Reconstruction simpliciale de variétés via l'estimation d'espaces tangents

EDDIE AAMARI

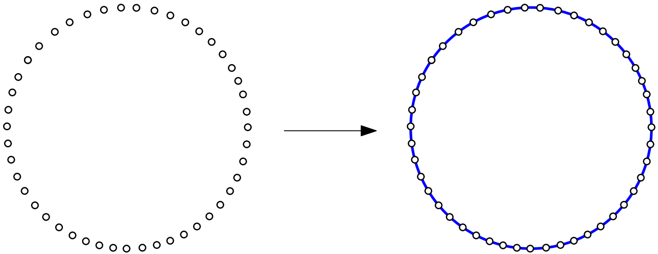
INRIA SACLAY, UNIVERSITÉ D'ORSAY

JOURNÉES MAS 2016, GRENOBLE

30/08/2016

COLLABORATION AVEC CLÉMENT LEVRARD (PARIS DIDEROT)

Manifold Reconstruction

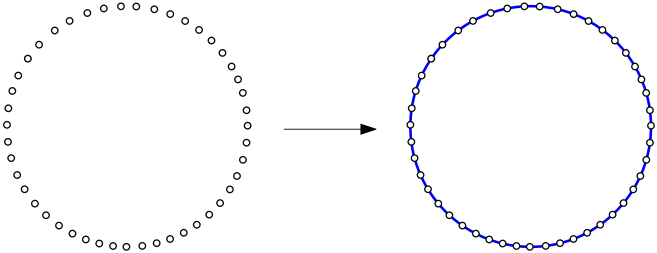


Input: observations $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ drawn *i.i.d.* on/nearby a manifold $M \subset \mathbb{R}^D$.

Goal: to give an estimator $\hat{M} \subset \mathbb{R}^D$ achieving

- topological guarantees.
- a good geometric approximation

Manifold Reconstruction



Input: a point cloud $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ drawn *i.i.d.* on/nearby a manifold $M \subset \mathbb{R}^D$.

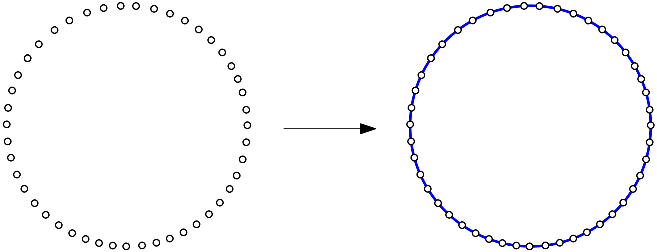
Goal: to give an estimator $\hat{M} \subset \mathbb{R}^D$ with:

- $\hat{M}$ isotopic to M ($\Rightarrow$ homeomorphic)
- Rates of convergence for the Hausdorff distance

$$\mathrm{d_H}(M, \hat{M}) = \left\| \mathrm{d}(\cdot, M) - \mathrm{d}(\cdot, \hat{M}) \right\|_\infty,$$

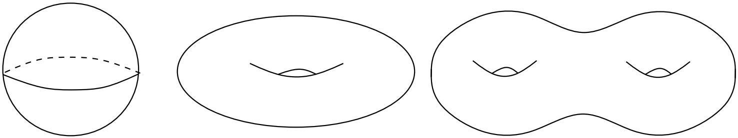
where $\mathrm{d}(x, K) = \inf_{p \in K} \|x - p\|$ is the distance to $K \subset \mathbb{R}^D$.

Manifold Reconstruction



Why ?

- Non-linear dimension reduction.
- Recover global data structure information: topology.



