3^d \frac{\log(n-1)}{(n-1)h}$. Applying Lemma 35, we derive that, for any fixed $m > 1$ and for n large enough

$$\mathbb{P}\left(\left\{\mathrm{d}(X_j, M) \geq h^2 \quad \text{and} \quad \sqrt{h}\frac{n}{\log(n)} P_n(S_{\hat{T}_j}(X_j)) \geq 1/m\right\}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d+1}} + \left(\frac{1}{n}\right)^{2D},$$

and

$$\mathbb{P}\left(\left\{X_j \in M \quad \text{and} \quad \sqrt{h}\frac{n}{\log(n)} P_n(S_{\hat{T}_j}(X_j)) \leq m\right\}\right) \leq 8\left(\frac{1}{n}\right)^{\frac{2}{d+1}} + \left(\frac{1}{n}\right)^{2D}.$$

Hence the result of Proposition 16.

□

G Proof of the Main Reconstruction Results

We now prove main results Theorem 6 in the noise-free model, and Theorem 7 in the clutter noise model.

G.1 Noise-Free Case

Proof of Corollary 14. Let $Q \in \mathcal{G}_{D,d,f_{min},f_{max},\rho}$. Write $\varepsilon = c_{d,f_{min},f_{max}} \left(\frac{\log n}{n-1}\right)^{1/d}$. Consider the event A defined by

$$A = \left\{ \max_{X_j \in \mathbb{Y}_n} \angle(T_{X_j} M, \hat{T}_j) \leq C_{d,f_{min},f_{max}} \frac{\varepsilon}{\rho} \right\} \cap \{ \mathrm{d_H}(M, \mathbb{Y}_n) \leq 2\varepsilon \} \cap \{ \mathbb{Y}_n \text{ is } \varepsilon\text{-sparse} \}.$$

The construction of $\mathbb{Y}_n$ from the farthest point sampling algorithm and a straightforward combination of Proposition 13 and Lemma 25 yields, for n large enough,

$$\mathbb{P}_Q(A^c) \leq 2 \left(\frac{\log n}{n}\right)^{1/d}.$$

□

Proof of Theorem 6. Following the above notation, we observe that on the event A, Theorem 12 holds with parameters $\tau_1 = 0$ and $\tau_2 = C_{d,f_{min},f_{max}}/\rho$, so that the first part of Theorem 6 is proved. Furthermore, for n large enough,

$$\begin{aligned}
\mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right)\right] &\leq \mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right) \mathbb{1}_A\right] + \mathbb{E}_Q\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC}}\right) \mathbb{1}_{A^c}\right] \\
&\leq (C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_2)\varepsilon^2 + \mathrm{diam}(M)\mathbb{P}_Q(A^c) \\
&\leq C'_{d,f_{min},f_{max},\rho}\varepsilon^2,
\end{aligned}$$

where for the last line we used the diameter bound of Proposition 2. □

G.2 Clutter Noise Case

Proof of Corollary 17. Let $P \in \mathcal{O}_{D,d,f_{min},f_{max},\rho,\beta}$. For n large enough, let us denote $\varepsilon = c_{d,f_{min},f_{max},\rho} \left(\frac{\log n}{\beta n}\right)^{1/(d+1)}$. Consider the event

$$\begin{aligned}
A^+ = \left\{ \max_{X_j \in \mathbb{Y}_n^+} \angle(T_{\pi(X_j)} M, \hat{T}_j) \leq C_{d,f_{min},f_{max},\rho}\varepsilon \right\} \cap \left\{ \max_{X_j \in \mathbb{Y}_n^+} \mathrm{d}(X_j, M) \leq \varepsilon^2 \right\} \\
\cap \left\{ \mathrm{d_H}(M, \mathbb{Y}_n^+) \leq 2\varepsilon \right\} \cap \left\{ \mathbb{Y}_n^+ \text{ is } \varepsilon\text{-sparse} \right\}.
\end{aligned}$$

From Lemma 25, Proposition 16 and the construction of $\mathbb{Y}_n^+$ with the farthest point sampling algorithm, it is clear that for n large enough,

$$\mathbb{P}_P(A^{+c}) \leq 8\left(\frac{1}{n}\right)^{2/d} + \left(\frac{1}{n}\right)^{2D} + \left(\frac{1}{n}\right)^{2/d} \leq 10\left(\frac{1}{n}\right)^{2/d}.$$

□

Proof of Theorem 7. Following the above notation, we observe that on the event A^+, Theorem 12 holds with parameters $\tau_1 = 1$ and $\tau_2 = C_{d,f_{min},f_{max},\rho}$, so that the first part of Theorem 7 is proved. Furthermore, for n large enough,

$$\begin{aligned}
\mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC+}}\right)\right] &\leq \mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC+}}\right)\mathbb{1}_{A^+}\right] + \mathbb{E}_P\left[\mathrm{d_H}\left(M, \hat{M}_{\texttt{TDC+}}\right)\mathbb{1}_{A^{+c}}\right] \\
&\leq (C_{d,\rho'_{\rho,\tau_1,\tau_2}} + \tau_1 + \tau_2)\varepsilon^2 + 2K_0\mathbb{P}_P(A^{+c}) \\
&\leq C'_{d,f_{min},f_{max},\rho,\beta}\varepsilon^2,
\end{aligned}$$

where for the second line we used Theorem 12 and the fact that $M \cup \hat{M}_{\texttt{TDC+}} \subset \mathcal{B}_0$, a ball of radius K_0. □