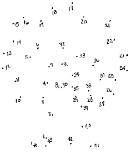
Simplicial Manifold Reconstruction *via* Tangent Space Estimation

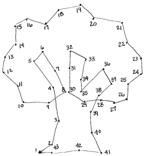
EDDIE AAMARI[1] CLÉMENT LEVRARD[2]

[1]INRIA Saclay, LMO [2]Université Paris Diderot

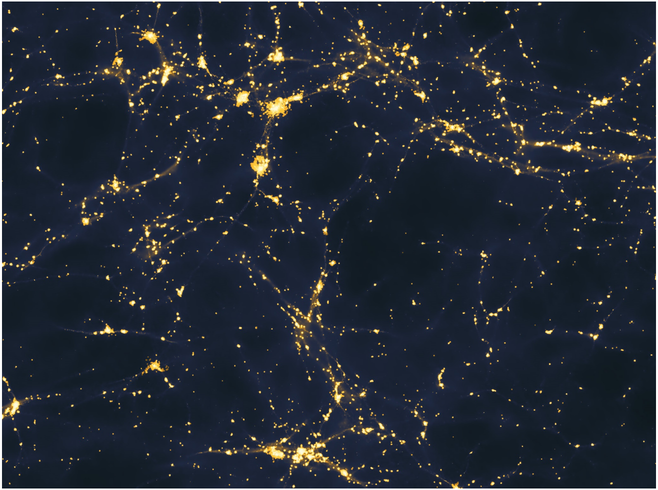
CIRM - Workshop Apprentissage

02/02/2016



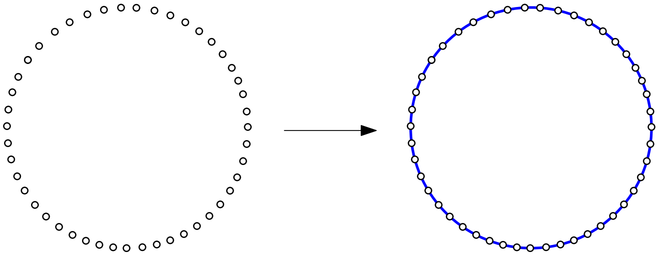


Motivation



"Large-scale structure of light distribution in the universe", Andrew Pontzen and Fabio Governato

Manifold reconstruction



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

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

- topological guarantees (homeomorphism),

- a good geometric approximation (Haussdorf distance).

A simplicial complex estimator

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

Figure: Sample points

A simplicial complex estimator

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

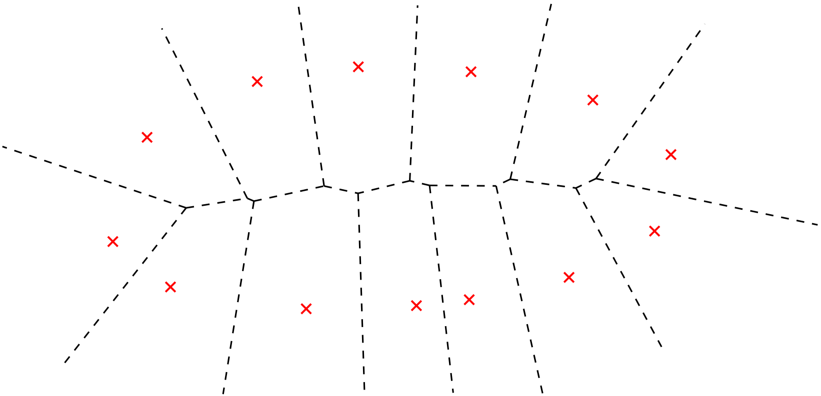


Figure: Voronoi diagram

A simplicial complex estimator

- $\tau = \{p_0, \ldots, p_k\}$ k-simplex,
- $\tau \in \mathrm{Del}(\mathcal{P})$ (Delaunay complex) iff $\bigcap_{p \in \tau} \mathrm{Vor}(p) \neq \emptyset$.