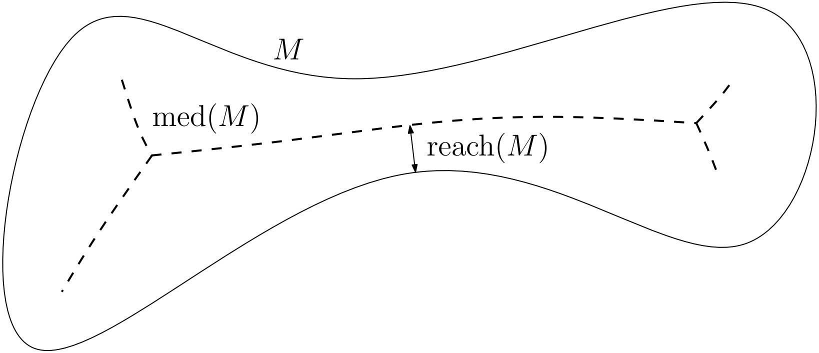
Regularity Assumption

$M \subset \mathbb{R}^D$ a d-dimensional submanifold.

The *reach* of M is the minimal distance to its *medial axis*:

$$\mathrm{reach}(M) = \inf_{x \in M} \mathrm{d}(x, \mathrm{med}(M)),$$

$$\mathrm{med}(M) = \{p \in \mathbb{R}^D, p \text{ has several nearest neighbors on } M\}.$$



Reach Condition

Assume $\mathrm{reach}(M) \geq \rho$ for some fixed $\rho > 0$

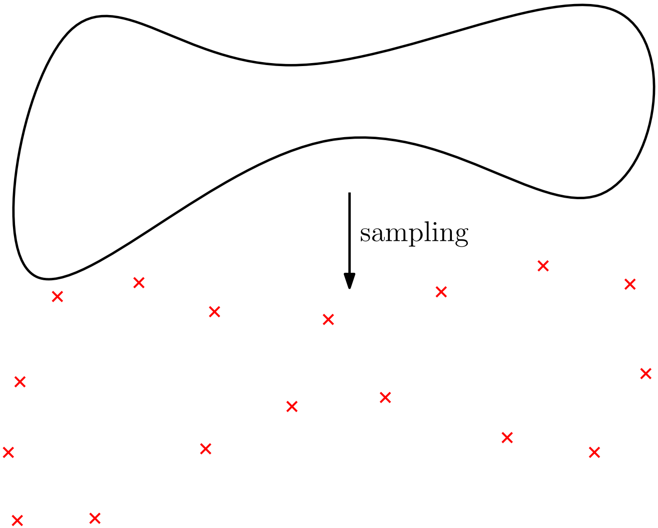


Figure : Reach and sampling

Reach Condition

Assume $\mathrm{reach}(M) \geq \rho$ for some fixed $\rho > 0$

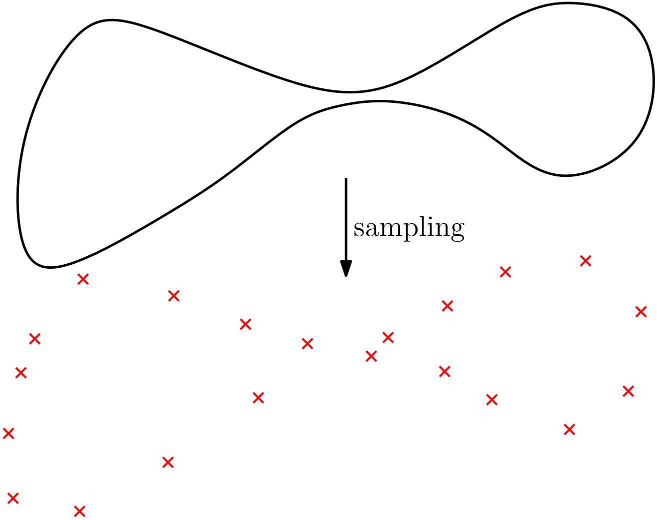


Figure : Reach and sampling

Building the estimator

Fix a finite set $\mathcal{P} \subset \mathbb{R}^D$.

