Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari, Clément Levrard

Noname manuscript No.
(will be inserted by the editor)

Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction

Eddie Aamari · **Clément Levrard**

Received: date / Accepted: date

Abstract We consider the problem of optimality in manifold reconstruction. A random sample $\mathbb{X}_n = \{X_1, \ldots, X_n\} \subset \mathbb{R}^D$ composed of points close to a d-dimensional submanifold M, with or without outliers drawn in the ambient space, is observed. Based on the Tangential Delaunay Complex [4], we construct an estimator $\hat{M}$ that is ambient isotopic and Hausdorff-close to M with high probability. The estimator $\hat{M}$ is built from existing algorithms. In a model with additive noise of small amplitude, we show that this estimator is asymptotically minimax optimal for the Hausdorff distance over a class of submanifolds satisfying a reach constraint. 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 decluttering procedure and a tangent space estimator both based on local principal component analysis (PCA) are studied.

Keywords Manifold reconstruction, Minimax optimality, Tangential Delaunay Complexes, Decluttering, Tangent space estimation

1 Introduction

Throughout many fields of applied science, data in $\mathbb{R}^D$ can naturally be modeled as lying on a d-dimensional submanifold M. 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 $\mathbb{X}_n = \{X_1, \ldots, X_n\}$. 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.

Eddie Aamari
Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, France
DataShape, Inria Saclay, Palaiseau, France
E-mail: eddie.aamari@inria.fr

Clément Levrard
Université Paris Diderot, Paris, France
E-mail: levrard@math.univ-paris-diderot.fr

The computational geometry community has now been active on manifold reconstruction for many years, mainly in deterministic frameworks. In dimension 3, [17] provides a survey of the state of the art. In higher dimension, the employed methods rely on variants of the ambient Delaunay triangulation [12, 4]. The geometric and topological guarantees are derived under the assumption that the point cloud — fixed and nonrandom — 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 going to infinity. To derive these bounds, a broad variety 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 [29,20]. These rates depend crudely on the regularity of the boundary of the convex set 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 bounds [14,24] yield rich classes of sets with good geometric properties. In particular, the reach, as introduced by Federer [22], appears to be a key regularity and scale parameter [10,24,28].

This paper mainly follows up the two articles [4,24], both dealing with the case of a d-dimensional submanifold $M \subset \mathbb{R}^D$ with a reach regularity condition and where the dimension d is known.

On one hand, [4] focuses on a deterministic analysis and proposes a provably faithful reconstruction. The authors introduce a weighted Delaunay triangulation restricted to tangent spaces, the so-called Tangential Delaunay Complex. This paper gives a reconstruction up to ambient isotopy with 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 [4] do not resist ambient perturbations. Indeed, it heavily relies on the knowledge of the tangent spaces at each point and on the absence of noise.

On the other hand, [24] takes a statistical approach in a model possibly corrupted by additive noise, or containing outlier points. The authors derive an estimator that is proved to be minimax optimal for the Hausdorff distance d_H. Roughly speaking, minimax optimality of the proposed estimator means that it performs best in the worst possible case up to numerical constants, when the sample size n is large enough. Although theoretically optimal, the proposed estimator appears to be intractable in practice. At last, [28] proposes a manifold estimator based on local linear patches that is tractable but fails to achieve the optimal rates.

Contribution

Our main contributions (Theorems 7, 8 and 9) make a two-way link between the approaches of [4] and [24].

From a geometric perspective, Theorem 7 shows that the Tangential Delaunay Complex of [4] can be combined with local PCA to provide a manifold estimator that is optimal in the sense of [24]. This remains possible even if data is corrupted with additive noise of small amplitude. Also, Theorems 8 and 9 show that, if outlier points are present (clutter noise), the Tangential Delaunay Complex of [4] still yields the optimal rates of [24], at the price of an additional decluttering procedure.

From a statistical point of view, our results show that the optimal rates described in [24] can be achieved by a tractable estimator $\hat{M}$ that (1) is a simplicial complex of which vertices are the data points, and (2) such that $\hat{M}$ is ambient isotopic to M with high probability.

In the process, a stability result for the Tangential Delaunay Complex (Theorem 14) is proved. Let us point out that this stability is derived using an interpolation result (Theorem 11) which is interesting in its own right. Theorem 11 states that if a point cloud $\mathscr{P}$ lies close to a submanifold M, and that estimated

tangent spaces at each sample point are given, then there is a submanifold M' (ambient isotopic, and close to M for the Hausdorff distance) that interpolates $\mathscr{P}$, with T_pM' agreeing with the estimated tangent spaces at each point $p \in \mathscr{P}$. Moreover, the construction can be done so that the reach of M' is bounded in terms of the reach of M, provided that $\mathscr{P}$ is sparse, points of $\mathscr{P}$ lie close to M, and error on the estimated tangent spaces is small. Hence, Theorem 11 essentially allows to consider a noisy sample with estimated tangent spaces as an exact sample with exact tangent spaces on a proxy submanifold. This approach can provide stability for any algorithm that takes point cloud and tangent spaces as input, such as the so-called *cocone* complex [12].

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 2.1) together with minimax optimality (Section 2.2). Then, the main results are stated (Section 2.3).

Two models are studied, one where $\mathbb{X}_n$ is corrupted with additive noise, and one where $\mathbb{X}_n$ contains outliers. We build a simplicial complex $\hat{M}_{\mathtt{TDC}}(\mathbb{X}_n)$ ambient isotopic to M and we derive the rate of approximation for the Hausdorff distance $d_H(M, \hat{M}_{\mathtt{TDC}})$, with bounds holding uniformly over a class of submanifolds satisfying a reach regularity condition. The derived rate of convergence is minimax optimal if the amplitude σ of the additive noise is small. With outliers, similar estimators $\hat{M}_{\mathtt{TDC}\delta}$ and $\hat{M}_{\mathtt{TDC}+}$ are built. $\hat{M}_{\mathtt{TDC}}$, $\hat{M}_{\mathtt{TDC}\delta}$ 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. 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). For ease of exposition, all the proofs are placed in the appendix.

Notation

In what follows, we consider a compact 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 $\mathscr{B}(p,r)$ denote the closed Euclidean ball of radius $r > 0$ centered at p. We let $\wedge$ and $\vee$ denote respectively the minimum and the maximum of real numbers. As introduced in [22], 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 $d(\cdot, M)$ the distance to M, the *medial axis* of M $\mathrm{med}(M) = \{x \in \mathbb{R}^D | \exists a \neq b \in M, \|x-a\| = \|x-b\| = d(x,M)\}$ is the set of points which have at least two nearest neighbors on M. Then, $\mathrm{reach}(M) = \inf_{p \in M} d(p, \mathrm{med}(M))$. 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 $d_a\Phi$ and $d_a^2\Phi$ 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 $\|A\|_{\mathscr{F}} = \sqrt{\mathrm{trace}\,(A^tA)}$ 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 the same dimension is measured by the sine $\angle(U,V) = \max_{u \in U} \max_{v' \in V^\perp} \frac{\langle u, v' \rangle}{\|u\| \, \|v'\|} = \|\pi_U - \pi_V\|_{\mathrm{op}}$ of their largest principal angle. The Hausdorff distance between two compact subsets K, K' of $\mathbb{R}^D$ is denoted by $d_H(K,K') = \sup_{x \in \mathbb{R}^D} |d(x,K) - d(x,K')|$. Finally, 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's sake, c_α and K_α may also be used when several constants are involved.

2 Minimax Risk and Main Results

2.1 Statistical Model

Let us describe the general statistical setting we will use to define optimality for manifold reconstruction. A *statistical model* $\mathscr{D}$ is a set of probability distributions on $\mathbb{R}^D$. In any statistical experiment, $\mathscr{D}$ is fixed and known. We observe an independent and identically distributed sample of size n (or i.i.d. n-sample) $\mathbb{X}_n = \{X_1, \dots, X_n\}$ drawn according to some unknown distribution $P \in \mathscr{D}$. 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 $\mathscr{D}$ by describing those of interest in this paper.

Let $\mathscr{M}_{D,d,\rho}$ be the set of all compact 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 [14]. From a reconstruction point of view, ρ gives a minimal feature size on M, and then a minimal scale for geometric information. Every $M \in \mathscr{M}_{D,d,\rho}$ inherits a measure induced by the d-dimensional Hausdorff measure on $\mathbb{R}^D \supset M$. We denote this induced measure by ν_M. Beyond the geometric restrictions induced by the lower bound ρ on the reach, it also requires the natural measure ν_M to behave like a d-dimensional measure, up to uniform constants. Denote by $\mathscr{U}_M(f_{min}, f_{max})$ the set of probability distributions Q having a density f with respect to ν_M such that $0 < f_{min} \leq f(x) \leq f_{max} < \infty$ for all $x \in M$. In particular, notice that such distributions $Q \in \mathscr{U}_M(f_{min}, f_{max})$ all have support M. Roughly speaking, when $Q \in \mathscr{U}_M(f_{min}, f_{max})$, points are drawn almost uniformly on M. This is to ensure that the sample visits all the areas of M with high probability. The noise-free model $\mathscr{G}_{D,d,f_{min},f_{max},\rho}$ 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 (Noise-free model) $\mathscr{G}_{D,d,f_{min},f_{max},\rho} = \bigcup_{M \in \mathscr{M}_{D,d,\rho}} \mathscr{U}_M(f_{min}, f_{max})$.

Notice that 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 $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho}$ with associated M,*

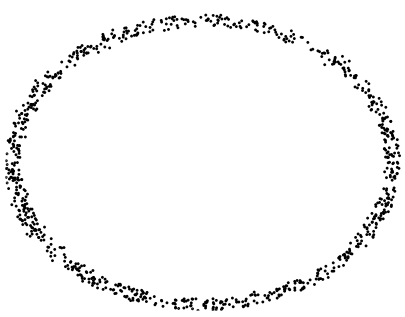
$$\mathrm{diam}(M) \leq \frac{C_d}{\rho^{d-1} f_{min}} =: K_{d,f_{min},\rho}.$$

Observed random variables with distribution belonging to the noise-free model $\mathscr{G}_{D,d,f_{min},f_{max},\rho}$ lie exactly on the submanifold of interest M. A more realistic model should allow some measurement error, as illustrated by Figure 1a. We formalize this idea with the following additive noise model.

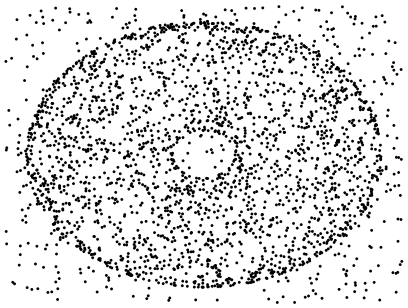
Definition 3 (Additive noise model) *For $\sigma < \rho$, we let $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$ denote the set of distributions of random variables $X = Y + Z$, where Y has distribution $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho}$, and $\|Z\| \leq \sigma$ almost surely.*

Let us emphasize that we do not require Y and Z to be independent, nor Z to be orthogonal to $T_Y M$, as done for the "perpendicular" noise model of [30, 24]. This model is also slightly more general than the one considered in [28]. Notice that the noise-free model can be thought of as a particular instance of the additive noise model, since $\mathscr{G}_{D,d,f_{min},f_{max},\rho} = \mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma=0}$.

Eventually, we may include distributions contaminated with outliers uniformly drawn in a ball $\mathscr{B}_0$ containing M, as illustrated in Figure 1b. Up to translation, we can always assume that $M \ni 0$. To avoid boundary effects, $\mathscr{B}_0$ will be taken to contain M amply, 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 $\mathscr{B}_0 = \mathscr{B}(0, K_0)$, where $K_0 = K_{d,f_{min},\rho} + \rho$. Notice that the larger the radius of $\mathscr{B}_0$, the easier to label the outlier points since they should be very far away from each other.



(a) Circle with noise: $d = 1$, $D = 2$, $\sigma > 0$.



(b) Torus with outliers: $d = 2$, $D = 3$, $\beta < 1$.

Fig. 1: Point clouds $\mathbb{X}_n$ drawn from distributions in $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$ (left) and $\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$ (right).

Definition 4 (Model with outliers/Clutter noise model) *For $0 < f_{min} \leq f_{max} < \infty$, $0 < \beta \leq 1$, and $\rho > 0$, we define $\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$ to be the set of mixture distributions*

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

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

Alternatively, a random variable X with distribution $P \in \mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$ 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 $\mathscr{G}_{D,d,f_{min},f_{max},\rho}$ and X'' has a uniform distribution over $\mathscr{B}_0$, and such that V, X', X'' are independent. In particular for $\beta = 1$, $\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta=1} = \mathscr{G}_{D,d,f_{min},f_{max},\rho}$.

2.2 Minimax Risk

For a probability measure $P \in \mathscr{D}$, 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 $\mathscr{D}$. For $n \geq 0$,

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

where the infimum is taken over all the estimators $\hat{M} = \hat{M}(X_1, \ldots, X_n)$ computed over an n-sample. $R_n(\mathscr{D})$ is the best risk that an estimator based on an n-sample can achieve uniformly over the class $\mathscr{D}$. It is clear from the definition that if $\mathscr{D}' \subset \mathscr{D}$ then $R_n(\mathscr{D}') \leq R_n(\mathscr{D})$. It follows the intuition that the broader the class of considered manifolds, the more difficult it is to estimate them uniformly well. Studying $R_n(\mathscr{D})$ for a fixed n is a difficult task that can rarely be carried out. We will focus on the semi-asymptotic behavior of this risk. As $R_n(\mathscr{D})$ cannot be surpassed, its rate of convergence to 0 as $n \to \infty$ may be seen as the best rate of approximation that an estimator can achieve. 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 5 *An estimator $\hat{M}$ is said to be (asymptotically) minimax optimal over $\mathscr{D}$ if*

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

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 provide an estimator and to study its quality uniformly on $\mathscr{D}$. On the other hand, bounding R_n from below amounts to study the worst possible case in $\mathscr{D}$. This part is usually achieved with standard Bayesian techniques [27]. For the models considered in the present paper, the rates were given in [24,26].

Theorem 6 (Theorem 3 of [26]) *We have,*

$$R_n\left(\mathscr{G}_{D,d,f_{min},f_{max},\rho}\right) \asymp \left(\frac{\log n}{n}\right)^{2/d}, \qquad \text{(Noise-free)}$$

and for $0<\beta\leq 1$ fixed,

$$R_n\left(\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}\right) \asymp \left(\frac{\log n}{\beta n}\right)^{2/d}. \qquad \text{(Clutter noise)}$$

Since the additive noise model $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$ has not yet been considered in the literature, the behavior of the associated minimax risk is not known. Beyond this theoretical result, an interesting question is to know whether these minimax rates can be achieved by a tractable algorithm. Indeed, that proposed in [24] especially rely on a minimization problem over the class of submanifolds $\mathscr{M}_{D,d,\rho}$, which is computationally costly. In addition, the proposed estimators are themselves submanifolds, which raises storage problems. Moreover, no guarantee is given on the topology of the estimators. Throughout the present paper, we will build estimators that address these issues.

2.3 Main Results

Let us start with the additive noise model $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$, that includes in particular the noise-free case $\sigma=0$. The estimator $\hat{M}_{\texttt{TDC}}$ is based on the Tangential Delaunay Complex (Section 3), with a tangent space estimation using a local PCA (Section 5).

Theorem 7 *$\hat{M}_{\texttt{TDC}}=\hat{M}_{\texttt{TDC}}(\mathbb{X}_n)$ is a simplicial complex with vertices included in $\mathbb{X}_n$ such that the following holds. There exists $\lambda_{d,f_{min},f_{max}}>0$ such that if $\sigma\leq\lambda\left(\frac{\log n}{n}\right)^{1/d}$ with $\lambda\leq\lambda_{d,f_{min},f_{max}}$, then*

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

Moreover, for n large enough,

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

It is interesting to note that the constants appearing in Theorem 7 do not depend on the ambient dimension D. Since $R_n\left(\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}\right) \geq R_n\left(\mathscr{G}_{D,d,f_{min},f_{max},\rho}\right)$, we obtain immediately from Theorem 7 that $\hat{M}_{\texttt{TDC}}$ achieves the minimax optimal rate $(\log n/n)^{2/d}$ over $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$ when $\sigma \leq c_{d,f_{min},f_{max}}(\log n/n)^{2/d}$. Note that the estimator of [28] achieves the rate $(\log n/n)^{2/(d+2)}$ when $\sigma \leq c_{d,f_{min},f_{max}}(\log n/n)^{2/(d+2)}$, so does the estimator of [25] for $\sigma < \rho$ if the noise is centered and perpendicular to the submanifold. As a consequence, $\hat{M}_{\texttt{TDC}}$ outperforms these two existing procedures whenever $\sigma \ll (\log n/n)^{2/(d+2)}$, with the additional feature of exact topology recovery. Still, for $\sigma \gg (\log n/n)^{1/d}$, $\hat{M}_{\texttt{TDC}}$ may perform poorly compared to [25]. This might be due to the fact that the vertices of $\hat{M}_{\texttt{TDC}}$ are sample points themselves, while for higher noise levels, a pre-process of the data based on local averaging could be more relevant.