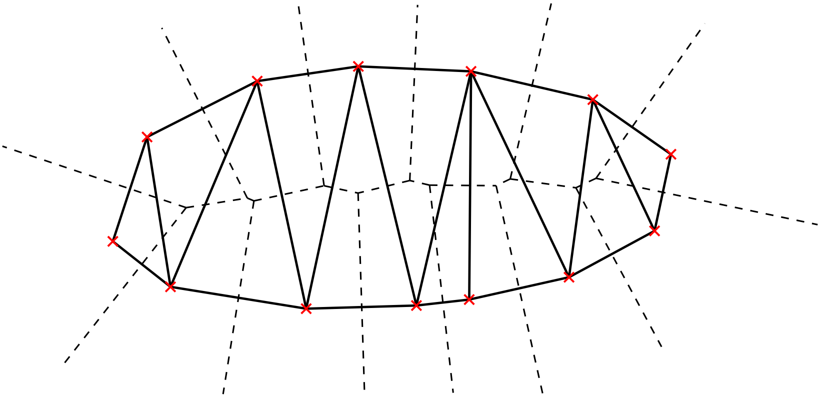


Figure: Delaunay complex

A simplicial complex estimator

- $\tau = \{p_1, \ldots, p_k\}$ k-simplex,
- $\tau \in \mathrm{Del}(\mathcal{P})$ (Delaunay complex) iff $\bigcap_{p \in \tau} \mathrm{Vor}(p) \neq \emptyset$,
- $\tau \in \mathrm{Del}(\mathcal{P}, T)$ iff $\bigcap_{p \in \tau} \mathrm{Vor}(p) \cap \left(\bigcup_{p \in \tau} T_p\mathcal{M} \right) \neq \emptyset$.

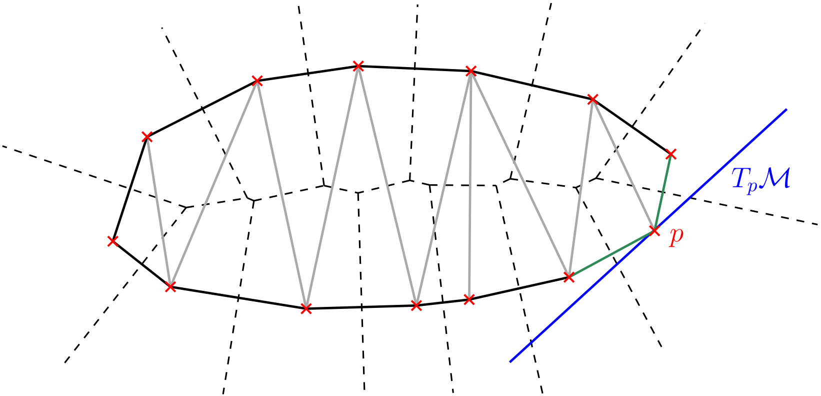
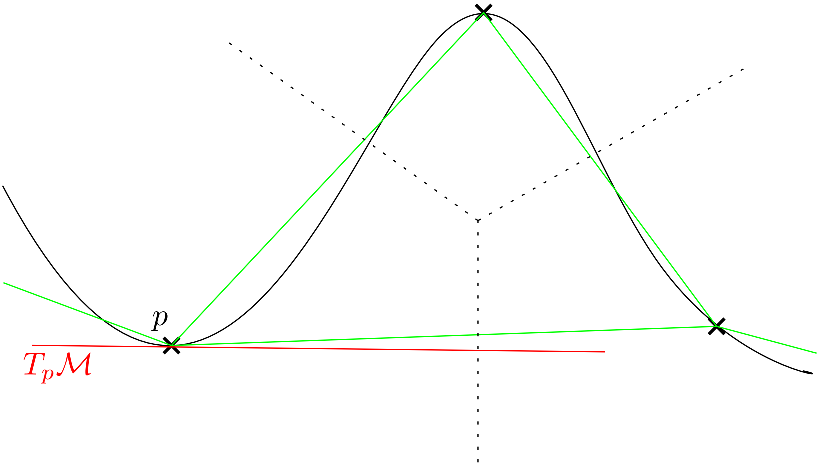