Figure : Sample points

Building the estimator

For $p \in \mathcal{P}$, the Voronoi cell $\mathrm{Vor}(p)$ is defined as

$$\mathrm{Vor}(p) = \{x \in \mathbb{R}^D : \|x - p\| \leq \|x - q\|, \forall q \in \mathcal{P}\}.$$

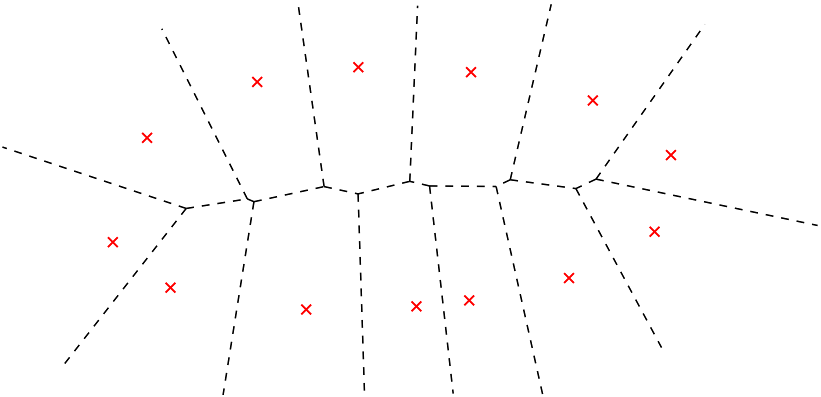


Figure : Voronoi diagram

Building the estimator

For a simplex $\tau \subset \mathcal{P}$, $\mathrm{Vor}(\tau) = \bigcap_{p \in \tau} \mathrm{Vor}(p)$.

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

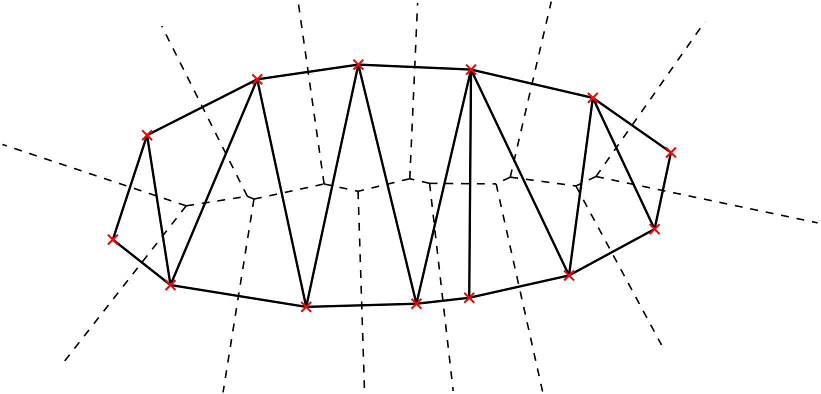


Figure : Delaunay complex

Building the estimator

For a simplex $\tau \subset \mathcal{P}$,

$$\tau \in \mathrm{Del}(\mathcal{P}, T) \Leftrightarrow \mathrm{Vor}(\tau) \cap \left(\bigcup_{p \in \tau} T_p M \right) \neq \emptyset.$$

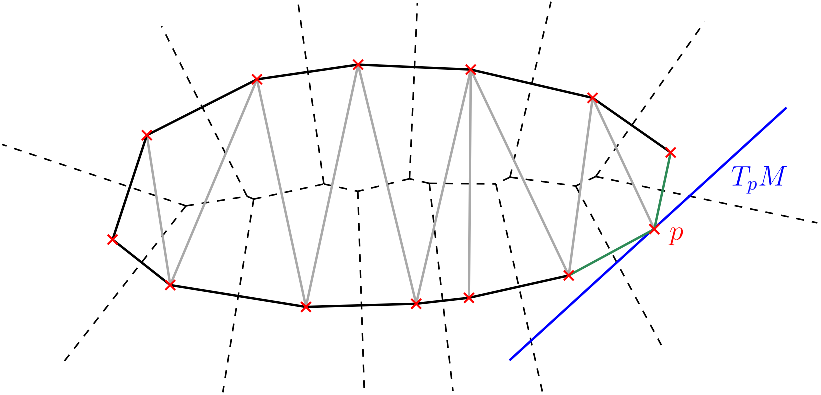


Figure : Tangential Delaunay complex [Boissonnat,Ghosh 2014]