In the model with outliers $\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$, with the same procedure used to derive Theorem 7 and an additional iterative preprocessing of the data based on local PCA to remove outliers (Section 5), we design an estimator of M that achieves a rate as close as wanted to the noise-free rate. Namely, for any positive $\delta < 1/(d(d+1))$, we build $\hat{M}_{\texttt{TDC}\delta}$ that satisfies the following similar statement.

Theorem 8 *$\hat{M}_{\texttt{TDC}\delta} = \hat{M}_{\texttt{TDC}\delta}(\mathbb{X}_n)$ is a simplicial complex with vertices included in $\mathbb{X}_n$ such that*

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

Moreover, for n large enough,

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

$\hat{M}_{\texttt{TDC}\delta}$ converges at the rate at least $(\log n/n)^{2/d-2\delta}$, which is not the minimax optimal rate according to Theorem 6, but that can be set as close as desired to it. To our knowledge, $\hat{M}_{\texttt{TDC}\delta}$ 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 8 do not depend on the ambient dimension D. The construction and computation of $\hat{M}_{\texttt{TDC}\delta}$ is the same as $\hat{M}_{\texttt{TDC}}$, with an extra pre-processing of the point cloud allowing to remove outliers. This decluttering procedure leads to compute, at each sample point, at most $\log(1/\delta)$ local PCA's, instead of a single one for $\hat{M}_{\texttt{TDC}}$.

From a theoretical point of view, there exists a (random) number of iterations of this decluttering process, from which an estimator $\hat{M}_{\texttt{TDC}+}$ can be built to satisfy the following.

Theorem 9 *$\hat{M}_{\texttt{TDC}+} = \hat{M}_{\texttt{TDC}+}(\mathbb{X}_n)$ is a simplicial complex of vertices contained in $\mathbb{X}_n$ such that*

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

Moreover, for n large enough,

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

$\hat{M}_{\texttt{TDC}+}$ may be thought of as a limit of $\hat{M}_{\texttt{TDC}\delta}$ when δ goes to 0. As it will be proved in Section 5, this limit will be reached for δ close enough to 0. Unfortunately this convergence threshold is also random, hence unknown.

The statistical analysis of the reconstruction 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 $\mathscr{P}$ be a finite subset of $\mathbb{R}^D$. In this section, we denote the point cloud $\mathscr{P}$ to emphasize the fact that it is considered nonrandom. For $\varepsilon, \delta > 0$, $\mathscr{P}$ is said to be ε-*dense* in M if $\sup_{x\in M} d(x,\mathscr{P}) \leq \varepsilon$, and δ-*sparse* if $d(p, \mathscr{P}\setminus\{p\}) \geq \delta$ for all $p \in \mathscr{P}$. A (δ,ε)-*net* (of M) is a δ-sparse and ε-dense point cloud.

3.1 Restricted Weighted Delaunay Triangulations

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

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

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

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

Note that for a constant weight assignment $\omega(p) \equiv \omega_0$, $\mathrm{Del}^\omega(\mathscr{P})$ is the usual Delaunay triangulation of $\mathscr{P}$. Under genericity assumptions on $\mathscr{P}$ and bounds on ω, $\mathrm{Del}^\omega(\mathscr{P})$ is an embedded triangulation with vertex set $\mathscr{P}$ [4]. The reconstruction method proposed in this paper is based on $\mathrm{Del}^\omega(\mathscr{P})$ for some weights ω to be chosen later. As it is a triangulation of the whole convex hull of $\mathscr{P}$ and fails to recover the geometric structure of M, we take restrictions of it in the following manner.

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

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

In particular, we define the *Tangential Delaunay Complex* $\mathrm{Del}^\omega(\mathscr{P},T)$ by taking $R = T = \{T_pM\}_{p\in\mathscr{P}}$, the family of tangent spaces taken at the points of $\mathscr{P} \subset M$ [4]. $\mathrm{Del}^\omega(\mathscr{P},T)$ is a pruned version of $\mathrm{Del}^\omega(\mathscr{P})$ 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 [4], computing $\mathrm{Del}^\omega(\mathscr{P},T)$ only requires to compute Delaunay triangulations in the tangent spaces that have dimension d. This reduces the computational complexity dependency on the ambient dimension $D > d$. The weight assignment ω gives degrees of freedom for the reconstruction. The extra degree of freedom ω permits to stabilize the triangulation and to remove the so-called *inconsistencies*, the points remaining fixed. For further details, see [5,4].

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 [4].

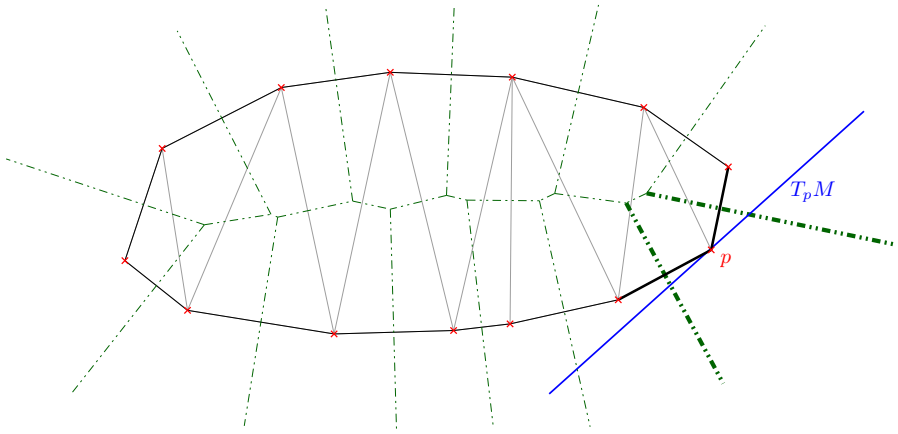


Fig. 2: Construction of $\mathrm{Del}^{\omega}(\mathscr{P},T)$ at p for $\omega \equiv 0$: p has three incident edges in the ambient Delaunay triangulation, but only two (bold) have dual Voronoi face intersecting T_pM.

Theorem 10 (Theorem 5.3 in [4]) *There exists $\varepsilon_d > 0$ such that for all $\varepsilon \leq \varepsilon_d \rho$ and all $M \in \mathscr{M}_{D,d,\rho}$, if $\mathscr{P} \subset M$ is an $(\varepsilon, 2\varepsilon)$-net, there exists a weight assignment $\omega_* = \omega_{*\mathscr{P},T}$ depending on $\mathscr{P}$ and $T = \{T_pM\}_{p\in\mathscr{P}}$ such that*

- $d_H(M, \mathrm{Del}^{\omega_*}(\mathscr{P},T)) \leq C_d\varepsilon^2/\rho$,
- *M and $\mathrm{Del}^{\omega_*}(\mathscr{P},T)$ are ambient isotopic.*

Computing $\mathrm{Del}^{\omega_*}(\mathscr{P},T)$ requires to determine the weight function $\omega_* = \omega_{*\mathscr{P},T}$. In [4], a greedy algorithm is designed for this purpose and has a time complexity $O\left(Dn^2 + D2^{O(d^2)}n\right)$.

Given an $(\varepsilon, 2\varepsilon)$-net $\mathscr{P}$ for ε small enough, $\mathrm{Del}^{\omega_*}(\mathscr{P},T)$ recovers M up to ambient isotopy and approximates it at the scale ε^2. The order of magnitude ε^2 with an input $\mathscr{P}$ of scale ε is remarkable. Another instance of this phenomenon is present in [13] 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 10 suffers two major imperfections. First, it requires the knowledge of the tangent spaces at each sample point — since $\omega_* = \omega_{*\mathscr{P},T}$ — and it is no longer usable if tangent spaces are only known up to some error. Second, the points are assumed to lie exactly on the submanifold M, and no noise is allowed. The analysis of $\mathrm{Del}^{\omega_*}(\mathscr{P},T)$ is sophisticated [4]. Rather than redo the whole study with milder assumptions, we tackle this question with an approximation theory approach (Theorem 11). Instead of studying if $\mathrm{Del}^{\omega_*}(\mathscr{P}',T')$ is stable when $\mathscr{P}'$ lies close to M and T' close to T, we examine what $\mathrm{Del}^{\omega_*}(\mathscr{P}',T')$ actually reconstructs, as detailed in Section 4.

3.3 On the Sparsity Assumption

In Theorem 10, $\mathscr{P}$ 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 $\mathrm{Del}^{\omega_*}(\mathscr{P},T)$ [5,4]. At this stage, we emphasize that the construction of a $(\varepsilon, 2\varepsilon)$-net can be carried out given an ε-dense sample. Given an ε-dense sample $\mathscr{P}$, the *farthest point sampling* algorithm prunes $\mathscr{P}$ and outputs an $(\varepsilon, 2\varepsilon)$-net $\mathscr{Q} \subset \mathscr{P}$ of M as follows. Initialize at $\mathscr{Q} = \{p_1\} \subset \mathscr{P}$, and while $\max_{p\in\mathscr{P}} d(p,\mathscr{Q}) > \varepsilon$, add to $\mathscr{Q}$ the farthest point to $\mathscr{Q}$ in $\mathscr{P}$, that is,

$\mathscr{Q} \leftarrow \mathscr{Q} \cup \{\underset{p\in\mathscr{P}}{\operatorname{argmax}}\, d(p,\mathscr{Q})\}$. The output $\mathscr{Q}$ is ε-sparse and satisfies $d_H(\mathscr{P},\mathscr{Q}) \leq \varepsilon$, so it is a $(\varepsilon, 2\varepsilon)$-net of M. Therefore, up to the multiplicative constant 2, sparsifying $\mathscr{P}$ 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 10 does not apply. To bypass this issue, we interpolate the data with another submanifold M' satisfying good properties, as stated in the following result.

Theorem 11 (Interpolation) *Let $M \in \mathscr{M}_{D,d,\rho}$. Let $\mathscr{P} = \{p_1,\dots,p_q\} \subset \mathbb{R}^D$ be a finite point cloud and $\tilde{T} = \{\tilde{T}_1,\dots,\tilde{T}_q\}$ 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*

- *$\mathscr{P}$ is δ-sparse: $\min_{i\neq j} \|p_j - p_i\| \geq \delta$,*
- *the p_j's are η-close to M: $\max_{1\leq j\leq q} d(p_j, M) \leq \eta$,*
- *$\max_{1\leq j\leq q} \angle(T_{\pi_M(p_j)}M, \tilde{T}_j) \leq \sin\theta$.*

Then, there exist a universal constant $c_0 \leq 285$ and a smooth submanifold $M' \subset \mathbb{R}^D$ such that

1. *$\mathscr{P} \subset M'$,*
2. *$\operatorname{reach}(M') \geq \left(1 - c_0\left(\frac{\eta}{\delta} + \theta\right)\frac{\rho}{\delta}\right)\rho$,*
3. *$T_{p_j}M' = \tilde{T}_j$ for all $1 \leq j \leq q$,*
4. *$d_H(M, M') \leq \delta\theta + \eta$,*
5. *M and M' are ambient isotopic.*

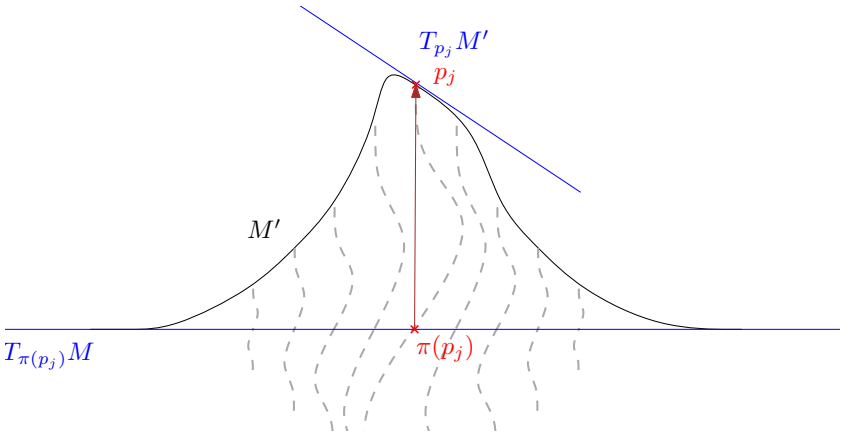


Fig. 3: An instance of the interpolating submanifold M'. Dashed lines correspond to the image of vertical lines by the ambient diffeomorphism Φ defining $M' = \Phi(M)$.

Theorem 11 fits a submanifold 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 for M. Indeed, if $\tilde{T}_1,\dots,\tilde{T}_q$ are estimated tangent spaces at the noisy base points $p_1,\dots,p_q$, M' has the virtue of being reconstructed by $\mathrm{Del}^{\omega_*}(\mathscr{P},\tilde{T})$

from Theorem 10. Since M' is topologically and geometrically close to M, we conclude that M is reconstructed as well by transitivity. In other words, Theorem 11 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, let us 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 $d_0\phi = 0$. 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 $\left\|R_j - I_D\right\|_{op} \leq 2\sin(\theta/2) \leq \theta$. 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 $d_{\pi(p_j)}\Phi = R_j$. 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 $\left\|a-\pi(p_j)\right\|$ increases. Notice that $d_a\psi_j = (R_j - I_D)$ and $\left\|\psi_j(a)\right\| \leq \ell\theta + \eta$ whenever $\phi\left(\frac{a-\pi(p_j)}{\ell}\right) \neq 0$. Defining $M' = \Phi(M)$, the facts that M' fits to $\mathscr{P}$ and $\tilde{T}$ and is Hausdorff-close to M follow by construction. Moreover, Theorem 4.19 of [22] (reproduced as Lemma 24 in this paper) states that the reach is stable with respect to $\mathscr{C}^2$-diffeomorphisms of the ambient space. The estimate on reach(M') relies on the following lemma stating differentials estimates on Φ.

Lemma 12 *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 $\mathscr{C}^\infty$-diffeomorphism. In addition, for all a in $\mathbb{R}^D$,*

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

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 13 *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 constant c_0 appearing in Theorem 11.*

4.2 Stability of the Tangential Delaunay Complex

Theorem 11 shows that even in the presence of noisy sample points 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 10 to some virtual submanifold M'. Denoting $\tilde{M} = \mathrm{Del}^{\omega_*}(\mathscr{P}, \tilde{T})$, since $d_H(M,\tilde{M}) \leq d_H(M,M') + d_H(M',\tilde{M})$ 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 10 and Theorem 11.

Theorem 14 (Stability of the Tangential Delaunay Complex) *There exists $\varepsilon_d > 0$ such that for all $\varepsilon \leq \varepsilon_d\rho$ and all $M \in \mathscr{M}_{D,d,\rho}$, the following holds. Let $\mathscr{P} \subset \mathbb{R}^D$ finite point cloud and $\tilde{T} = \left\{\tilde{T}_p\right\}_{p\in\mathscr{P}}$ be a family of d-dimensional linear subspaces of $\mathbb{R}^D$ such that*

- $\max_{p\in\mathscr{P}} d(p,M) \leq \eta$,
- $\max_{p\in\mathscr{P}} \angle(T_{\pi_M(p)}M, \tilde{T}_p) \leq \sin\theta$,
- $\mathscr{P}$ *is* ε*-sparse,*
- $\max_{x\in M} d(x,\mathscr{P}) \leq 2\varepsilon$.

If $\theta \leq \varepsilon/(1140\rho)$ *and* $\eta \leq \varepsilon^2/(1140\rho)$*, then,*

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

Indeed, applying the reconstruction algorithm of Theorem 10 even in the presence of noise and uncertainty on the tangent spaces actually recovers the submanifold M' built in Theorem 11. M' is isotopic to M and the quality of the approximation of M is at most impacted by the term $d_H(M,M') \leq \varepsilon\theta + \eta$. The lower bound on $\mathrm{reach}(M')$ is crucial, as constants appearing in Theorem 10 are not bounded for arbitrarily small reach.

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. For instance, a stability result similar to Theorem 14 could be derived readily for the so-called *cocone* complex [12] using the interpolating submanifold of Theorem 11.

5 Tangent Space Estimation and Decluttering Procedure

5.1 Additive Noise Case

We now focus on the estimation of tangent spaces in the model with additive noise $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$. The proposed method is similar to that of [2,28]. 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 is likely to recover the main directions spanned by M at p, and therefore yield a good approximation of T_pM. For $j = 1,\dots,n$ and $h > 0$ to be chosen later, define the local covariance matrix at X_j by

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

where $\bar{X}_j = \frac{1}{N_j}\sum_{i\neq j} X_i \mathbb{1}_{\mathscr{B}(X_j,h)}(X_i)$ is the barycenter of sample points contained in the ball $\mathscr{B}(X_j,h)$, and $N_j = |\mathscr{B}(X_j,h)\cap\mathbb{X}_n|$. Let us emphasize the fact that the normalization $1/(n-1)$ in the definition of $\hat{\Sigma}_j$ stands for technical convenience. In fact, any other normalization would yield the same guarantees on tangent spaces since only the principal directions of $\hat{\Sigma}_j$ play a role. Set $\hat{T}_j(h)$ to be the linear space spanned by the d eigenvectors associated with the d largest eigenvalues of $\hat{\Sigma}_j(h)$. Computing a basis of $\hat{T}_j(h)$ can be performed naively using a singular value decomposition of the full matrix $\hat{\Sigma}_j(h)$, although fast PCA algorithms [31] may lessen the computational dependence on the ambient dimension. We also denote by $\mathtt{TSE}(.,h)$ the function that maps any vector of points to the vector of their estimated tangent spaces, with

$$\hat{T}_j(h) = \mathtt{TSE}(\mathbb{X}_n,h)_j.$$

Proposition 15 *Set* $h = \left(c_{d,f_{min},f_{max}}\frac{\log n}{n-1}\right)^{1/d}$ *for* $c_{d,f_{min},f_{max}}$ *large enough. Assume that* $\sigma/h \leq 1/4$*. Then for n large enough, for all* $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$*,*

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

with probability larger than $1-4\left(\frac{1}{n}\right)^{\frac{2}{d}}$.