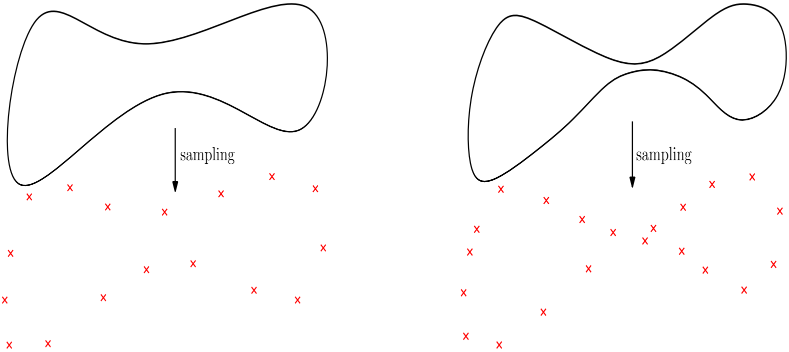
Figure: Tangential Delaunay complex [Boissonnat,Ghosh 2014]

Geometric condition



$\rightarrow$ Bound on curvature.

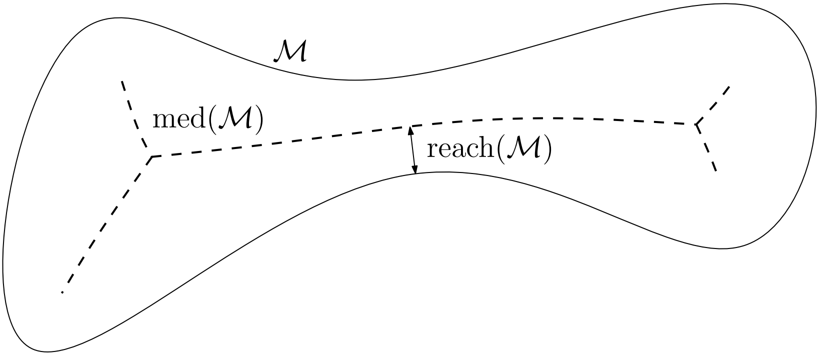
Geometric condition



→ No infinitely small "bottleneck".

Geometric condition

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



Geometric regularity condition: $\operatorname{reach}(\mathcal{M}) > 0$.

A Reconstruction Theorem

Theorem (Boissonnat,Ghosh 2014)

If $\mathrm{reach}(\mathcal{M}) > 0$*, there exists* ε_0 *such that for all* $\varepsilon \leq \varepsilon_0$*, if* $\mathcal{P} \subset \mathcal{M}$ *is*

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

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

- $\mathrm{Del}^\omega(\mathcal{P}, T)$ *and* $\mathcal{M}$ *are homeomorphic,*
- $\mathrm{d_H}\left(\mathrm{Del}^\omega(\mathcal{P}, T), \mathcal{M}\right) \leq C\varepsilon^2$*, where* $C = C(d)$.

A Reconstruction Theorem

Theorem (Boissonnat,Ghosh 2014)

If $\mathrm{reach}(\mathcal{M}) > 0$*, there exists* ε_0 *such that for all* $\varepsilon \leq \varepsilon_0$*, if* $\mathcal{P} \subset \mathcal{M}$ *is*

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

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

- $\mathrm{Del}^\omega(\mathcal{P}, T)$ *and* $\mathcal{M}$ *are homeomorphic,*
- $\mathrm{d_H}\left(\mathrm{Del}^\omega(\mathcal{P}, T), \mathcal{M}\right) \leq C\varepsilon^2$*, where* $C = C(d)$*.*

Problem:

- The $T_p\mathcal{M}$'s are unknown.