A Reconstruction Theorem

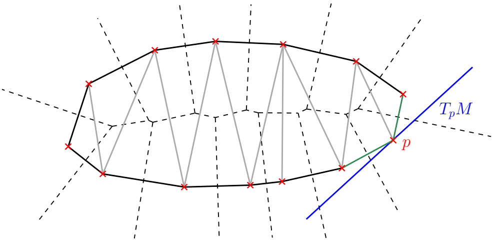
Theorem (Boissonnat,Ghosh 2014)

There exists $\varepsilon_0 = \varepsilon_0(\rho)$ *such that for all* $\varepsilon \leq \varepsilon_0$*, if* $\mathcal{P} \subset M$ *is*

- 2ε*-dense:* $\quad \mathrm{d_H}(\mathcal{P}, M) \leq 2\varepsilon$,
- ε*-sparse:* $\quad \mathrm{d}(p, \mathcal{P} \setminus \{p\}) \geq \epsilon$ *for all* $p \in \mathcal{P}$,

there exists as computable perturbation $\mathrm{Del}^\omega(\mathcal{P}, T)$ *of* $\mathrm{Del}(\mathcal{P}, T)$ *such that:*

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



Stability

Theorem (A.,Levrard, 2016)

The result still holds if:

- ***Small Noise:*** *For all* $p \in \mathcal{P}, \mathrm{d}(p, M) \lesssim \varepsilon^2$.
- ***Approximate Tangent Spaces:*** *For all* $p \in \mathcal{P}$, *we use* $\hat{T}_p$ *instead of* T_pM, *with* $\angle(T_pM, \hat{T}_p) \lesssim \varepsilon$.

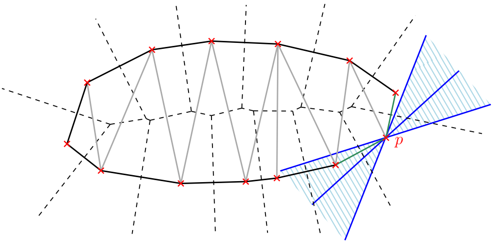


Figure : Tangent Space Stability

Statistical Model

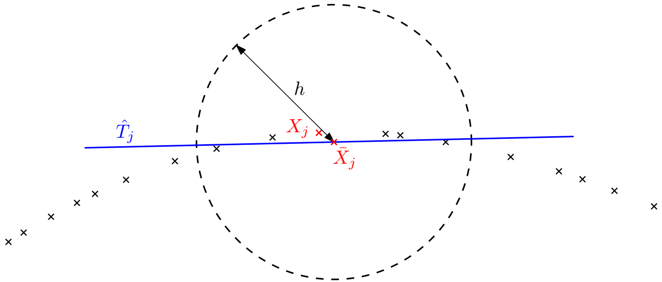
$X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$, where $M = \mathrm{supp}(P) \subset \mathbb{R}^D$ is a connected d-submanifold that satisfies:

- M has no boundary,
- $\mathrm{reach}(M) \geq \rho > 0$,
- P has a density f with respect to the uniform measure on M, with

$$0 < f_{min} \leq f(x) \leq f_{max} < \infty$$

Same model studied in *Minimax Manifold Estimation*, 2012 by Genovese, Perone-Pacifico, Verdinelli & Wasserman.

Tangent Space Estimation: Local P.C.A.

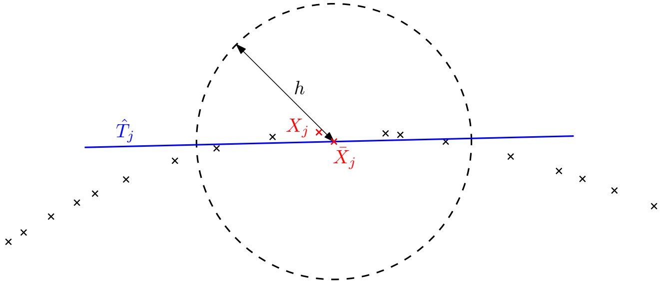


Define $\hat{T}_j$ as the span of the d first eigenvectors of

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

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

Tangent Space Estimation: Local P.C.A.