An important feature given by Proposition 15 is that the statistical error of our tangent space estimation procedure does not depend on the ambient dimension D. The intuition behind Proposition 15 is the following: if we assume that the true tangent space $T_{X_j}M$ is spanned by the first d vectors of the canonical basis, we can decompose $\hat{\Sigma}_j$ as

$$\hat{\Sigma}_j(h) = \left(\begin{array}{c|c} \hat{A}_j(h) & 0 \\ \hline 0 & 0 \end{array}\right) + \hat{R},$$

where $\hat{R}$ comes from the curvature of the submanifold along with the additive noise, and is of order $N_j(h)(h^3/(\rho(n-1)) + h\sigma) \lesssim h^{d+2}(h/\rho + \sigma/h)$, provided that h is roughly smaller than $(\log(n)/(n-1))^{1/d}$. On the other hand, for a bandwidth h of order $(\log(n)/(n-1))^{1/d}$, $\hat{A}_j(h)$ can be proved (Lemma 36) to be close to its deterministic counterpart

$$A_j(h) = \mathbb{E}\left(\left(\pi_{T_{X_j}M}(X) - \mathbb{E}\pi_{T_{X_j}M}(X)\right)\left(\pi_{T_{X_j}M}(X) - \mathbb{E}\pi_{T_{X_j}M}(X)\right)^t \mathbb{1}_{\mathscr{B}(X_j,h)}(X)\right),$$

where $\pi_{T_{X_j}M}$ denotes orthogonal projection onto $T_{X_j}M$ and expectation is taken conditionally on X_j. The bandwidth $(\log(n)/(n-1))^{1/d}$ may be thought of as the smallest radius that allows enough sample points in balls to provide an accurate estimation of the covariance matrices. Then, since $f_{min} > 0$, Lemma 35 shows that the minimum eigenvalue of $A(h)$ is of order h^{d+2}. At last, an eigenvalue perturbation result (Proposition 38) shows that $\hat{T}_j(h)$ must be close to $T_{X_j}M$ up to $(h^{d+3}/\rho + h^{d+1}\sigma)/(h^{d+2}) \approx h/\rho + \sigma/h$. The complete derivation is provided in Section E.1.

Then, it is shown in Lemma 32, based on the results of [11], that letting $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \rho\sigma/h)$ for $c_{d,f_{min},f_{max}}$ large enough, entails $\mathbb{X}_n$ is ε-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 ε, and for the stability reasons described in Section 3, we sparsify it with the farthest point sampling algorithm (Section 3.3) with scale parameter ε. Let $\mathbb{Y}_n$ denote the output of the algorithm. If $\sigma \leq h/4$, and $c_{d,f_{min},f_{max}}$ is large enough, we have the following.

Corollary 16 *With the above notation, for n large enough, with probability at least* $1 - 5\left(\frac{1}{n}\right)^{2/d}$,

- $\max_{X_j \in \mathbb{Y}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\rho}$,
- $\max_{X_j \in \mathbb{Y}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\rho}$,
- $\mathbb{Y}_n$ *is* ε*-sparse,*
- $\max_{x \in M} d(x, \mathbb{Y}_n) \leq 2\varepsilon$.

In other words, the previous result shows that $\mathbb{Y}_n$ satisfies the assumptions of Theorem 14 with high probability. We may then define $\hat{M}_{\mathtt{TDC}}$ to be the Tangential Delaunay Complex computed on $\mathbb{Y}_n$ and the collection of estimated tangent spaces $\mathtt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n}$, that is elements of $\mathtt{TSE}(\mathbb{X}_n, h)$ corresponding to elements of $\mathbb{Y}_n$, where h is the bandwidth defined in Proposition 15.

Definition 17 *With the above notation, define* $\hat{M}_{\mathtt{TDC}} = \mathrm{Del}^{\omega_*}(\mathbb{Y}_n, \mathtt{TSE}(\mathbb{X}_n, h)_{\mathbb{Y}_n})$.

Combining Theorem 14 and Corollary 16, it is clear that $\hat{M}_{\mathtt{TDC}}$ satisfies Theorem 7.

5.2 Clutter Noise Case

Let us now focus on the model with outliers $\mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$. We address problem of decluttering the sample $\mathbb{X}_n$, that is, to remove outliers. We follow ideas from [24]. 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 $S(x,T,h) = \{x\} \oplus \mathscr{B}_T(0,k_1 h) \oplus \mathscr{B}_{T^\perp}(0,k_2 h^2) \subset \mathbb{R}^D$, where $\oplus$ denotes the Minkovski sum, and $\mathscr{B}_T, \mathscr{B}_{T^\perp}$ are the Euclidean balls in T and $T^\perp$ respectively.

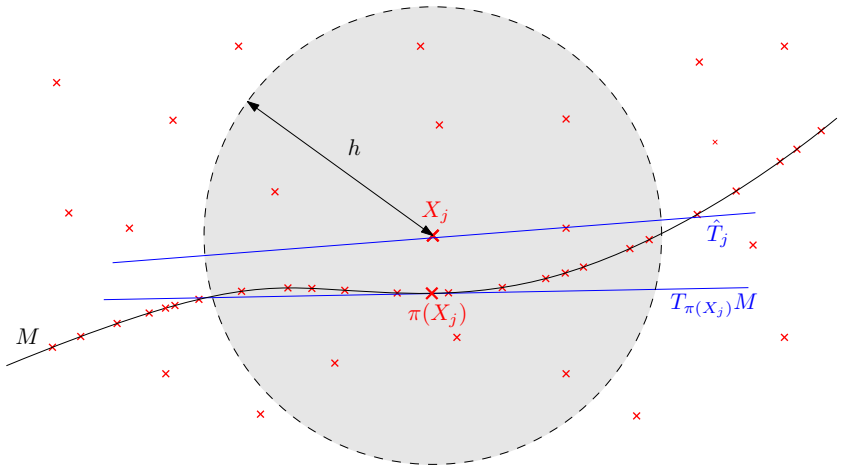


Fig. 4: Local PCA at an outlier point $X_j \in \mathbb{X}_n$.

Following notation of Section 2.1, for $P \in \mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$, let us write $P = \beta Q + (1-\beta)U_{\mathscr{B}_0}$. For h small enough, by definition of the slabs, $U_{\mathscr{B}_0}\left(S(x,T_{\pi(x)}M,h)\right) \asymp (k_1 h)^d (k_2 h^2)^{D-d} \asymp h^{2D-d}$. Furthermore, Figure 5 indicates that for k_1 and k_2 small enough, $Q\left(S(x,T_{\pi(x)}M,h)\right) \asymp Vol\left(S(x,T_{\pi(x)}M,h) \cap M\right) \asymp h^d$ if $d(x,M) \leq h^2$, and $Q\left(S(x,T_{\pi(x)}M,h)\right) = 0$ if $d(x,M) > h^2$. Coming back to $P = \beta Q + (1-\beta)U_{\mathscr{B}_0}$, we roughly get

$$\begin{array}{lll} P\left(S(x,T_{\pi(x)}M,h)\right) \asymp \beta h^d + (1-\beta)h^{2D-d} \asymp h^d & \text{if} & d(x,M) \leq h^2, \\ P\left(S(x,T_{\pi(x)}M,h)\right) \asymp \quad 0 \quad + (1-\beta)h^{2D-d} \asymp h^{2D-d} & \text{if} & d(x,M) > h^2, \end{array}$$

as h goes to 0, for k_1 and k_2 small enough. Since $h^{2D-d} \ll h^d$, the measure $P(S(x,T,h))$ of the slabs clearly is discriminatory for decluttering, provided that tangent spaces are known.

Based on this intuition, we define the elementary step of our decluttering procedure as the map $\mathtt{SD}_t(.,.,h)$, that sends a vector $P = (p_1,\dots,p_r) \subset \mathbb{R}^D$ and a corresponding vector of (estimated) tangent spaces $T_{\mathscr{P}} = (T_1,\dots,T_r)$ onto a subvector of $\mathscr{P}$ according to the rule

$$p_j \in \mathtt{SD}_t(\mathscr{P}, T_{\mathscr{P}}, h) \quad \Leftrightarrow \quad |S(p_j,T_j,h) \cap \mathscr{P}| \geq t(n-1)h^d,$$

where t is a threshold to be fixed. This procedure relies on counting how many sample points lie in the slabs of direction the estimated tangent spaces (see Figure 5).

Since tangent spaces are unknown, the following result gives some insight on the relation between the accuracy of the tangent space estimation and the decluttering performance that can be reached.

Lemma 18 *Let $K > 0$ be fixed. There exist constants $k_1(K)$ and $k_2(\rho,K)$ such that for every $h \leq 1$ and x in $\mathbb{R}^D$, $S(x,T,h) \subset \mathscr{B}(x,h/2)$. Moreover, for every $h \leq h_+ \wedge 1$ we have*

$$h/\sqrt{2} \geq d(x,M) \geq h^2/\rho \quad \text{and} \quad \angle\left(T_{\pi_M(x)}M, T\right) \leq Kh/\rho \Rightarrow S(x,T,h) \subset S'(x,T_{\pi_M(x)}M,h),$$

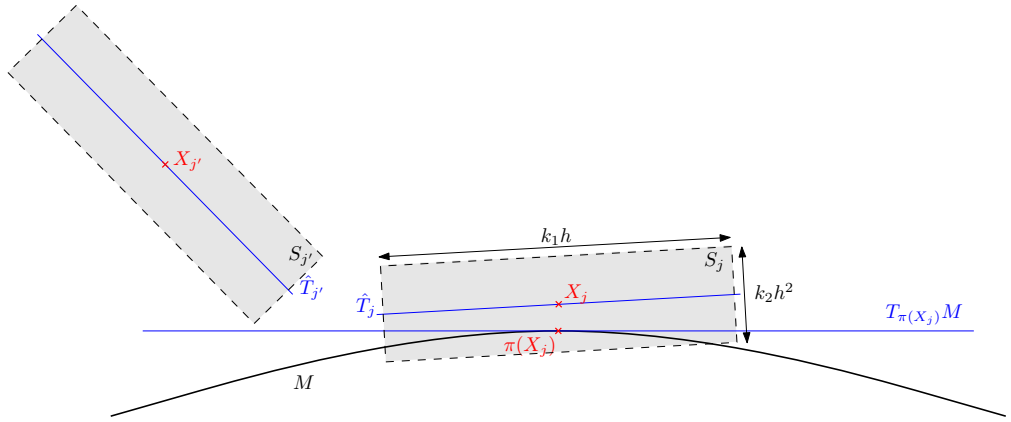


Fig. 5: The slab $S(X_j, \hat{T}_j, h)$ is centered at X_j and has size $k_1 h$ in the d directions spanned by $\hat{T}_j$, and size $k_2 h^2$ in the $D-d$ directions normal to $\hat{T}_j$.

where $S'(x, T_{\pi_M(x)}M, h)$ is a larger slab with parameters $k_1'(\rho, K)$ and $k_2'(\rho, K)$, and satisfies $S'(x, T_{\pi_M(x)}M, h) \cap M = \emptyset$. In addition, there exists $k_3(\rho, K)$ such that for all x and y are in M,

$$\angle(T_x M, T) \leq Kh/\rho \quad \text{and} \quad \|x-y\| \leq k_3 h \quad \Rightarrow \quad y \in S(x, T, h).$$

Possible values for k_1 and k_2 are, respectively, $\frac{1}{16(K \vee 1)}$ and $\frac{1}{16(\rho \vee K \vee 1)}$, and k_3 can be taken as $k_1 \wedge \frac{\rho k_2}{1+2K}$.

The proof of Lemma 18, mentioned in [24], follows from elementary geometry, combined with the definition of the reach and Proposition 25.

Roughly, Lemma 18 states that the decluttering performance is of order the square of the tangent space precision, hence will be closely related to the performance of the tangent space estimation procedure TSE. Unfortunately, a direct application of TSE to the corrupted sample $\mathbb{X}_n$ leads to slightly worse precision bounds, in terms of angle deviation. Typically, the angle deviation would be of order $n^{-1/(d+1)}$. However, this precision is enough to remove outliers points which are at distance at least $n^{-2/(d+1)}$ from M. Then running our TSE on this refined sample $\mathtt{SD}_t(\mathbb{X}_n, \mathtt{TSE}(\mathbb{X}_n), n^{-1/(d+1)})$ leads to better angle deviation rates, hence better decluttering performance, and so on.

Let us introduce an iterative decluttering procedure in a more formal way. We choose the initial bandwidth $h_0 = \left(c_{d,f_{min},f_{max},\rho} \frac{\log n}{\beta(n-1)}\right)^{\gamma_0}$, with $\gamma_0 = 1/(d+1)$, and define the first set $\mathbb{X}^{(-1)} = \mathbb{X}_n$ as the whole sample. We then proceed recursively, setting $h_{k+1} = \left(c_{d,f_{min},f_{max},\rho} \frac{\log n}{\beta(n-1)}\right)^{\gamma_{k+1}}$, with γ_{k+1} satisfying $\gamma_{k+1} = (2\gamma_k + 1)/(d+2)$. This recursion formula is driven by the optimization of a trade-off between imprecision terms in tangent space estimation, as may be seen from (5). An elementary calculation shows that

$$\gamma_k = \frac{1}{d} - \frac{1}{d(d+1)} \left(\frac{2}{d+2}\right)^k.$$

With this updated bandwidth we define

$$\mathbb{X}^{(k+1)} = \mathtt{SD}_t(\mathbb{X}^{(k)}, \mathtt{TSE}(\mathbb{X}^{(k)}, h_{k+1}), h_{k+1}).$$

In other words, at step $k+1$ we use a smaller bandwidth h_{k+1} in the tangent space estimation procedure TSE. Then we use this better estimation of tangent spaces to run the elementary decluttering step SD. The performance of this procedure is guaranteed by the following proposition. With a slight abuse of notation, if X_j is in $\mathbb{X}^{(k)}$, $\mathtt{TSE}(\mathbb{X}^{(k)}, h)_j$ will denote the corresponding tangent space of $\mathtt{TSE}(\mathbb{X}^{(k)}, h)$.

Proposition 19 *In the clutter noise model, for t, $c_{d,f_{min},f_{max},\rho}$ and n large enough, k_1 and k_2 small enough, the following properties hold with probability larger than $1-7\left(\frac{1}{n}\right)^{2/d}$ for all $k \geq 0$.*

***Initialization**:*

- *For all $X_j \in \mathbb{X}^{(-1)}$ such that $d(X_j,M) \leq h_0/\sqrt{2}$, $\angle\big(\mathtt{TSE}(\mathbb{X}^{(-1)},h_0)_j, T_{\pi(X_j)}M\big) \leq C_{d,f_{min},f_{max}} h_0/\rho$.*
- *For every $X_j \in M \cap \mathbb{X}^{(-1)}$, $X_j \in \mathbb{X}^{(0)}$.*
- *For every $X_j \in \mathbb{X}^{(-1)}$, if $d(X_j,M) > h_0^2/\rho$, then $X_j \notin \mathbb{X}^{(0)}$.*

***Iterations**:*

- *For all $X_j \in \mathbb{X}^{(k)}$ such that $d(X_j,M) \leq h_{k+1}/\sqrt{2}$, $\angle\big(\mathtt{TSE}(\mathbb{X}^{(k)},h_{k+1})_j, T_{\pi(X_j)}M\big) \leq C_{d,f_{min},f_{max}} h_{k+1}/\rho$.*
- *For every $X_j \in M \cap \mathbb{X}^{(k)}$, $X_j \in \mathbb{X}^{(k+1)}$.*
- *For every $X_j \in \mathbb{X}^{(k)}$, if $d(X_j,M) > h_{k+1}^2/\rho$, then $X_j \notin \mathbb{X}^{(k+1)}$.*

This result is threefold. Not only can we distinguish data and outliers within a decreasing sequence of offsets of radii h_k^2/ρ around M, but we can also ensure that no point of M is removed during the process with high probability. Moreover, it also provides a convergence rate for the estimated tangent spaces $\mathtt{TSE}(\mathbb{X}_k, h_{k+1})$.

Now fix a precision level δ. If k is larger than $(\log(1/\delta)-\log(d(d+1))/(\log(d+2)-\log(2))$, then $1/d > \gamma_k \geq 1/d-\delta$. Let us define k_δ as the smallest integer satisfying $\gamma_k \geq 1/d-\delta$, and denote by $\mathbb{Y}_n^\delta$ the output of the farthest point sampling algorithm applied to $\mathbb{X}^{(k_\delta)}$ with parameter $\varepsilon = c_{d,f_{min}f_{max}} h_{k_\delta}$, for $c_{d,f_{min}f_{max}}$ large enough. Define also $\hat{T}^\delta$ as the restriction of $\mathtt{TSE}(\mathbb{X}^{(k_\delta)}, h_{k_\delta})$ to the elements of $\mathbb{Y}_n^\delta$.