  $\Rightarrow$ We replace each $T_p\mathcal{M}$ by an estimated version $\hat{T}_p$.

- How to deal with noise?

Statistical Model

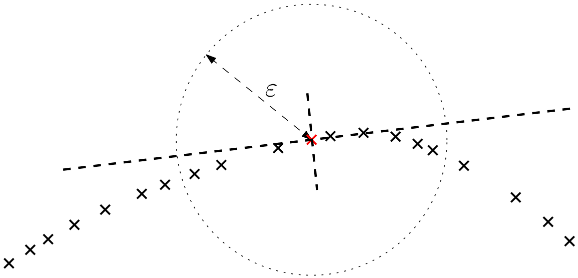
Geometric assumptions:

- $\mathcal{M}$ is a closed and connected d-submanifold of $\mathbb{R}^D$,
- $\mathrm{reach}(\mathcal{M}) := \rho > 0$.

Statistical assumptions: $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$,

- $P \sim f \mathrm{d}\lambda_{\mathcal{M}}$,
- $0 < f_{min} \leq f(x) \leq f_{max}$,

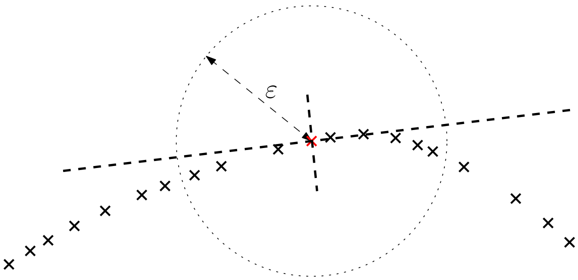
Tangent Space Estimation: Local PCA



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

$$\hat{O}_j = \frac{1}{n-1} \sum_{i \neq j} \mathbf{1}_{\|X_i - X_j\| \leq \varepsilon} \left(X_i - \bar{X}_j \right) \left(X_i - \bar{X}_j \right)^T .$$

Tangent Space Estimation: Local PCA



Proposition

Taking $\varepsilon \asymp \left(\frac{\log(n)}{n}\right)^{1/d}$*, for* n *large enough, yields, with probability larger than* $1 - \left(\frac{1}{n}\right)^{2/d}$,

$$\begin{cases} \max_j \angle(T_{X_j}\mathcal{M}, \hat{T}_j) & \leq & c\varepsilon \\ \mathrm{d_H}\left(\{X_1, \ldots, X_n\}, \mathcal{M}\right) & \leq & C\varepsilon. \end{cases}$$

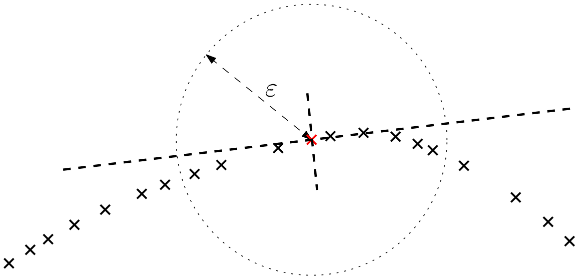
Tangent Space Estimation: Local PCA/Sketch of Proof

$$\hat{O}_j = \varepsilon^{d+2}\left[\left(\begin{array}{c|c} A > 0 & 0 \\ \hline 0 & 0 \end{array}\right) + Bias + \left(\begin{array}{c|c} Dev_{1,1} & Dev_{1,2} \\ \hline Dev_{2,1} & Dev_{2,2} \end{array}\right)\right]$$

$\rightarrow$ $Bias \lesssim \varepsilon/\rho$

$\rightarrow$ $\angle(T_{X_j}\mathcal{M}, \hat{T}_j) \approx Bias_{2,1} + Dev_{2,1}$ (for n large enough).

Tangent Space Estimation: Local PCA/Sketch of Proof



→ $\mathit{Bias} \lesssim \varepsilon/\rho$

→ $\angle(T_{X_j}\mathcal{M}, \hat{T}_j) \approx \mathit{Bias}_{2,1} + \mathit{Dev}_{2,1}$ (for n large enough).