Theorem

Taking $h \asymp \left(\frac{\log n}{n}\right)^{1/d}$, for n large enough, with probability at least $1 - \left(\frac{1}{n}\right)^{2/d}$,

$$\begin{cases} \max_j \angle(T_{X_j}M, \hat{T}_j) \leq ch \\ \mathrm{d_H}\left(\mathbb{X}_n, M\right) \leq Ch. \end{cases}$$

Estimation Procedure & Convergence Rate

1. Estimate the $T_{X_j}M$'s with local PCA.
2. Take as estimator $\hat{M}$, the Tangential Delaunay Complex of $\mathbb{X}_n$ restricted to the estimated tangent spaces $\hat{T}_j$'s.

Theorem (A., Levrard 2015)

$$\lim_{n\to\infty} \mathbb{P}\left(\mathrm{d_H}(M,\hat{M}) \leq c\left(\frac{\log n}{n}\right)^{2/d} \text{ and } M \cong \hat{M}\right) = 1,$$

where $\cong$ denotes the isotopy equivalence.
Moreover, for n large enough,

$$\mathbb{E}\mathrm{d_H}(M,\hat{M}) \leq C\left(\frac{\log n}{n}\right)^{2/d}.$$

This rate is minimax optimal (Genovese *et al.* 2011).

A Noisy Model: Clutter Noise

$$X \sim \beta P + (1-\beta)\mathcal{U},$$

with $0 < \beta < 1$, P as previously and $\mathcal{U} \sim \mathit{Uniform}(\mathcal{B}_{\mathbb{R}^D})$.

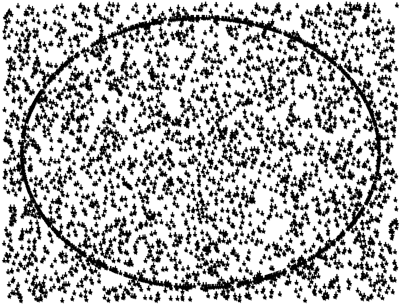
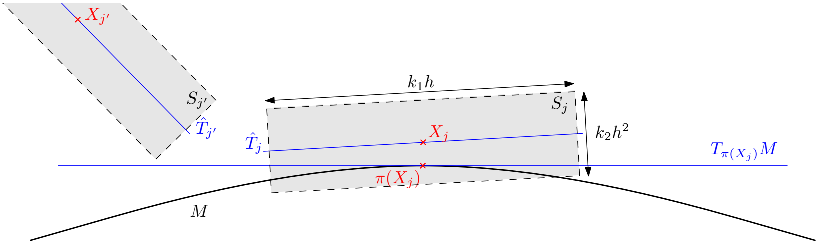


Figure : A realization of the clutter model

Clustering Before Estimation: Slab Denoising

We define boxes S_j centered at each X_j:



To determine if $X_j \in M$, consider $P_n(S_j) = |S_j \cap \{X_1, \ldots, X_n\}|$.
As $h \to 0$,

$$P_n(S_j) \sim \begin{cases} h^{2D-d} & \text{if } \quad X_j \text{ is far from } M \\ h^d \gg h^{2D-d} & \text{if } \quad X_j \in M \end{cases}$$

Clustering Result

Proposition

There exist constants $k(d, D, \beta)$ and $t(d, D, \rho)$ such that, for n large enough, if

$$h = k \left(\frac{\log n}{n} \right)^{\frac{1}{d+1}},$$

then, with probability larger than $1 - \left(\frac{1}{n}\right)^{\frac{2}{d}} - \left(\frac{1}{n}\right)^{2D}$*, we have*

$$\left(\frac{n}{\log n} \right) P_n(S_j) \begin{cases} \leq & t \quad \text{if} \quad d(X_j, M) \geq h^2 \\ > & t \quad \text{if} \quad X_j \in M \end{cases}$$