According to Proposition 19, the decluttering procedure removes no data point on M with high probability. In other words, $\mathbb{X}^{(k_\delta)} \cap M = \mathbb{X}_n \cap M$, and as a consequence, $\max_{x\in M} d(x,\mathbb{X}^{(k_\delta)}) \leq c_{d,f_{min}} \left(\frac{\log n}{\beta n}\right)^{1/d} \ll h_{k_\delta}$ with high probability (see Lemma 32). As a consequence, we obtain the following.

Corollary 20 *With the above notation, for n large enough, with probability larger than $1-8\left(\frac{1}{n}\right)^{2/d}$,*

- $\max_{X_j\in\mathbb{Y}_n^\delta} d(X_j,M) \leq \frac{\varepsilon^2}{1140\rho}$,
- $\max_{X_j\in\mathbb{Y}_n^\delta} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\rho}$,
- *$\mathbb{Y}_n^\delta$ is ε-sparse,*
- $\max_{x\in M} d(x,\mathbb{Y}_n^\delta) \leq 2\varepsilon$.

We are now able to define the estimator $\hat{M}_{\mathtt{TDC}\delta}$.

Definition 21 *With the above notation, define $\hat{M}_{\mathtt{TDC}\delta} = \mathrm{Del}^{\omega_*}\left(\mathbb{Y}_n^\delta, \hat{T}^\delta\right)$.*

Combining Theorem 14 and Corollary 20, it is clear that $\hat{M}_{\mathtt{TDC}\delta}$ satisfies Theorem 8.

Finally, we turn to the asymptotic estimator $\hat{M}_{\mathtt{TDC}+}$. Set $h_\infty = \left(c_{d,f_{min},f_{max},\rho}\frac{\log n}{\beta(n-1)}\right)^{1/d}$, and let $\hat{k}$ denote the smallest integer such that $\min\{d(X_j,M) | d(X_j,M) > h_\infty^2/\rho\} > h_{\hat{k}}^2/\rho$. Since $\mathbb{X}_n$ is a (random) finite set, we can always find such a random integer $\hat{k}$ that provides a sufficient number of iterations to obtain the asymptotic decluttering rate. For this random iteration $\hat{k}$, we can state the following result.

Proposition 22 *Under the assumptions of Corollary 20, for every $X_j \in X^{(\hat{k}+1)}$, we have*

$$\angle(\mathtt{TSE}(\mathbb{X}^{(\hat{k}+1)}, h_\infty)_j, T_{\pi(X_j)}M) \leq C_{d,f_{min},f_{max}} h_\infty/\rho.$$

As before, taking $\mathbb{Y}_n^+$ as the result of the farthest point sampling algorithm based on $\mathbb{X}^{(\hat{k}+1)}$, and T^+ the vector of tangent spaces $\mathtt{TSE}(\mathbb{X}^{(\hat{k}+1)}, h_\infty)_j$ such that $\mathbb{X}_j^{(\hat{k}+1)} \in \mathbb{Y}_n^+$, we can construct our last estimator.

Definition 23 *With the above notation, define $\hat{M}_{\mathtt{TDC}+} = \mathrm{Del}^{\omega_*}(\mathbb{Y}_n^+, T^+)$.*

In turn, Proposition 22 implies that $\hat{M}_{\mathtt{TDC}+}$ satisfies Theorem 9.

6 Conclusion

In this work, we gave results on explicit manifold reconstruction with simplicial complexes. We built estimators $\hat{M}_{\mathtt{TDC}}$, $\hat{M}_{\mathtt{TDC}\delta}$ and $\hat{M}_{\mathtt{TDC}+}$ in two statistical models. We proved minimax rates of convergence for the Hausdorff distance and consistency results for ambient isotopic reconstruction. Since $\hat{M}_{\mathtt{TDC}}$ is minimax optimal in the additive noise model for σ small, and uses the Tangential Delaunay Complex of [4], the latter is proved to be optimal. Moreover, rates of [24] 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 decluttering method both based on local PCA were studied.

In the model with outliers, the proposed reconstruction method achieves a rate of convergence that can be as close as desired to the minimax rate of convergence, depending on the number of iterations of the decluttering procedure. Though this procedure seems to be well adapted to our reconstuction scheme — which is based on tangent spaces estimation — we believe that it could be of interest in the context of other applications. Also, further investigation may be carried out to compare this decluttering procedure to existing ones [9, 19].

As briefly mentioned below Theorem 7, our approach is likely to be suboptimal in cases where noise level σ is large. In such cases, with additional structure on the noise such as *centered* and *independent from the source*, other statistical procedures such as deconvolution [24] could be adapted to provide vertices to the Tangential Delaunay Complex. Tangential properties of deconvolution are still to be studied.

The effective construction of $\hat{M}_{TDC\delta}$ can be performed using existing algorithms. Namely, Tangential Delaunay Complex, farthest point sampling, local PCA and point-to-linear subspace distance computation for slab counting. A crude upper bound on the time complexity of a naive step-by-step implementation is

$$O\left(nD\left[2^{O(d^2)}+\log(1/\delta)D(D+n)\right]\right),$$

since the precision δ requires no more than $\log(1/\delta)$ iterations of the decluttering procedure. It is likely that better complexity bounds may be obtained using more refined algorithms, such as fast PCA [31], that lessens the dependence on the ambient dimension D. An interesting development would be to investigate a possible precision/complexity tradeoff, as done in [3] for community detection in graphs for instance.

Even though Theorem 11 is applied to submanifold estimation, we 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 future paper.

Acknowledgements We 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. We are also grateful to the reviewers whose comments helped enhancing substantially this paper.

This work was supported by ANR project TopData ANR-13-BS01-0008 and by the Advanced Grant of the European Research Council GUDHI (Geometric Understanding in Higher Dimensions). E. Aamari was supported by the Conseil régional d'Île-de-France under a doctoral allowance of its program Réseau de Recherche Doctoral en Mathématiques de l'Île-de-France (RDM-IdF).

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. Ery Arias-Castro, Gilad Lerman, and Teng Zhang. Spectral clustering based on local PCA. J. Mach. Learn. Res., 18:Paper No. 9, 57, 2017.
3. Ery Arias-Castro and Nicolas Verzelen. Community detection in dense random networks. Ann. Statist., 42(3):940–969, 2014.
4. Jean-Daniel Boissonnat and Arijit Ghosh. Manifold reconstruction using tangential Delaunay complexes. Discrete Comput. Geom., 51(1):221–267, 2014.
5. 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.
6. Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. ESAIM Probab. Stat., 9:323–375, 2005.
7. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. Concentration inequalities. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.
8. 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.
9. Mickaël Buchet, Tamal K. Dey, Jiayuan Wang, and Yusu Wang. Declutter and resample: Towards parameter free denoising. CoRR, abs/1511.05479, 2015.
10. 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.
11. Frédéric Chazal, Marc Glisse, Catherine Labruère, and Bertrand Michel. Convergence rates for persistence diagram estimation in topological data analysis. Journal of Machine Learning Research, 16:3603–3635, 2015.
12. 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.
13. 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.
14. Antonio Cuevas and Alberto Rodríguez-Casal. On boundary estimation. Adv. in Appl. Probab., 36(2):340–354, 2004.
15. Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. SIAM J. Numer. Anal., 7:1–46, 1970.
16. Giuseppe De Marco, Gianluca Gorni, and Gaetano Zampieri. Global inversion of functions: an introduction. NoDEA Nonlinear Differential Equations Appl., 1(3):229–248, 1994.
17. 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.
18. Manfredo Perdigão do Carmo. Riemannian geometry. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, 1992. Translated from the second Portuguese edition by Francis Flaherty.
19. D. L. Donoho. De-noising by soft-thresholding. IEEE Trans. Inf. Theor., 41(3):613–627, May 1995.
20. Lutz Dümbgen and Günther Walther. Rates of convergence for random approximations of convex sets. Adv. in Appl. Probab., 28(2):384–393, 1996.
21. 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.
22. Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93:418–491, 1959.
23. Herbert Federer. Geometric measure theory. Die Grundlehren der mathematischen Wissenschaften, Band 153. Springer-Verlag New York Inc., New York, 1969.
24. 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.
25. Christopher R. Genovese, Marco Perone-Pacifico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. J. Mach. Learn. Res., 13:1263–1291, 2012.
26. 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.
27. L. LeCam. Convergence of estimates under dimensionality restrictions. Ann. Statist., 1:38–53, 1973.
28. Mauro Maggioni, Stanislav Minsker, and Nate Strawn. Multiscale dictionary learning: non-asymptotic bounds and robustness. J. Mach. Learn. Res., 17:Paper No. 2, 51, 2016.
29. E. Mammen and A. B. Tsybakov. Asymptotical minimax recovery of sets with smooth boundaries. Ann. Statist., 23(2):502–524, 1995.
30. 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.
31. 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 $\mathscr{C}^2$-diffeomorphisms.

Lemma 24 (Theorem 4.19 in [22]) *Let $A \subset \mathbb{R}^D$ with $\mathrm{reach}(A) \geq \rho > 0$ and $\Phi : \mathbb{R}^D \longrightarrow \mathbb{R}^D$ is a $\mathscr{C}^1$-diffeomorphism such that Φ, Φ^{-1}, and $d\Phi$ 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 $b_1 = \sup_x \|d_x\phi\|$, $b_2 = \sup_x \left\|d_x^2\phi\right\|_{\mathrm{op}}$, 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 12) First notice that the sum appearing in (1) consists of at most one term. Indeed, since $\phi \equiv 0$ outside $\mathscr{B}(0,1)$, if $\phi_\ell\left(a - \pi(p_j)\right) \neq 0$ for some $j \in \{1,\dots,q\}$, then $\left\|a - \pi(p_j)\right\| \leq \ell$. Consequently, for all $i \neq j$,

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

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

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

where the last line follows from $b_1 \leq 5/2$, $6\eta \leq \ell$ and $\theta \leq \pi/64$. Therefore, $d_a\Phi$ is invertible for all $a \in \mathbb{R}^D$, and $(d_a\Phi)^{-1} = \sum_{i=0}^{\infty}(I_D - d_a\Phi)^i$. 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 $\mathscr{C}^\infty$-diffeomorphism by Hadamard-Cacciopoli theorem [16].

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

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

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

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

where the first inequality holds since $\|d_a\Phi - I_D\|_{\mathrm{op}} < 1$, and $\|\cdot\|_{\mathrm{op}}$ is sub-multiplicative.

(iii) Second order: Again, since the sum (1) includes at most one term,

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

Proof (of Theorem 11) 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 $d_x\phi_l|_{x=0} = 0$, for all $j \in \{1,\dots,q\}$, $d_a\Phi|_{a=\pi(p_j)} = R_j$. Thus,

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

  by definition of R_j.
- *Proximity to M:* The bound on $d_H(M,M') = d_H\big(M,\Phi(M)\big)$ follows from the correspondence

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

- *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 $0 \le t \le 1$. Since $\Phi_{(t)} - I_D = t\big(\Phi - I_D\big)$, the arguments above show that $\Phi_{(t)}$ is a global diffeomorphism of $\mathbb{R}^D$ for all $t \in [0,1]$. Moreover $\Phi_{(0)} = I_D$, 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 $d\Phi$. Applying Lemma 24 leads to

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

  Now, replace ℓ by its value $\delta/3$, and write $c_1 = 3C_1 \le 21/2 \le 11$ and $c_2 = 3^2 C_2 \le 252$. We derive

$$\begin{aligned} \operatorname{reach}\left(M'\right) &\ge \left(1 - 2c_1\left(\frac{\eta}{\delta} + \theta\right)\right)\left(1 - c_1\left(\frac{\eta}{\delta} + \theta\right) - c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\rho\right)\rho \\ &\ge \left(1 - 3c_1\left(\frac{\eta}{\delta} + \theta\right) - c_2\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\rho\right)\rho \\ &\ge \left(1 - (3c_1 + c_2)\left(\frac{\eta}{\delta^2} + \frac{\theta}{\delta}\right)\rho\right)\rho, \end{aligned}$$

  where for the last line we used that $\delta/\rho \le 1$. The desired lower bound follows taking $c_0 = 3c_1 + c_2 \le 285$.

B Some Geometric Properties under Reach Regularity Condition

B.1 Reach and Projection on the Submanifold

In this section we state intermediate results that connect the reach condition to orthogonal projections onto the tangent spaces. They are based on the following fundamental result.

Proposition 25 (Theorem 4.18 in [22]) *For all x and y in M,*

$$\|(y-x)_\perp\| \le \frac{\|y-x\|^2}{2\rho},$$

where $(y-x)_\perp$ denotes the projection of $y-x$ onto $T_xM^\perp$.

From Proposition 25 we may deduce the following property about trace of Euclidean balls on M.

Proposition 26 *Let $x \in \mathbb{R}^D$ be such that $d(x,M) = \Delta \leq h \leq \frac{\rho}{8}$, and let y denote $\pi(x)$. Then,*

$$\mathscr{B}\left(y, r_h^-\right) \cap M \subset \mathscr{B}(x,h) \cap M \subset \mathscr{B}\left(y, r_h^+\right) \cap M,$$

where $r_h^2 + \Delta^2 = h^2$, $(r_h^-)^2 = \left(1 - \frac{\Delta}{\rho}\right) r_h^2$, and $(r_h^+)^2 = \left(1 + \frac{2\Delta}{\rho}\right) r_h^2$.

Proof (of Proposition 26) Let z be in $M \cap \mathscr{B}(x,h)$, and denote by δ the quantity $\|z-y\|$. We may write

$$\|z-x\|^2 = \delta^2 + \Delta^2 + 2\langle z-y, y-x\rangle, \tag{2}$$

hence $\delta^2 \leq h^2 - \Delta^2 - 2\langle z-y, y-x\rangle$. Denote, for u in $\mathbb{R}^D$, by $u_\perp$ its projection onto $T_y M^\perp$. Since $\langle z-y, y-x\rangle = \langle (z-y)_\perp, y-x\rangle$, Proposition 25 ensures that

$$\delta^2 \left(1 - \frac{\Delta}{\rho}\right) \leq r_h^2.$$

Since $\Delta \leq h \leq \rho/8$, it comes $\delta^2 \leq (1 + 2\frac{\Delta}{\rho}) r_h^2$. On the other hand, (2) and Proposition 25 also yield

$$\|z-x\|^2 \leq \delta^2 \left(1 + \frac{\Delta}{\rho}\right) + \Delta^2.$$

Hence, if $\delta^2 \leq \left(1 - \frac{\Delta}{\rho}\right) r_h^2$, we have

$$\|z-x\|^2 \leq r_h^2 + \Delta^2 = h^2.$$

Also, the following consequence of Proposition 25 will be of particular use in the decluttering procedure.

Proposition 27 *Let h and h_k be bandwidths satisfying $h_k^2/\rho \leq h \leq h_k$. Let x be such that $d(x,M) \leq h/\sqrt{2}$ and $\pi_M(x) = 0$, and let z be such that $\|z-x\| \leq h$ and $d(z,M) \leq h_k^2/\rho$. Then*

$$\|z_\perp\| \leq \frac{6h_k^2}{\rho},$$

where $z_\perp$ denotes the projection of z onto $T_0 M^\perp$.

Proof (of Proposition 27) Let y denote $\pi_M(z)$. A triangle inequality yields $\|y\| \leq \|y-z\| + \|z-x\| + \|x\| \leq h_k^2/\rho + (1 + 1/\sqrt{2})h \leq 3h_k$. Proposition 25 ensures that $\|y_\perp\| \leq \|y\|^2/(2\rho) \leq (9h_k^2)/(2\rho)$. Since $\|z_\perp\| \leq \|y_\perp\| + h_k^2/\rho$, we have $\|z_\perp\| \leq 6h_k^2/\rho$.

At last, let us prove Lemma 18, that gives properties of intersections of ambient slabs with M.

Proof (of Lemma 18) Set $k_1 = \frac{1}{16(K \vee 1)}$, $k_2 = \frac{1}{16(K \vee \rho \vee 1)}$, and $k_3 = k_1 \wedge \frac{\rho k_2}{1+2K} \wedge 1$. For all $h > 0$, and $z \in S(x,T,h)$, triangle inequality yields $\|z-x\| \leq \|\pi_T(z-x)\| + \|\pi_{T^\perp}(z-x)\| \leq (k_1 + k_2 h)h$. Since $h \leq 1$ and $k_1 + k_2 \leq 1/2$, we get $z \in \mathscr{B}(x, h/2)$.