→ $\mathit{Dev}_{2,1} \lesssim \frac{\varepsilon/\rho}{\sqrt{(n-1)\varepsilon^d}}$

What about $\mathrm{Del}(\mathcal{P}, \hat{T})$?

Two ways of resolution:

(1)

Prove that

$$\mathrm{Del}(\mathcal{P}, \hat{T}) = \mathrm{Del}(\mathcal{P}, T).$$

(2)

Find $\mathcal{M}' \cong \mathcal{M}$ such that

$$\mathrm{d_H}\left(\mathrm{Del}(\mathcal{P}, \hat{T}), \mathcal{M}'\right) \lesssim \varepsilon^2,$$

and

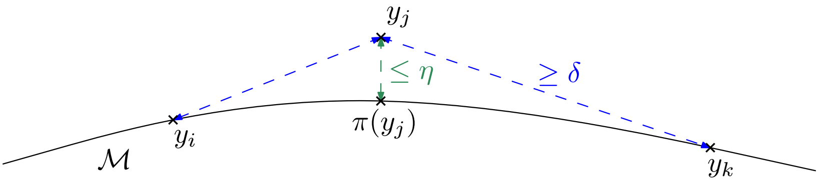
$$\mathrm{d_H}\left(\mathcal{M}', \mathcal{M}\right) \lesssim \varepsilon^2.$$

Interpolation Theorem

Theorem (Aamari, L. 2015)

Let $\mathbb{Y} = \{y_1, \ldots, y_q\} \subset \mathbb{R}^D$ and $T_1, \ldots, T_q$ be a collection of d-dimensional linear subspaces of $\mathbb{R}^D$.

- $\mathbb{Y}$ *is δ-sparse:* $\min_{i \neq j} \|y_j - y_i\| \geq \delta > 0$ *for all* j,
- *the* y_j*'s are η-close to $\mathcal{M}$:* $\max_{1 \leq j \leq q} \mathrm{d}(y_j, \mathcal{M}) < \eta$,
- $\max_{1 \leq j \leq q} \angle(T_{\pi(y_j)}\mathcal{M}, T_j) \leq \theta$.

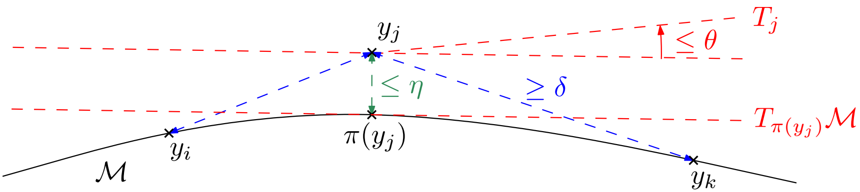


Interpolation Theorem

Theorem (Aamari, L. 2015)

Let $\mathbb{Y} = \{y_1, \ldots, y_q\} \subset \mathbb{R}^D$ *and* $T_1, \ldots, T_q$ *be a collection of d-dimensional linear subspaces of* $\mathbb{R}^D$*.*

- $\mathbb{Y}$ *is* δ*-sparse:* $\min_{i \neq j} \|y_j - y_i\| \geq \delta > 0$ *for all* j,
- *the* y_j*'s are* η*-close to* $\mathcal{M}$*:* $\max_{1 \leq j \leq q} \mathrm{d}(y_j, \mathcal{M}) < \eta$,
- $\max_{1 \leq j \leq q} \angle(T_{\pi(y_j)}\mathcal{M}, T_j) \leq \theta$.