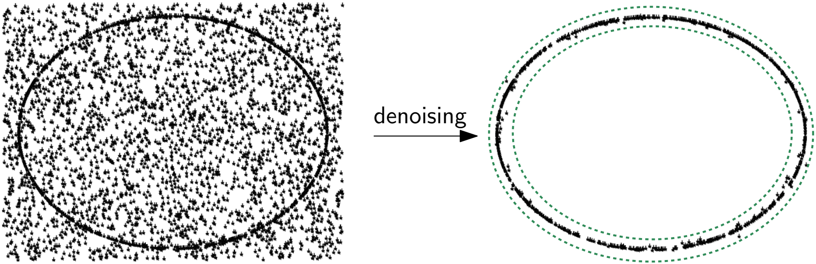
Moreover, on the same event, for every X_j *such that* $d(X_j, M) \leq Ch$*, we have*

$$\angle(\hat{T}_j, T_{\pi(X_j)} M) \leq ch$$

Clustering Result

Keeping the sample point X_{j_0} if and only if $P_n(S_{j_0}) > t_n$, w.h.p.

- no point $X_j \in M$ are removed;
- all false negative lie in a neighbourhood of M.



Convergence Result

1. Partition the sample into noise/data with slab counting,
2. Take $\hat{M}$ to be the Tangential Delaunay of the denoised points, restricted to the estimated tangent spaces $\hat{T}_j$'s.

Theorem (A., Levrard 2016)

$$\lim_{n\to\infty} \mathbb{P}\left(\mathrm{d_H}(M,\hat{M}) \leq c\left(\frac{\log n}{n}\right)^{2/(d+1)} \text{ and } M \cong \hat{M}\right) = 1,$$

where $\cong$ denotes the isotopy equivalence.
Moreover, for n large enough,

$$\mathbb{E}\mathrm{d_H}(M,\hat{M}) \leq C\left(\frac{\log n}{n}\right)^{2/(d+1)}.$$

Convergence Result

1. Partition the sample into noise/data with slab counting,
2. Take $\hat{M}$ to be the Tangential Delaunay of the denoised points, restricted to the estimated tangent spaces $\hat{T}_j$'s.

Theorem (A., Levrard 2016)

$$\lim_{n \to \infty} \mathbb{P}\left(\mathrm{d_H}(M, \hat{M}) \leq c \left(\frac{\log n}{n} \right)^{2/(d+1)} \text{ and } M \cong \hat{M} \right) = 1,$$

where $\cong$ denotes the isotopy equivalence.
Moreover, for n large enough,

$$\mathbb{E}\mathrm{d_H}(M, \hat{M}) \leq C \left(\frac{\log n}{n} \right)^{2/(d+1)}.$$

This rate is not minimax optimal (Genovese *et al.* 2011)

Iteration: Denoising + Tangent Space Estimation

We iterate $m \geq 1$ times the process of tangent space estimation + slab denoising with (appropriate) decreasing bandwidths.

Theorem (A., Levrard 2016)

If $m \geq C_d \log(1/\delta)$,

$$\lim_{n\to\infty} \mathbb{P}\left(\mathrm{d_H}(M, \hat{M}) \leq c\left(\frac{\log n}{n}\right)^{2/d-2\delta} \text{ and } M \cong \hat{M}\right) = 1,$$

where $\cong$ denotes the isotopy equivalence.
Moreover, for n large enough,

$$\mathbb{E}\mathrm{d_H}(M, \hat{M}) \leq C\left(\frac{\log n}{n}\right)^{2/d-2\delta}.$$

References

- Aamari, Levrard — Stability and Minimax Optimality of Tangential Delaunay Complexes for Manifold Reconstruction (Preprint)
- Boissonnat, Ghosh — Manifold reconstruction using tangential Delaunay complexes
- Genovese, Perone-Pacifico, Verdinelli, Wasserman — Minimax Manifold Estimation