Now, suppose that $h/\sqrt{2} \geq d(x,M) \geq h^2/\rho$ and $\angle\left(T_{\pi(x)}M, T\right) \leq Kh/\rho$. For short we write $T_0 = T_{\pi(x)}M$. Let $z \in S(x,T,h)$, since $h \leq 1$, it comes

$$\|\pi_{T_0}(z-x)\| \leq \|z-x\| \leq (k_1 + k_2)h = k_1' h,$$

with $k_1' = k_1 + k_2$. On the other hand

$$\|\pi_{T_0^\perp}(z-x)\| \leq \|\pi_{T_0^\perp}\pi_T(z-x)\| + \|\pi_{T_0^\perp}\pi_{T^\perp}(z-x)\| \leq (Kh/\rho)(k_1 h) + k_2 h^2 = k_2' h^2,$$

with $k_2' = k_1 K/\rho + k_2$. Hence $S(x,T,h) \subset S'(x,T_0,h)$, for the constants k_1' and k_2' defined above. It remains to prove that $S'(x,T_0,h) \cap M = \emptyset$. To see this, let $z \in S'(x,T_0,h)$, and $y = \pi(x)$. Since $k_1' + k_2' \leq 1/4$, we have $\|y-z\| \leq \|y-x\| + \|x-z\| \leq h(1/\sqrt{2} + 1/4)$. For the normal part, we may write

$$\|\pi_{T_0^\perp}(z-y)\| \geq \|\pi_{T_0^\perp}(y-x)\| - \|\pi_{T_0^\perp}(x-z)\| \geq h^2(1/\rho - k_2').$$

Since $k_2' \leq 1/(8\rho)$, we have $\|\pi_{T_0^\perp}(z-y)\| > \|y-z\|^2/(2\rho)$, hence Proposition 25 ensures that $z \notin M$.

At last, suppose that $x \in M$ and $y \in \mathscr{B}(x, k_3 h) \cap M$. Since $k_3 \leq k_1$, we have $\|\pi_T(y-x)\| \leq k_1 h$. Next, we may write

$$\|\pi_{T^\perp}(y-x)\| \leq \|\pi_{T^\perp}\pi_{T_0}(y-x)\| + \|\pi_{T^\perp}\pi_{T_0^\perp}(y-x)\|.$$

Since $y \in M$, Proposition 25 entails $\|\pi_{T_0^\perp}(y-x)\| \leq \|y-x\|^2/(2\rho) \leq k_3^2 h^2/(2\rho)$. It comes

$$\|\pi_{T^\perp}(y-x)\| \leq \frac{h^2}{\rho}\left(k_3 K + \frac{k_3^2}{2}\right) \leq k_2 h^2.$$

Hence $y \in S(x,T,h)$.

B.2 Reach and Exponential Map

In this section we state results that connect Euclidean and geodesic quantities under reach regularity condition. We start with a result linking reach and principal curvatures.

Proposition 28 (Proposition 6.1 in [30]) *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 29 (Corollary 1.4 in [1]) *The injectivity radius of M is at least $\pi\rho$.*

Denoting by $d_M(\cdot,\cdot)$ 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 30 *For all $x, y \in M$ such that $\|x-y\| \leq \rho/4$,*

$$\|x-y\| \leq d_M(x,y) \leq \alpha \|x-y\|.$$

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 $\|R(r,v)\| \leq \frac{r^2}{2\rho}$.

Proof (of Proposition 30) The first statement is a direct consequence of Proposition 6.3 in [30]. 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 $r \leq \alpha \|x-y\|$.

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

Proposition 31 *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 $y = \exp_x(rv)$, $d_yV = J(r,v)drdv$. Then*

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

where $c_d = 2^{-d}$ and $C_d = 2^d$. As a consequence, if $\mathscr{B}_M(x,r)$ denotes the geodesic ball of radius r centered at x, then, if $r \leq \frac{\rho}{4}$,

$$c'_d r^d \leq Vol(\mathscr{B}_M(x,r)) \leq C'_d r^d,$$

with $c'_d = c_d V_d$ and $C'_d = C_d V_d$, where V_d denotes the volume of the unit d-dimensional Euclidean ball.

Proof (of Proposition 31) Denoting $A_{r,v} = d_{rv}\exp_x$, the Area Formula [23, Section 3.2.5] asserts that $J(r,v) = r^{d-1}\sqrt{\det\left(A^t_{r,v}A_{r,v}\right)}$. Note that from Proposition 6.1 in [30] together with the Gauss equation [18, p. 130], the sectional curvatures in M are bounded by $|\kappa| \leq 2/\rho^2$. Therefore, the Rauch theorem [21, Lemma 5] 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,

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

Since $Vol(\mathscr{B}_M(x,r)) = \int_{s=0}^r \int_{v \in \mathscr{S}_{d-1}} J(s,v)dsdv$, where $\mathscr{S}_{d-1}$ denotes the unit $d-1$-dimensional sphere, the bounds on the volume easily follows.

C Some Technical Properties of the Statistical Model

C.1 Covering and Mass

Lemma 32 *Let $Q_0 \in \mathscr{U}_M(f_{min}, f_{max})$. Then for all $p \in M$ and $r \leq \rho/4$,*

$$Q_0\big(\mathscr{B}(p,r)\big) \geq a_d f_{min} r^d,$$

where $a_d > 0$. As a consequence, for n large enough and for all $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$, with probability larger that $1 - \left(\frac{1}{n}\right)^{2/d}$,

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

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

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

Proof (of Lemma 32) The first statement is a direct corollary of Proposition 31, since for all $r \leq \rho/4$,

$$Q_0\big(\mathscr{B}(p,r)\big) = \int_{\mathscr{B}(p,r)} f d\mathscr{H}^d \geq f_{min} Vol\left(\mathscr{B}(p,r) \cap M\right) \geq a_d f_{min} r^d,$$

where a_d can be taken to be equal to c'_d of Proposition 31. Let us now prove the second statement. By definition, sample $X_i \in \mathbb{X}_n$, that has distribution $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$ can be written as $X_i = Y_i + Z_i$, with Y_i having distribution $Q_0 \in \mathscr{G}_{D,d,f_{min},f_{max},\rho}$, and $\|Z_i\| \leq \sigma$. From the previous point, letting $a = a_d f_{min}$, Q_0 fulfils the so-called (a,d)-standard assumption of [11] for $r \leq \rho/4$. Looking carefully at the proof of Lemma 10 in [11] shows that its conclusion still holds for measures satisfying the (a,d)-standard assumption for small radii only. Therefore, writing $\mathbb{Y}_n = \{Y_1, \ldots, Y_n\}$, for $r \leq \rho/8$ we obtain

$$\mathbb{P}_{Q_0}\left(d_H(M, \mathbb{Y}_n) > r\right) \leq \frac{4^d}{a r^d} \exp\left(-n \frac{a}{2^d} r^d\right).$$

The statement then follows using that $d_H(\mathbb{X}_n, \mathbb{Y}_n) \leq \sigma$, and setting $r = C_{d,f_{min}} \left(\frac{\log n}{n}\right)^{1/d}$ with $C^d_{d,f_{min}} \frac{a}{2^{d+1}} \geq 1 + 2/d$.

To prove the last point, notice that 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 Q_0. Therefore,

$$\begin{aligned}
\mathbb{P}_P\left(d_H(M, \mathbb{X}_n \cap M) > r \,|\, |\mathbb{X}_n \cap M| = k\right) &= \mathbb{P}_{Q_0}\left(d_H(M, \mathbb{X}_k \cap M) > r\right) \\
&\leq \frac{4^d}{a r^d} \exp\left(-k \frac{a}{2^d} r^d\right).
\end{aligned}$$

Hence,

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

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

We now focus on proving Lemma 2. For its proof, we need the following piece of notation. For all bounded subset $K \subset \mathbb{R}^D$ and $\varepsilon > 0$, we let $\mathrm{cv}_K(\varepsilon)$ denote the Euclidean covering number of K. That is, $\mathrm{cv}_K(\varepsilon)$ is the minimal number k of Euclidean open balls of radii ε and centered at elements of K that are needed to cover K.

Lemma 33 *Let $K \subset \mathbb{R}^D$ be a bounded subset. If K is path connected, then for all $\varepsilon > 0$, $\operatorname{diam}(K) \leq 2\varepsilon \mathrm{cv}_K(\varepsilon)$.*

Proof (of Lemma 33) Let $p, q \in K$ and $\gamma : [0,1] \to K$ be a continuous path joining $\gamma(0) = p$ and $\gamma(1) = q$. Writing $N = \mathrm{cv}_K(\varepsilon)$, let $x_1, \ldots, x_N \in \mathbb{R}^D$ be the centers of a covering of K by open balls of radii ε. We let U_i denote $\{t, \|\gamma(t) - x_i\| < \varepsilon\} \subset [0,1]$. By construction of the covering, there exists $x_{(1)} \in \{x_1, \ldots, x_N\}$ such that $\|p - x_{(1)}\| < \varepsilon$. Then $U_{(1)} \ni \gamma(0) = p$ is a non-empty open subset of $[0,1]$, so that $t_{(1)} = \sup U_{(1)}$ is positive. If $t_{(1)} = 1$, then $\|q - x_{(1)}\| \leq \varepsilon$, and in particular $\|q - p\| \leq 2\varepsilon$. If $t_{(1)} < 1$, since $U_{(1)}$ is an open subset of $[0,1]$, we see that $\gamma(t_{(1)}) \notin U_{(1)}$. But $\cup_{i=1}^N U_i$ is an open cover of $[0,1]$, which yields the existence $U_{(2)}$ such that $\gamma(t_{(1)}) \in U_{(2)}$, and for all $t < t_{(1)}$, $\gamma(t) \notin U_{(2)}$. Then consider $t_{(2)} = \sup U_{(2)}$, and so on. Doing so, we build by induction a sequence of numbers $0 < t_{(1)} < \ldots < t_{(k)} \leq 1$ and distinct centers $x_{(1)}, \ldots, x_{(k)} \in \{x_1, \ldots, x_N\}$ ($k \leq N$) such that $\|p - x_{(1)}\| < \varepsilon$, $\|q - x_{(k)}\| \leq \varepsilon$, with $\|\gamma(t_{(i)}) - x_{(i)}\| \leq \varepsilon$ for $1 \leq i \leq k$ and $\|\gamma(t_{(i)}) - x_{(i+1)}\| < \varepsilon$ for $1 \leq i \leq k-1$. In particular, $\|x_{(i)} - x_{(i+1)}\| \leq 2\varepsilon$ for all $1 \leq i \leq k-1$. To conclude, write

$$\begin{aligned}
\|p - q\| &\leq \|p - x_{(1)}\| + \|x_{(1)} - x_{(k)}\| + \|q - x_{(k)}\| \\
&\leq \varepsilon + \sum_{i=1}^{k-1} \|x_{(i)} - x_{(i+1)}\| + \varepsilon \\
&\leq 2k\varepsilon \leq 2\varepsilon \mathrm{cv}_K(\varepsilon).
\end{aligned}$$

Since this bound holds for all $p, q \in K$, we get the announced bound on the diameter of K.

We are now in position to prove Lemma 2.

Proof (of Lemma 2) Let $\varepsilon \leq \rho/4$, and $x_1, \ldots, x_{\mathrm{cv}_M(\varepsilon)}$ be a minimal covering of M. According to Lemma 32, for all k,

$$Q(\mathscr{B}_M(x_k, \varepsilon)) \geq a_d f_{min} \varepsilon^d$$

for some $a_d > 0$. A straightforward packing argument [11, Section B.1] yields that the covering number of the support M of Q satisfies

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

for all $\varepsilon \leq \rho/4$, where $c_d = 2^d / a_d$. Applying this bound with $\varepsilon = \rho/4$, together with Lemma 33 yields

$$\begin{aligned}
\operatorname{diam}(M) &\leq 2\frac{\rho}{4} \mathrm{cv}_M\left(\frac{\rho}{4}\right) \\
&\leq \frac{\rho}{2} \frac{c_d}{f_{min}\left(\frac{\rho}{4}\right)^d} \\
&= \frac{C_d}{f_{min} \rho^{d-1}},
\end{aligned}$$

where $C_d = 2^{3d-1}/a_d$.

Now we allow for some outliers. We consider a random variable X with distribution P, that can be written as $X = V(Y+Z) + (1-V)X''$, with $\|Z\| \leq sh$, $s \leq 1/4$, such that $\mathbb{P}(V=1) = \beta$ and V is independent from (Y, Z, X''), Y has law Q in $\mathscr{G}_{D,d,f_{min},f_{max},\rho}$, and X'' has uniform distribution on $\mathscr{B}(0, K_0)$ (recall that K_0 is defined below Lemma 2). Note that $s = 0$ corresponds to the clutter noise case, whereas $\beta = 1$ corresponds to the additive noise case.

For a fixed point x, let $p(x,h)$ denote $P(\mathscr{B}(x,h))$. We have $\mathbb{P}(VY \in \mathscr{B}(x,(1-s)h)) \leq \mathbb{P}(VX \in \mathscr{B}(x,h)) \leq \mathbb{P}(VY \in \mathscr{B}(x,2h))$. Hence we may write

$$\beta q(x, 3/4h) + (1-\beta) q'(x,h) \leq p(x,h) \leq \beta q(x,2h) + (1-\beta) q'(x,h),$$

where $q(x,h) = Q(\mathscr{B}(x,h))$, and $q'(x,h) = (h/K_0)^D$. Bounds on the quantities above are to be found in the following lemma.

Lemma 34 *There exists $h_+(\rho, \beta, f_{min}, f_{max}, d) \leq \rho/\sqrt{12d}$ such that, if $h \leq h_+$, for every x such that $d(x,M) \leq h$, we have*
- *$\mathscr{B}(x,2h) \cap M \subset \mathscr{B}(\pi_M(x), 4h) \cap M$,*
- *$q(x,2h) \leq C_d f_{max} h^d$.*

Moreover, if $d(x,M) \leq h/\sqrt{2}$, we have
- *$\mathscr{B}(\pi_M(x), h/8) \cap M \subset \mathscr{B}(x, 3h/4)$,*
- *$c_d f_{min} h^d \leq q(x, 3h/4)$,*
- *$p(x,h) \leq 2\beta q(x,2h)$.*

Proof (of Lemma 34) Set $h_1(\rho) = \rho/(16\alpha)$, and let x be such that $d(x,M) \leq h$, and $h \leq h_1$. According to Proposition 26, $\mathscr{B}(x,2h) \cap M \subset \mathscr{B}\left(\pi_M(x), r_{2h}^+\right) \cap M$, with $r_{2h}^+ = \sqrt{(1 + 2\Delta/\rho)} r_{2h} \leq 2r_{2h} \leq 4h$. According to Proposition 30, if $y \in \mathscr{B}(\pi_M(x), 4h) \cap M$, then $d_M(\pi_M(x), y) \leq 4\alpha h \leq \rho/4$. Proposition 31 then yields $q(x,2h) \leq C_d f_{max} h^d$.

Now if $d(x,M) \leq h/\sqrt{2}$, $\mathscr{B}\left(\pi_M(x), r_{3h/4}^-\right) \cap M \subset \mathscr{B}(x, 3h/4) \cap M$ according to Proposition 26, with $r_{3h/4}^- = \sqrt{(1 - \Delta/\rho)} r_{3h/4} \geq r_{3h/4}/2 \geq h/8$. Since $\mathscr{B}_M(\pi_M(x), h/8) \subset \mathscr{B}(\pi_M(x), h/8) \cap M$, a direct application of Proposition 31 entails $c_d f_{min} h^d \leq q(x, 3h/4)$.

Applying Proposition 31 again, there exists $h_2(f_{min}, d, D, \beta, \rho)$ such that if $h \leq h_1 \wedge h_2$, then for any x such that $d(x,M) \leq h/\sqrt{2}$ we have $(1-\beta) q'(x,h) \leq \beta c_{d,f_{min}} h^d$, along with $q(x,2h) \geq q(x,3h/4) \geq c_{d,f_{min}} h^d$. We deduce that $p(x,h) \leq 2\beta q(x,2h)$. Taking $h_+ = h_1 \wedge h_2 \wedge \rho/\sqrt{12d}$ leads to the result.

C.2 Local Covariance Matrices

In this section we describe the shape of the local covariance matrices involved in tangent space estimation. Without loss of generality, the analysis will be conducted for $\hat{\Sigma}_1$ (at sample point X_1), abbreviated as $\hat{\Sigma}$. We further assume that $d(X_1,M) \leq h/\sqrt{2}$, $\pi_M(X_1) = 0$, and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$.

The two models (additive noise and clutter noise) will be treated jointly, by considering a random variable X of the form

$$X = V(Y+Z) + (1-V)X'',$$

where $\mathbb{P}(V=1) = \beta$ and V is independent from (Y,Z,X''), Y has distribution in $\mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$, $\|Z\| \leq \sigma$, and X'' has uniform law on $\mathscr{B}(0,K_0)$ (recall that K_0 is defined above Definition 4). For short we denote by s the quantity σ/h, and recall that we take $s \leq 1/4$, along with $h \leq h_+$ (defined in Lemma 34).

Let $U(X_i,h)$, $i = 2,\ldots,n$, denote $\mathbb{1}_{\mathscr{B}(X_1,h)}(X_i)$, let $Y_i \in M$ and Z_i such that $X_i = Y_i + Z_i$, with $\|Z_i\| \leq sh$, and let $V_2,\ldots V_n$ denote random variables such that $V_i = 1$ if X_i is drawn from the signal distribution (see page 5). It is immediate that the $(U(X_i,h),V_i)$'s are independent and identically distributed, with distribution $(U(X,h),V)$.

With a slight abuse of notation, we will denote by $\mathbb{P}$ and $\mathbb{E}$ conditional probability and expectation with respect to X_1. The following expectations will be of particular interest.

$$\begin{aligned} m(h) &= \mathbb{E}(XU(X,h)V)/\mathbb{E}(VU(X,h)), \\ \Sigma(h) &= \mathbb{E}(X-m(h))_\top (X-m(h))_\top^t U(X,h)V, \end{aligned}$$

where for any x in $\mathbb{R}^D$ $x_\top$ and $x_\perp$ denote respectively the projection of x onto T_0M and $T_0M^\perp$.

The following lemma gives useful results on both $m(h)$ and $\Sigma(h)$, provided that X_1 is close enough to M.

Lemma 35 *If $d(X_1,M) \leq h/\sqrt{2}$, for $h \leq h_+$, then*

$$\Sigma(h) = \left(\begin{array}{c|c} A(h) & 0 \\ \hline 0 & 0 \end{array}\right),$$

with

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

Furthermore,

$$\begin{aligned} \|m_\top(h)\| &\leq 2h, \\ \|m_\perp(h)\| &\leq \frac{2h^2}{\rho} + sh. \end{aligned}$$

Proof (of Lemma 35) Let $x = y+z$ be in $B(X_1,h)$, with $y \in M$ and $\|z\| \leq sh$. Since $s \leq 1/4$, $\|y\| \leq 2h$. According to Proposition 26 combined with Proposition 30, we may write, for $h \leq h_+$ and y in $\mathscr{B}(X_1,2h) \cap M$,

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

in local polar coordinates. Moreover, if $y \in \mathscr{B}(X_1,(1-s)h)$, then $x \in \mathscr{B}(X_1,h)$. Then, according to Proposition 26, we have $\mathscr{B}(\pi_M(X_1),r^-_{3h/4}) \cap M \subset \mathscr{B}(X_1,(1-s)h) \cap M$. Let u be a unit vector in T_0M. Then $\langle u, x - m_\top(h)\rangle^2 = \langle u, rv + R(r,v) + z - m_\top(h)\rangle^2$ $\geq \langle u, rv - m_\top(h)\rangle^2/2 - 3(R(r,v)+z)^2 \geq \langle u, rv - m_\top(h)\rangle^2/2 - 6r^4/(4\rho^2) - 6s^2h^2$ according to Proposition 30. Hence we may write

$$\begin{aligned} \langle Au, u\rangle &= \beta \int_{\mathscr{B}(X_1,h)\cap M} \langle u, rv + R(r,v) - m_\top(h)\rangle^2 J(r,v) f(r,v) dr dv \\ &\geq \beta f_{min} c_d \int_{r=0}^{r^-_{3h/4}} \int_{\mathscr{S}_{d-1}} r^{d-1} \left[\langle u, rv - m_\top(h)\rangle^2/2 - 3r^4/(2\rho^2) - 6s^2h^2\right] dr dv, \end{aligned}$$

according to Proposition 31 (bound on $J(r,v)$) and Proposition 26 (the geodesic ball $\mathscr{B}_M(\pi_M(X_1),r^-_{3h/4})$ is included in the Euclidean ball $\mathscr{B}(\pi_M(X_1),r^-_{3h/4}) \subset \mathscr{B}(X_1,(1-s)h) \cap M$). Then

$$\int_{r=0}^{r^-_{3h/4}} \int_{\mathscr{S}_{d-1}} \frac{r^{d-1}\langle u, rv - m_\top(h)\rangle^2}{2} dr dv \geq \int_{r=0}^{r^-_{3h/4}} \int_{\mathscr{S}_{d-1}} \frac{r^{d-1}\langle u, rv\rangle^2}{2} dr dv = \frac{\sigma_{d-1}}{2d} \int_{r=0}^{r^-_{3h/4}} r^{d+1} dr = \frac{\sigma_{d-1}(r^-_{3h/4})^{d+2}}{2d(d+2)},$$

where σ_{d-1} denotes the surface of the $d-1$-dimensional unit sphere. On the other hand,

$$\int_{r=0}^{r^-_{3h/4}} \int_{\mathscr{S}_{d-1}} \frac{3r^{d+3}}{2\rho^2} + 6s^2h^2 r^{d-1} dr dv = \sigma_{d-1}(r^-_{3h/4})^{d+2} \left(\frac{3(r^-_{3h/4})^2}{2(d+4)\rho^2} + \frac{6s^2h^2}{d}\right).$$

Since $r^-_{3h/4} \leq h \leq h_+ \leq \rho/\sqrt{12d}$, we conclude that

$$\langle Au, u\rangle \geq \beta c_d f_{min} (r^-_{3h/4})^{d+2} \geq \beta c_d f_{min} h^{d+2},$$

since, for $d(X_1, M) \leq h/\sqrt{2}$ and $h \leq h_+$, $r^-_{3h/4} \geq r_{3h/4}/2 \geq h/8$, according to Proposition 26.

Now, since for any $x = y + z \in \mathscr{B}(X_1, h)$, $y \in M \cap \mathscr{B}(0, 2h)$ and $\|z\| \leq sh$, we have $\|y_\perp\| \leq 2h^2/\rho$, according to Proposition 25. Jensen's inequality yields that $\|m(h)_\perp\| \leq 2h^2/\rho + sh$ and $\|m(h)_\top\| \leq \|m(h)\| \leq 2h$.

The following Lemma 36 is devoted to quantify the deviations of empirical quantities such as local covariance matrices, means and number of points within balls from their deterministic counterparts. To this aim we define $N_0(h)$ and $N_1(h)$ as the number of points drawn from respectively noise and signal in $\mathscr{B}(X_1, h) \cap M$, namely

$$N_0(h) = \sum_{i \geq 2} U(X_i, h)(1 - V_i),$$
$$N_1(h) = \sum_{i \geq 2} U(X_i, h) V_i.$$

Lemma 36 *Recall that* $h_0 = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{1/(d+1)}$ *(as defined page 15), and* $h_\infty = h_0^{(d+1)/d}$*, for* κ *to be fixed later.*

If $h_0 \leq h_+$ *and* $d(X_1, M) \leq h_+/\sqrt{2}$*, then, with probability larger than* $1 - 4\left(\frac{1}{n}\right)^{2/d+1}$*, the following inequalities hold, for all* $h \leq h_0$*.*

$$\begin{aligned} \tfrac{N_0(h)}{n-1} &\leq 2(1-\beta) q'(h) + \tfrac{10(2+2/d)\log n}{n-1}, \\ \tfrac{N_1(h)}{n-1} &\leq \quad 2\beta q(2h) + \tfrac{10(2+2/d)\log n}{n-1}. \end{aligned}$$

Moreover, for all $(h_\infty \vee \sqrt{2} d(X_1, M)) \leq h \leq h_0$*, and n large enough,*

$$\left\| \frac{1}{n-1} \sum_{i \geq 2} (X_i - m(h))_\top (X_i - m(h))_\top^t U(X_i, h) V_i - \Sigma(h) \right\|_{\mathscr{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(2h) h^2,$$
$$\frac{1}{n-1} \left\| \sum_{i \geq 2} (X_i - m(h))_\top U(X_i, h) V_i \right\|_{\mathscr{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(2h) h.$$

Proof (of Lemma 36) The first two inequalities are straightforward applications of Theorem 5.1 in [6]. The proofs of the two last results are detailed below. They are based on Talagrand-Bousquet's inequality (see, e.g., Theorem 2.3 in [8]) combined with the so-called peeling device.

Define $h_- = (h_\infty \vee \sqrt{2} d(X_1, M))$, where we recall that in this analysis X_1 is fixed, and let $f_{T,h}$ denote the function

$$f_{T,h}(x, v) = \left\langle T, (x - m(h))_\top (x - m(h))_\top^t U(x, h) v \right\rangle,$$

for $h_- \leq h \leq h_0$, T a $d \times d$ matrix such that $\|T\|_{\mathscr{F}} = 1$, x in $\mathbb{R}^D$, v in $\{0, 1\}$, and $\langle T, B\rangle = \mathrm{trace}(T^t A)$, for any square matrices T and A. Now we define the weighted empirical process

$$Z = \sup_{T,h} \sum_{i \geq 2} \frac{f_{T,h}(X_i, V_i) - \mathbb{E} f_{T,h}(X, V)}{r(h)},$$

with $r(h) = \beta q(2h) h^2$, along with the constrained empirical processes

$$Z(u) = \sup_{T, h \leq u} \sum_{i \geq 2} f_{T,h}(X_i, V_i) - \mathbb{E} f_{T,h}(X, V),$$

for $h_- \leq u \leq h_0$. Since $\|f_{T,h}\|_\infty \leq sup_{x \in M} \|x - m(h)\|^2 U(x, h) \leq 4h^2$, and

$$Var(f_{T,h}(X, V)) \leq \mathbb{E}\left(\|X - m(h)\|^2 U(X, h) V\right) \leq 16 \beta h^4 \mathbb{P}(VX \in \mathscr{B}(X_1, h) \leq 16 \beta h^4 \mathbb{P}(VY \in \mathscr{B}(X_1, 2h),$$

for $s \leq 1/4$, a direct application of Theorem 2.3 in [8] yields, with probability larger than $1 - e^{-x}$,

$$Z(u) \leq 3\mathbb{E}Z(u) + \sqrt{\frac{32\beta q(2u) u^4 x}{n-1}} + \frac{20u^2 x}{3(n-1)}.$$

To get a bound on $\mathbb{E}Z(u)$, we introduce some independent Rademacher random variables $\sigma_2,\dots,\sigma_n$, i.e. $\mathbb{P}(\sigma_j=1)=\mathbb{P}(\sigma_j=-1)=1/2$. With a slight abuse of notation, expectations with respect to the (X_i,V_i)'s and σ_i's, $i=2,\dots,n$, will be denoted by $\mathbb{E}_{(X,V)}$ and $\mathbb{E}_\sigma$ in what follows. According to the symmetrization principle (see, e.g., Lemma 11.4 in [7]), we have

$$\begin{aligned}
(n-1)\mathbb{E}Z(u) &\le 2\mathbb{E}_{(X,V)}\mathbb{E}_{\sigma_i}\sup_{h\le u,T}\sum_{i\ge 2}\left\langle T,\sigma_i V_i U(X_i,h)((X_i-m(h))_\top(X_i-m(h))_\top^t)\right\rangle\\
&\le 2\mathbb{E}_{(X,V)}\mathbb{E}_\sigma\sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_iU(X_i,h)X_iX_i^t,T\right\rangle+2\mathbb{E}_{(X,V)}\mathbb{E}_\sigma\sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_iU(X_i,h)X_im(h)^t,T\right\rangle\\
&\quad+2\mathbb{E}_{(X,V)}\mathbb{E}_\sigma\sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_iU(X_i,h)m(h)X_i^t,T\right\rangle+2\mathbb{E}_{(X,V)}\mathbb{E}_\sigma\sup_{h\le u,T}\sum_{i\ge 2}\sigma_i\left\langle V_iU(X_i,h)m(h)m(h)^t,T\right\rangle\\
&:= 2\mathbb{E}_{(X,V)}(E_1+E_2+E_3+E_4).
\end{aligned}$$

For a fixed sequence (X_i,V_i), $i=2,\dots,n$, we may write

$$\begin{aligned}
E_1 &\le \mathbb{E}_\sigma\sup_{h\le u}\left(\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}-\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}\right)+\sup_{h\le u}\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}\\
&:= E_{11}+E_{12}.
\end{aligned}$$

Jensen's inequality ensures that

$$\begin{aligned}
E_{12} &\le \sup_{h\le u}\sqrt{\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}^2}\\
&\le 4u^2\sqrt{N_1(u)},
\end{aligned}$$

hence

$$\mathbb{E}_{(X,V)}E_{12}\le 4u^2\sqrt{\beta(n-1)q(2u)}.$$

For the term E_{11}, note that, when $(X_i,V_i)_{i=2,\dots,n}$ is fixed, $\sup_{h\le u}\left(\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}-\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_iX_i^t\right\|_{\mathscr{F}}\right)$ is in fact a supremum of at most $N_1(u)$ processes. According to the bounded difference inequality (see, e.g., Theorem 6.2 of [7]), each of these processes is subGaussian with variance bounded by $16h^4N_1(u)$ (see Theorem 2.1 of [7]). Hence a maximal inequality for subGaussian random variables (see Section 2.5, p.31, of [7]) ensures that

$$E_{11}\le 4h^2\sqrt{2N_1(u)\log(N_1(u))}\le 4h^2\sqrt{2N_1(u)\log(n-1)}.$$

Hence $\mathbb{E}_{(X,V)}E_{11}\le 4h^2\sqrt{2\beta(n-1)q(2u)\log(n-1)}$. E_2 may also be decomposed as

$$\begin{aligned}
E_2 &= \mathbb{E}_\sigma\sup_{h\le u}\left\|\left(\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_i\right)m(h)^t\right\|_{\mathscr{F}}\\
&\le 2u\mathbb{E}_\sigma\sup_{h\le u}\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_i\right\|\\
&\le 2u\left(\mathbb{E}_\sigma\sup_{h\le u}\left(\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_i\right\|-\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_i\right\|\right)+\sup_{h\le u}\mathbb{E}_\sigma\left\|\sum_{i\ge 2}\sigma_iV_iU(X_i,h)X_i\right\|\right)\\
&:= 2u(E_{21}+E_{22}).
\end{aligned}$$

Jensen's inequality yields that $E_{22}\le 2u\sqrt{N_1(u)}$, and the same argument as for E_{11} (expectation of a supremum of $n-1$ subGaussian processes with variance bounded by $4u^2N_1(u)$) gives $E_{22}\le 2u\sqrt{2N_1(u)\log(n-1)}$. Hence

$$\mathbb{E}_{(X,V)}E_2\le 4u^2\sqrt{\beta(n-1)q(2u)}\left(\sqrt{2\log(n-1)}+1\right).$$

Similarly, we may write

$$\mathbb{E}_{(X,V)}E_3\le 4u^2\sqrt{\beta(n-1)q(u)}\left(\sqrt{2\log(n-1)}+1\right).$$

At last, we may decompose E_4 as

$$\begin{aligned}
E_4 &\le \mathbb{E}_\sigma 4u^2\sup_{h\le u}\left|\sum_{i\ge 2}V_iU(X_i,h)\right|\\
&\le 4u^2\left[\mathbb{E}_\sigma\sup_{h\le u}\left(\left|\sum_{i\ge 2}V_iU(X_i,h)\right|-\mathbb{E}_\sigma\left|\sum_{i\ge 2}V_iU(X_i,h)\right|\right)+\sup_{h\le u}\mathbb{E}_\sigma\left|\sum_{i\ge 2}V_iU(X_i,h)\right|\right]\\
&\le 4u^2\sqrt{N_1(u)}\left(\sqrt{2\log(n-1)}+1\right),
\end{aligned}$$

using the same argument. Combining all these terms leads to

$$\mathbb{E}Z(u) \leq \frac{32\sqrt{\beta q(2u)}}{\sqrt{n-1}}\left(\sqrt{2\log(n-1)}+1\right),$$

hence we get

$$\mathbb{P}\left(Z(u) \geq \frac{192\sqrt{2}u^2\sqrt{\beta q(2u)\log(n-1)}}{\sqrt{n-1}}\left(1+\frac{1}{48}\sqrt{\frac{x}{\log(n-1)}}\right)+\frac{20u^2x}{n-1}\right) \leq e^{-x}.$$

To derive a bound on the weighted process Z, we make use of the so-called peeling device (see, e.g., Section 13.7, p.387, of [7]). Set $p = \lceil \log(h_0/h_\infty)\rceil \leq 1+\log(h_0/h_\infty)$, so that $e^{-p}h_0 \leq h_-$. According to Lemma 34, if I_j denotes the slice $[e^{-j}h_0, e^{-(j-1)}h_0] \cap [h_-, h_0]$, then, for every h in I_j, we have

$$r(h) \geq r(h_{j-1})c_d\frac{f_{min}}{f_{max}},$$

where c_d depends only on the dimension, provided that $h_0 \leq h_+$. Now we may write

$$\begin{aligned}
&\mathbb{P}\left(Z \geq \frac{192 f_{max}\sqrt{2}}{f_{min}c_d\sqrt{\beta q(2h_-)(n-1)}}\left(1+\frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right)+\frac{20 f_{max}(x+\log(p))}{(n-1)\beta c_d f_{min} q(2h_-)}\right)\\
&\leq \sum_{j=1}^{p}\mathbb{P}\left(\sup_{T,h\in I_j}\frac{\sum_{i\geq 2} f_{T,h}(X_i,V_i)-\mathbb{E}f_{T,h}(X,V)}{r(h)} \geq \frac{192 f_{max}\sqrt{2}}{f_{min}c_d\sqrt{\beta q(2h_-)(n-1)}}\left(1+\frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right)+\frac{20 f_{max}(x+\log(p))}{(n-1)f_{min}c_d\beta q(2h_-)}\right)\\
&\leq \sum_{j=1}^{p}\mathbb{P}\left(Z(h_{j-1}) \geq \frac{192\sqrt{2}r(h_{j-1})}{\sqrt{\beta q(2h_-)(n-1)}}\left(1+\frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right)+\frac{20 r(h_{j-1})(x+\log(p))}{(n-1)\beta q(2h_-)}\right).
\end{aligned}$$

Since $q(2h_{j-1}) \geq q(2h_-)$, we deduce that

$$\begin{aligned}
\mathbb{P}\left(Z \geq \frac{192 f_{max}\sqrt{2}}{f_{min}c_d\sqrt{\beta q(2h_-)(n-1)}}\left(1+\frac{1}{48}\sqrt{\frac{x+\log(p)}{n-1}}\right)+\frac{20 f_{max}(x+\log(p))}{(n-1)c_d f_{min}\beta q(2h_-)}\right) &\leq pe^{-(x+\log(p))}\\
&= e^{-x}.
\end{aligned}$$

Now, according to Lemma 34, $\beta q(2h_-) \geq c_d\kappa\log n/(n-1)$. On the other hand, $p \leq 1+\log(h_0/h_\infty) \leq \log(\beta(n-1)/\kappa)/d \leq \log n/d$, for $\kappa \geq 1$. For n large enough, taking $x = (1+2/d)\log n$ in the previous inequality, we get

$$\mathbb{P}\left(Z \geq C_d\frac{f_{max}}{f_{min}\sqrt{\kappa}}\right) \leq \left(\frac{1}{n}\right)^{1+2/d}.$$

The last concentration inequality of Lemma 36 may be derived the same way, considering the functions

$$g_{T,h}(x,v) = \langle (x-m(h))U(x,h)v, T\rangle,$$

where T is an element of $\mathbb{R}^d$ satisfying $\|T\| \leq 1$.

C.3 Decluttering Rate

In this section we prove that, if the angle between tangent spaces is of order h, then we can distinguish between outliers and signal at order h^2. We recall that the slab $S(x,T,h)$ is the set of points y such that $\|\pi_T(y-x)\| \leq k_1 h$ and $\|\pi_{T^\perp}(y-x)\| \leq k_2 h^2$, k_1 and k_2 defined in Lemma 18, and where π_T denotes the orthogonal projection onto T.

Lemma 37 *Recall that $h_0 = \left(\kappa\frac{\log n}{\beta(n-1)}\right)^{1/(d+1)}$, and $h_\infty = h_0^{(d+1)/d}$. Let K be fixed, and k_1, k_2 defined accordingly from Lemma 18. If $h_0 \leq h_+$, for κ large enough (depending on d, ρ and f_{min}) and n large enough, there exists a threshold t such that, for all $h_\infty \leq h \leq h_0$, we have, with probability larger than $1-3\left(\frac{1}{n}\right)^{2/d+1}$,*

$$\begin{aligned}
X_1 \in M \quad \text{and} \quad \angle\left(T, T_{X_1}M\right) \leq Kh/\rho &\Rightarrow |S(X_1,T,h)\cap\{X_2,\dots,X_n\}| \geq t(n-1)h^d,\\
d(X_1,M) \geq h^2/\rho \quad \text{and} \quad \angle\left(T, T_{\pi(X_1)}M\right) \leq Kh/\rho &\Rightarrow |S(X_1,T,h)\cap\{X_2,\dots,X_n\}| < t(n-1)h^d,\\
d(X_1,M) \geq h/\sqrt{2} &\Rightarrow |S(X_1,T,h)\cap\{X_2,\dots,X_n\}| < t(n-1)h^d.
\end{aligned}$$

Proof (of Lemma 37) Suppose that $d(X_1,M) \geq h/\sqrt{2}$. Then, according to Lemma 18, $S(X_1,T,h) \subset \mathscr{B}(X_1,h/2)$, with $\mathscr{B}(X_1,h/2) \cap M = \emptyset$, hence $P_n(S(X_1,T,h)) \leq P_n(\mathscr{B}(X_1,h/2))$. Theorem 5.1 in [6] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1-\left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(\mathscr{B}(X_1,h/2)) \leq 2P(\mathscr{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1}.$$

Since $\log(n)/(n-1) \leq \beta h^d/\kappa$, we may write

$$\begin{aligned} P_n(S(X_1,T,h)) &\leq 2Q'(\mathscr{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1} \leq 2(1-\beta)\frac{h^D}{(2K_0)^D} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq (1-\beta)C_{d,D,\rho,f_{min}}h^{d+1} + \frac{4\,(2/d+1)\log(8n)}{n-1} \leq h^d\left((1-\beta)C_{d,D,\rho,f_{min}}h + \frac{C_d\beta}{\kappa}\right), \end{aligned}$$

for n large enough so that $h \leq 1$.

If $h/\sqrt{2} \geq d(X_1,M) \geq h^2/\rho$ and $\angle\left(T_{\pi(X_1)}M, T\right) \leq Kh/\rho$, then Lemma 18 provides a big slab $S'(x,T_{\pi(x)}M,h)$ so that $S(x,T,h) \subset S'(x,T_{\pi(x)}M,h)$ and $S'(x,T_{\pi(x)}M,h) \cap M = \emptyset$. Thus, $P_n(S(x,T,h)) \leq P_n(S'(x,T_{\pi(x)}M,h))$. An other application of Theorem 5.1 in [6] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1-\left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(S'(x,T_{\pi(x)}M,h)) \leq 2P(S'(x,T_{\pi(x)}M,h)) + \frac{4\,(2/d+1)\log(8n)}{n-1},$$

hence, denoting by ω_r the volume of the r-dimensional unit ball, we get

$$\begin{aligned} P_n(S(X_1,T,h)) &\leq 2Q'(\mathscr{B}(X_1,h/2)) + \frac{4\,(2/d+1)\log(8n)}{n-1} \leq \frac{2(1-\beta)\omega_d\omega_{D-d}}{K_0^D\omega_D}(k_1'h)^d(k_2'h^2)^{D-d} + \frac{4\,(2/d+1)\log(8n)}{n-1} \\ &\leq (1-\beta)C_{d,D,f_{min},\rho}h^{d+1} + \frac{4\,(2/d+1)\log(8n)}{n-1} \leq h^d\left((1-\beta)C_{d,D,\rho,f_{min}}h + \frac{C_d\beta}{\kappa}\right), \end{aligned}$$

when n is large enough.

Now, if $X_1 \in M$ and $\angle\left(T_{\pi(X_1)}M, T\right) \leq Kh/\rho$, Lemma 18 entails that $\mathscr{B}(X_1,k_3h) \cap M \subset S(X_1,T,h)$, hence $P_n(S(X_1,T,h)) \geq P_n(\mathscr{B}(X_1,k_3h)\cap M)$. A last application of Theorem 5.1 in [6] yields that, for all $h_\infty \leq h \leq h_0$, with probability larger than $1-\left(\frac{1}{n}\right)^{2/d+1}$,

$$P_n(\mathscr{B}(X_1,k_3h)\cap M) \geq \frac{1}{2}P(\mathscr{B}(X_1,k_3h)) - \frac{2\,(2/d+1)\log(8n)}{n-1}.$$

Thus we deduce that

$$\begin{aligned} P_n(S(X_1,T,h)) \geq \frac{\beta}{2}Q(\mathscr{B}(X_1,k_3h)) - \frac{2\,(2/d+1)\log(8n)}{n-1} &\geq \frac{\beta}{2}q(k_3h) - C_d\frac{\beta h^d}{\kappa} \\ &\geq h^d\left(\beta c_{d,f_{min},\rho} - C_d\frac{\beta}{\kappa}\right), \end{aligned}$$

according to Lemma 34 (since $k_3 \leq 1$). Choosing κ large enough (depending on d, ρ and f_{min}) and then n large enough leads to the result.

D Matrix Decomposition and Principal Angles

In this section we expose a standard matrix perturbation result, adapted to our framework. For real symmetric matrices, we let $\mu_i(\cdot)$ denote their i-th largest eigenvalue and $\mu_{min}(\cdot)$ the smallest one.

Theorem 38 (Sin θ theorem [15], this version from Lemma 19 in [2]) *Let $O \in \mathbb{R}^{D\times D}$, $B \in \mathbb{R}^{d\times d}$ be positive semi-definite symmetric matrices such that*

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

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). Then

$$\angle(T_0,T) \leq \frac{\sqrt{2}\|E\|_{op}}{\mu_{min}(B)}.$$

E Local PCA for Tangent Space Estimation and Decluttering

This section is dedicated to the proofs of Section 5. We begin with the case of additive noise (and no outliers), that is Proposition 15.

E.1 Proof of Proposition 15

Without loss of generality, the local PCA analysis will be conducted at base point X_1, the results on the whole sample then follow from a standard union bound. For convenience, we assume that $\pi_M(X_1) = 0$ and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. We recall that $X_i = Y_i + Z_i$, with $Y_i \in M$ and $\|Z_i\| \leq sh$, for $s \leq 1/4$. In particular, $\|X_1\| \leq \|Z_1\| \leq sh \leq h/4$.

We adopt the following notation for the local covariance matrix based on the whole sample $\mathbb{X}_n$.

$$\begin{aligned}\hat{\Sigma}(h) &= \tfrac{1}{n-1}\textstyle\sum_{j\geq 2}(X_j - \bar{X}(h))(X_j-\bar{X}(h))^t U(X_i,h),\\ \bar{X}(h) &= \tfrac{1}{N(h)}\textstyle\sum_{i\geq 2} X_i U(X_i,h),\\ N(h) &= \textstyle\sum_{i\geq 2} U(X_i,h).\end{aligned}$$

Note that the tangent space estimator $\texttt{TSE}(\mathbb{X}_n,h)_1$ is the space spanned by the first d eigenvectors of $\hat{\Sigma}(h)$. From now on we suppose that all the inequalities of Lemma 36 are satisfied, defining then a global event of probability larger than $1-4\left(\frac{1}{n}\right)^{2/d+1}$.

We consider $h = h_0 \leq h_+$, so that Lemma 34 and 35 hold. We may then decompose the local covariance matrix as follows.

$$\begin{aligned}\hat{\Sigma}(h) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i-m(h))^t U(X_i,h) - \frac{N(h)}{n-1}(\bar{X}(h)-m(h))(\bar{X}(h)-m(h))^t\\ &:= \hat{\Sigma}_1 + \hat{\Sigma}_2. \end{aligned} \tag{3}$$

The first term may be written as

$$\begin{aligned}\hat{\Sigma}_1 = \frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))(X_i-m(h))^t U(X_i,h) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))_\top (X_i-m(h))_\top^t U(X_i,h) + R_1\\ &= \Sigma(h) + R_1 + R_2,\end{aligned}$$

where

$$\Sigma(h) = \begin{pmatrix} A(h) & 0\\ 0 & 0\end{pmatrix}.$$

According to Lemma 35 (with $\beta = 1$), $\mu_{min}(A(h)) \geq c_d f_{min} h^{d+2}$. On the other hand, using Proposition 25 and Lemma 35 we may write

$$\begin{aligned}(n-1)\|R_1\|_{\mathscr{F}}/N(h) &\leq 2\sup_{y+z\in\mathscr{B}(X_1,h)}\|(y+z-m(h))_\top\|\|(y+z-m(h))_\perp\| + \sup_{y+z\in\mathscr{B}(X_1,h)}\|(y+z-m(h))_\perp\|^2\\ &\leq 2\sup_{y+z\in\mathscr{B}(X_1,h)}\|(y+z-m(h))\|\left(\|(y-m(h))_\perp\|+sh\right) + \sup_{y\in\mathscr{B}(0,2h)\cap M}\left(\|(y-m(h))_\perp\|+sh\right)^2\\ &\leq 8h\left(\frac{4h^2}{\rho}+2sh\right) + \left(\frac{4h^2}{\rho}+2sh\right)^2\\ &\leq \frac{34h^3}{\rho} + 20sh^2,\end{aligned}$$

since $h \leq h_+$ and $s \leq 1/4$. In addition, we can write

$$R_2 = \begin{pmatrix} R_2 & 0\\ 0 & 0\end{pmatrix},$$

with $\|R_2\|_{\mathscr{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} q(2h)h^2$ according to Lemma 36 (with $\beta = 1$).

In turn, the term $\hat{\Sigma}_2$ may be decomposed as

$$\hat{\Sigma}_2 = \begin{pmatrix} R_4 & 0\\ 0 & 0\end{pmatrix} + R_3,$$

with

$$\begin{aligned}\|R_4\|_{\mathscr{F}} &\leq \frac{N(h)}{n-1}\|(\bar{X}(h)-m(h))_\top\|\|(\bar{X}(h)-m(h))\| \\ &\leq \frac{2h}{n-1}\left\|\sum_{i\geq 2}(X_i-m(h))_\top U(X_i,h)\right\| \\ &\leq \frac{2C_d q(2h)h^2 f_{max}}{f_{min}\sqrt{\kappa}},\end{aligned}$$

according to Lemma 36. A similar bound on R_3 may be derived,

$$\begin{aligned}\|R_3\|_{\mathscr{F}} &\leq \frac{N(h)}{n-1}\|(\bar{X}(h)-m(h))_\perp\|(\bar{X}(h)-m(h))\| \\ &\leq \frac{4h}{n-1}\left\|\sum_{i\geq 2}(Y_i+Z_i-m(h))_\perp U(X_i,h)\right\| \\ &\leq \frac{8hN(h)\left(2h^2/\rho+sh\right)}{n-1} \\ &\leq \frac{N(h)h^2}{n-1}\left(\frac{16h}{\rho}+8s\right),\end{aligned}$$

according to Proposition 25 and Lemma 35. If we choose $h=\left(\kappa\frac{\log n}{n-1}\right)^{1/d}$, for κ large enough (depending on d, f_{min} and f_{max}), we have

$$\frac{\|R_2+R_4\|_{\mathscr{F}}}{\mu_{min}(A(h))} \leq 1/4.$$

Now, provided that $\kappa \geq 1$, according to Lemma 36, we may write

$$\frac{\|R_1+R_3\|_{\mathscr{F}}}{\mu_{min}(A(h))} \leq K_{f_{max},f_{min},d}\,(h/\rho+s),$$

which, for n large enough, leads to

$$\angle(T_0M,\hat{T}_{X_1}M) \leq \sqrt{2}K_{f_{max},f_{min},d}\,(h/\rho+s),$$

according to Proposition 38.

E.2 Proof of Proposition 19

The proof of Proposition 19 follows the same path as the derivation of Proposition 15, with some technical difficulties due to the outliers ($\beta<1$). We emphasize that in this framework, there is no additive noise ($\sigma=0$). As in the previous section, the analysis will be conducted for $X_1\in\mathbb{X}^{(k)}$, for some fixed $k\geq -1$, $k=-1$ referring to the initialization step. Results on the whole sample then follow from a standard union bound. As before, we assume that $\pi_M(X_1)=0$ and that T_0M is spanned by the d first vectors of the canonical basis of $\mathbb{R}^D$. In what follows, denote by $\hat{t}$ the map from $\mathbb{R}^D$ to $\{0,1\}$ such that $\hat{t}(X_i)=1$ if and only if X_i is in $\mathbb{X}^{(k)}$.

We adopt the following notation for the local covariance matrix based on $\mathbb{X}^{(k)}$ (after $k+1$ iterations of the outlier filtering procedure).

$$\begin{aligned}\hat{\Sigma}^{(k)}(h) &= \tfrac{1}{n-1}\textstyle\sum_{j\geq 2}(X_j-\bar{X}(h)^{(k)})(X_j-\bar{X}(h)^{(k)})^t U(X_i,h)\hat{t}(X_i),\\ \bar{X}^{(k)}(h) &= \tfrac{1}{N^{(k)}(h)}\textstyle\sum_{i\geq 2}X_i U(X_i,h)\hat{t}(X_i),\\ N^{(k)}(h) &= \textstyle\sum_{i\geq 2}U(X_i,h)\hat{t}(X_i).\end{aligned}$$

Also recall that we define $N_0(h)$ and $N_1(h)$ as the number of points drawn from respectively clutter and signal in $\mathscr{B}(X_1,h)\cap M$ (based on the whole sample $\mathbb{X}_n$). At last, we suppose that all the inequalities of Lemma 36 and Lemma 37 are satisfied, defining then a global event of probability larger than $1-7\left(\frac{1}{n}\right)^{2/d+1}$.

We recall that we consider $h_\infty \leq h \leq h_k$, $k \geq -1$ (with $h_{-1} = h_0$), and X_1 in $\mathbb{X}^{(k)}$ such that $d(X_1, M) \leq h/\sqrt{2}$. We may then decompose the local covariance matrix as

$$\begin{aligned}
\hat{\Sigma}^{(k)}(h) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)\hat{t}(X_i) - \frac{N^{(k)}(h)}{n-1}(\bar{X}^{(k)}(h) - m(h))(\bar{X}(h)^{(k)} - m(h))^t \\
&= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)\hat{t}(X_i)V_i(X_i) + \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)(1-V_i)\hat{t}(X_i) \\
&\qquad - \frac{N^{(k)}(h)}{n-1}(\bar{X}(h)^{(k)} - m(h))(\bar{X}(h)^{(k)} - m(h))^t, \\
&:= \hat{\Sigma}_1^{(k)} + \hat{\Sigma}_2^{(k)} + \hat{\Sigma}_3^{(k)}. \qquad (4)
\end{aligned}$$

The proof of Proposition 19 will follow by induction.

Initialization step ($k = -1$):

In this case $\mathbb{X}^{(k)} = \mathbb{X}_n$, $h = h_0$, $d(X_1, M) \leq h_0/\sqrt{2}$, and $\hat{t}$ is always equal to 1. Then the first term $\hat{\Sigma}_1^{(k)}$ of (4) may be written as

$$\begin{aligned}
\frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))(X_i - m(h))^t U(X_i,h)V_i &= \frac{1}{n-1}\sum_{i\geq 2}(X_i - m(h))_\top (X_i - m(h))_\top^t U(X_i,h)V_i + R_1 \\
&= \Sigma(h) + R_1 + R_2,
\end{aligned}$$

where

$$\Sigma(h) = \begin{pmatrix} A(h) & 0 \\ 0 & 0 \end{pmatrix}.$$

According to Lemma 35, $\mu_{min}(A(h)) \geq c_d f_{min} \beta h^{d+2}$, and $\|R_1\|_{\mathscr{F}} \leq 34\frac{N_1(h)h^3}{\rho(n-1)}$ according to Proposition 25. Moreover, we can write

$$R_2 = \begin{pmatrix} R_2 & 0 \\ 0 & 0 \end{pmatrix},$$

with $\|R_2\|_{\mathscr{F}} \leq C_d \frac{f_{max}}{f_{min}\sqrt{\kappa}} \beta q(2h) h^2$ according to Lemma 36.

Term $\hat{\Sigma}_2^{(k)}$ in inequality (4) may be bounded by

$$\|\hat{\Sigma}_2^{(k)}\|_{\mathscr{F}} \leq \frac{16h^2 N_0(h)}{n-1}.$$

In turn, term $\hat{\Sigma}_3^{(k)}$ may be decomposed as

$$\frac{N^{(k)}(h)}{n-1}(\bar{X}(h)^{(k)} - m(h))(\bar{X}(h)^{(k)} - m(h))^t = \begin{pmatrix} R_6 & 0 \\ 0 & 0 \end{pmatrix} + R_5,$$

with

$$\begin{aligned}
\|R_6\|_{\mathscr{F}} &\leq \frac{N(h)^{(k)}}{n-1}\|(\bar{X}(h)^{(k)} - m(h))_\top\|\|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1}\left(\|\sum_{i\geq 2}(X_i - m(h))_\top U(X_i,h)V_i\| + \|\sum_{i\geq 2}(X_i - m(h))_\top U(X_i,h)(1-V_i)\|\right) \\
&\leq \frac{4C_d \beta q(2h) h^2 f_{max}}{f_{min}\sqrt{\kappa}} + \frac{16h^2 N_0(h)}{n-1},
\end{aligned}$$

according to Lemma 36. We may also write

$$\begin{aligned}
\|R_5\|_{\mathscr{F}} &\leq \frac{N(h)^{(k)}}{n-1}\|(\bar{X}(h)^{(k)} - m(h))_\perp\|\|(\bar{X}(h)^{(k)} - m(h))\| \\
&\leq \frac{4h}{n-1}\left(\|\sum_{i\geq 2}(X_i - m(h))_\perp U(X_i,h)V_i\| + \|\sum_{i\geq 2}(X_i - m(h))_\perp U(X_i,h)(1-V_i)\|\right) \\
&\leq \frac{16N_1(h)h^3}{(n-1)\rho} + \frac{16N_0(h)h^2}{(n-1)},
\end{aligned}$$

according to Proposition 25 and Lemma 35. As in the additive noise case (see proof of Proposition 15), provided that κ is large enough (depending on d, f_{min}, and f_{max}), we have

$$\frac{\|R_2+R_6\|_{\mathscr{F}}}{\mu_{min}(A(h))} \leq 1/4.$$

Since $(n-1)h_0^d = \frac{\kappa \log n}{\beta h}$, if we ask $\kappa \geq \rho$, then for n large enough we eventually get

$$\frac{\|\hat{\Sigma}_2^{(k)}+R_1+R_5\|_{\mathscr{F}}}{\mu_{min}(A(h))} \leq K_{d,f_{min},f_{max},\beta}\frac{h_0}{\rho},$$

according to Lemma 36. Then, Proposition 38 can be applied to obtain

$$\angle(\mathtt{TSE}(\mathbb{X}^{(-1)},h_0)_1, T_{\pi(X_1)}M) \leq \sqrt{2}K^{(0)}_{d,f_{min},f_{max},\beta}h_0/\rho.$$

According to Lemma 37, we may choose κ large enough (with respect to $K=\sqrt{2}K^{(0)}$, d, f_{min} and ρ) and then a threshold t so that, if $X_1 \in M$, then $X_1 \in \mathbb{X}^{(0)}$, and if $d(X_1,M) \geq h_0^2/\rho$, then $X_1 \notin \mathbb{X}^{(0)}$.

Iteration step Now we assume that $k \geq 0$, and that $d(X_i,M) \geq h_k^2/\rho$ implies $\hat{t}(X_i)=0$, with $h_k = \left(\kappa\frac{\log n}{\beta(n-1)}\right)^{\gamma_k}$, γ_k being between $1/(d+1)$ and $1/d$. Let $h_\infty \leq h \leq h_k$, and suppose that $d(X_1,M) \leq h_k/\sqrt{2}$. As in the initialization step, $\hat{\Sigma}_1^{(k)}$ may be written as

$$\begin{pmatrix} A(h) & 0 \\ 0 & 0 \end{pmatrix} + R_1 + R_2,$$

with $\mu_{min}(A(h)) \geq c_d f_{min}\beta h^{d+2}$, $\|R_1\|_{\mathscr{F}} \leq 34\frac{N_1(h)h^3}{\rho(n-1)}$, and $\|R_2\|_{\mathscr{F}} \leq C_d\frac{f_{max}}{f_{min}\sqrt{\kappa}}\beta q(2h)h^2$.

We can decompose $\hat{\Sigma}_2$ as

$$\begin{aligned}\frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))(X_i-m(h))^t U(X_i,h)(1-V_i)\hat{t}(X_i) &= \frac{1}{n-1}\sum_{i\geq 2}(X_i-m(h))_\top(X_i-m(h))^t_\top U(X_i,h)(1-V_i)\hat{t}(X_i)+R_3\\ &= \begin{pmatrix} R_4 & 0 \\ 0 & 0\end{pmatrix}+R_3,\end{aligned}$$

with $\|R_4\|_{\mathscr{F}} \leq \frac{16N_0(h)h^2}{n-1}$ and $\|R_3\| \leq \frac{128N_0(h)hh_k^2}{(n-1)\rho}$, according to Proposition 27, for n large enough so that $h_0^2/\rho \leq h_\infty$. Term $\hat{\Sigma}_3^{(k)}$ may also be written as

$$\frac{N(h)^{(k)}}{n-1}(\bar{X}(h)^{(k)}-m(h))(\bar{X}(h)^{(k)}-m(h))^t = \begin{pmatrix} R_6 & 0 \\ 0 & 0\end{pmatrix}+R_5,$$

with

$$\begin{aligned}\|R_6\|_{\mathscr{F}} &\leq \frac{N(h)^{(k)}}{n-1}\|(\bar{X}(h)^{(k)}-m(h))_\top\|\|(\bar{X}(h)^{(k)}-m(h))\|\\ &\leq \frac{4h}{n-1}\left(\|\sum_{i\geq 2}(X_i-m(h))_\top U(X_i,h)V_i\|+\|\sum_{i\geq 2}(X_i-m(h))_\top U(X_i,h)(1-V_i)\hat{t}(X_i)\|\right)\\ &\leq \frac{4C_d\beta q(2h)h^2 f_{max}}{f_{min}\sqrt{\kappa}}+\frac{16h^2N_0(h)}{(n-1)},\end{aligned}$$

according to Lemma 36. We may also write

$$\begin{aligned}\|R_5\|_{\mathscr{F}} &\leq \frac{N(h)^{(k)}}{n-1}\|(\bar{X}(h)^{(k)}-m(h))_\perp\|\|(\bar{X}(h)^{(k)}-m(h))\|\\ &\leq \frac{4h}{n-1}\left(\|\sum_{i\geq 2}(X_i-m(h))_\perp U(X_i,h)V_i\|+\|\sum_{i\geq 2}(X_i-m(h))_\perp U(X_i,h)(1-V_i)\hat{t}(X_i)\|\right)\\ &\leq \frac{16N_1(h)h^3}{(n-1)\rho}+\frac{32N_0(h)hh_k^2}{\rho(n-1)},\end{aligned}$$

according to Proposition 25, Proposition 27 and Lemma 35. As done before, we may choose κ large enough (depending on d, f_{min} and f_{max}, but not on k) such that

$$\frac{\|R_2+R_4+R_6\|_{\mathscr{F}}}{\mu_{min}(A(h))} \leq 1/4.$$

Now choose $h = h_{k+1} = \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{(2\gamma_k+1)/(d+2)}$, with $\kappa \geq 1$. This choice is made to optimize residual terms of the form $h/\rho + h_k^2 N_0(h)/h$ coming from $\|R_1+R_3+R_5\|_{\mathscr{F}}/\mu_{min}(A(h_{k+1}))$. Then we get, according to Lemma 36,

$$\begin{aligned}\frac{\|R_1+R_3+R_5\|_{\mathscr{F}}}{\mu_{min}(A(h_{k+1}))} &\leq C_d \frac{f_{max} h_{k+1}}{\rho f_{min}} + \frac{C'_d}{\beta \rho f_{min}} \left(\kappa \frac{\log n}{\beta(n-1)}\right)^{\gamma_{k+1}+2\gamma_k-(2\gamma_k+1)+1} \\ &\leq K_{d,f_{min},f_{max},\beta} \frac{h_{k+1}}{\rho},\end{aligned} \tag{5}$$

where again, $K_{d,f_{min},f_{max},\beta}$ does not depend on k. At last, we may apply Proposition 38 to get

$$\begin{aligned}\angle(\mathtt{TSE}(\mathbb{X}^{(k)}, h_{k+1})_1, T_{\pi(X_1)}M) \leq \sqrt{2} K_{d,f_{min},f_{max},\beta} h_{k+1}/\rho &\leq \sqrt{2}\left(K_{d,f_{min},f_{max},\beta} \vee K^{(0)}_{d,f_{min},f_{max},\beta}\right) h_{k+1}/\rho \\ &:= C_{d,\beta,f_{max},f_{min}} h_{k+1}/\rho.\end{aligned}$$

Then, according to Lemma 37, we may choose κ large enough (not depending on k) and t (not depending on k either) so that if $X_1 \in M$, then $X_1 \in \mathbb{X}^{(k+1)}$, and if $d(X_1, M) \geq h_k^2/\rho$, then $X_1 \notin \mathbb{X}^{(k+1)}$. Proposition 19 then follows from a straightforward union bound on the sample $\{X_1, \ldots, X_n\}$.

E.3 Proof of Proposition 22

In this case, we have $d(X_j, M) \leq h_\infty^2/\rho$, for every X_j in $\mathbb{X}^{(\hat{k})}$. The proof of Proposition 22 follows from the same calculation as in the proof of Proposition 19, replacing h_k^2/ρ by its upper bound h_∞^2/ρ and taking $h_{k+1} = h_\infty$ in the iteration step.

F Proof of the Main Reconstruction Results

We now prove main results Theorem 7 (additive noise model), and Theorems 8 and 9 (clutter noise model).

F.1 Additive Noise Model

Proof (of Corollary 16) Let $Q \in \mathscr{G}_{D,d,f_{min},f_{max},\rho,\sigma}$. Write $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \rho\sigma/h)$ for $c_{d,f_{min},f_{max}}$ large enough, an consider the event A defined by

$$A = \left\{\max_{X_j \in \mathbb{X}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq C_{d,f_{min},f_{max}}\left(\frac{h}{\rho} + \frac{\sigma}{h}\right)\right\} \cap \left\{\sup_{x \in M} d(x, \mathbb{X}_n) \leq \sigma\right\} \cap \left\{\sup_{X_j \in \mathbb{X}_n} d(X_j, M) \leq C_{d,f_{min}}\left(\frac{\log n}{n}\right)^{1/d}\right\}.$$

Then from Proposition 15 and Lemma 32, $\mathbb{P}_Q(A) \geq 1 - 5\left(\frac{1}{n}\right)^{2/d}$, and from the definition of ε and the construction of $\mathbb{Y}_n$, for n large enough,

$$\begin{aligned}A &\subset \left\{\max_{X_j \in \mathbb{X}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\rho}\right\} \cap \left\{\sup_{x \in M} d(x, \mathbb{X}_n) \leq \varepsilon\right\} \cap \left\{\sup_{X_j \in \mathbb{X}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\rho}\right\} \\ &\subset \left\{\max_{X_j \in \mathbb{Y}_n} \angle(T_{\pi_M(X_j)}M, \hat{T}_j(h)) \leq \frac{\varepsilon}{2280\rho}\right\} \cap \left\{\sup_{x \in M} d(x, \mathbb{Y}_n) \leq 2\varepsilon\right\} \cap \{\mathbb{Y}_n \text{ is } \varepsilon\text{-sparse}\} \cap \left\{\sup_{X_j \in \mathbb{Y}_n} d(X_j, M) \leq \frac{\varepsilon^2}{1140\rho}\right\},\end{aligned}$$

which yields the result.

Proof (of Theorem 7) Following the above notation, we observe that on the event A, Theorem 14 holds for $\varepsilon = c_{d,f_{min},f_{max}}(h \vee \rho\sigma/h)$, $\theta = \varepsilon/(1140\rho)$ (where we used that $\theta \leq 2\sin\theta$) and $\eta = \varepsilon^2/(1140\rho)$ with high probability, so that the first part of Theorem 7 is proved. Furthermore, for n large enough,

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

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

F.2 Clutter Noise Model

Proof (of Corollary 20) Let $P \in \mathscr{O}_{D,d,f_{min},f_{max},\rho,\beta}$. For n large enough, write $\varepsilon = c_{d,f_{min}f_{max}} h_{k_\delta}$ for $c_{d,f_{min}f_{max}}$ large enough, and consider the event

$$A^\delta = \left\{ \max_{X_j \in \mathbb{X}^{(k_\delta)}} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq C_{d,f_{min},f_{max}} \frac{h_{k_\delta}}{\rho} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}^{(k_\delta)}) \leq \frac{h_{k_\delta}^2}{\rho} \right\} \cap \left\{ \sup_{X_j \in \mathbb{X}^{(k_\delta)}} d(X_j, M) \leq C_{d,f_{min}} \left(\frac{\log n}{n} \right)^{1/d} \right\}.$$

From Proposition 19 and Lemma 32, $\mathbb{P}_P(A^\delta) \geq 1 - 8\left(\frac{1}{n}\right)^{2/d}$ and from the definition of ε and the construction of $\mathbb{Y}_n^\delta$, for n large enough,

$$\begin{aligned} A^\delta &\subset \left\{ \max_{X_j \in \mathbb{X}^{(k_\delta)}} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\rho} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{X}^{(k_\delta)}) \leq \varepsilon \right\} \cap \left\{ \sup_{X_j \in \mathbb{X}^{(k_\delta)}} d(X_j, M) \leq \frac{\varepsilon^2}{1140\rho} \right\} \\ &\subset \left\{ \max_{X_j \in \mathbb{Y}_n^\delta} \angle(T_{\pi_M(X_j)}M, \hat{T}_j^\delta) \leq \frac{\varepsilon}{2280\rho} \right\} \cap \left\{ \sup_{x \in M} d(x, \mathbb{Y}_n^\delta) \leq 2\varepsilon \right\} \cap \{\mathbb{Y}_n \text{ is } \varepsilon\text{-sparse}\} \cap \left\{ \sup_{X_j \in \mathbb{Y}_n^\delta} d(X_j, M) \leq \frac{\varepsilon^2}{1140\rho} \right\}, \end{aligned}$$

which yields the result.

Proof (of Theorem 8) Following the above notation, we observe that on the event A^δ, Theorem 14 holds for $\varepsilon = c_{d,f_{min}f_{max}} h_{k_\delta}$, $\theta = \varepsilon/(1140\rho)$ and $\eta = \varepsilon^2/(1140\rho)$, so that the first part of Theorem 8 is proved. As a consequence, for n large enough,

$$\begin{aligned} \mathbb{E}_P\left[d_H\left(M, \hat{M}_{\mathtt{TDC}\delta}\right)\right] &= \mathbb{E}_P\left[d_H\left(M, \hat{M}_{\mathtt{TDC}\delta}\right) \mathbb{1}_{A^\delta}\right] + \mathbb{E}_P\left[d_H\left(M, \hat{M}_{\mathtt{TDC}\delta}\right) \mathbb{1}_{(A^\delta)^c}\right] \\ &\leq C_d \frac{\varepsilon^2}{\rho} + \left(1 - \mathbb{P}_P(A^\delta)\right) \times 2K_0 \\ &\leq C'_{d,f_{min},f_{max},\rho} \varepsilon^2, \end{aligned}$$

where for the second line we used the fact that $M \cup \hat{M}_{\mathtt{TDC}\delta} \subset \mathscr{B}_0$, a ball of radius $K_0 = K_0(d, f_{min}, \rho)$.

Finally, Theorem 9 is obtained similarly using Proposition 22.