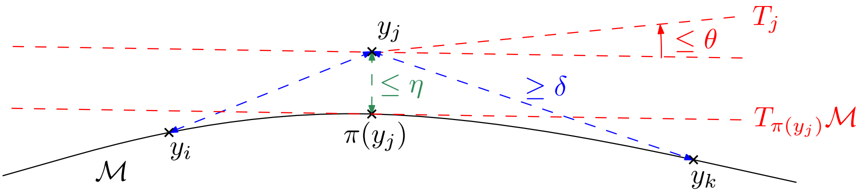


Interpolation Theorem

Theorem (Aamari, L. 2015)

If $\eta \asymp \delta^2 \ll 1$ and $\theta \asymp \delta$, there exists a smooth sub-manifold $\mathcal{M}' \subset \mathbb{R}^D$ and $C > 0$ such that

- *$\mathcal{M}' \supset \mathbb{Y}$ and $\mathcal{M}'$ has the T_j's as tangent spaces,*
- *$\mathrm{d_H}(\mathcal{M}, \mathcal{M}') \leq \eta + \delta\theta$,*
- *$\mathcal{M}$ and $\mathcal{M}'$ are ambient isotopic,*
- *$\mathrm{reach}(\mathcal{M}') \geq C\mathrm{reach}(\mathcal{M})$.*



Estimation Procedure & Convergence Rate

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

With $\varepsilon \asymp \left(\frac{\log(n)}{n}\right)^{\frac{1}{d}}$, we have

- $\mathrm{d_H}\left(\{X_j's\}, \mathcal{M}\right) \lesssim \varepsilon$
- $\max_j \angle(T_{X_j}\mathcal{M}, \hat{T}_j) \leq c\varepsilon$

Estimation Procedure & Convergence Rate

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

Theorem (Aamari, L. 2015)

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

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

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

- This rate is minimax optimal (Genovese 2011, Kim 2013)

A Noisy Model: Clutter Noise

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

with $0 < \beta < 1$, P as previously and $\mathcal{U} \sim \mathit{Uniform}(\mathcal{B}(0, M))$.

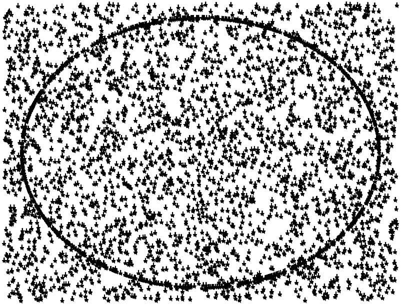
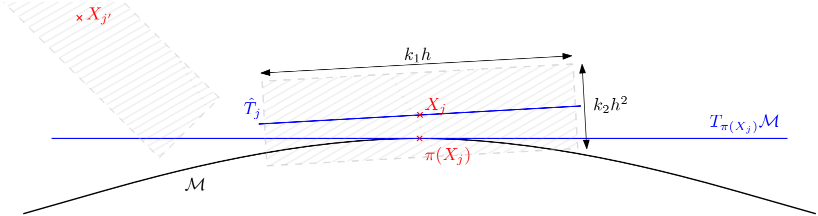


Figure: Clutter noise model

A denoising procedure

Define slabs S_j centered at each X_j:



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

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

Clustering Result

Proposition

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

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

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

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

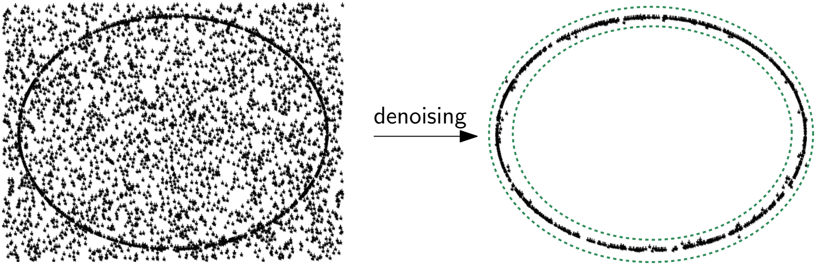
Moreover, on the same event, for every X_j such that $d(X_j, \mathcal{M}) \leq C\varepsilon$, we have

$$\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c\varepsilon.$$

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 $X_j \in \mathcal{M}$ are removed,
- all false negative lie in a neighbourhood of $\mathcal{M}$.



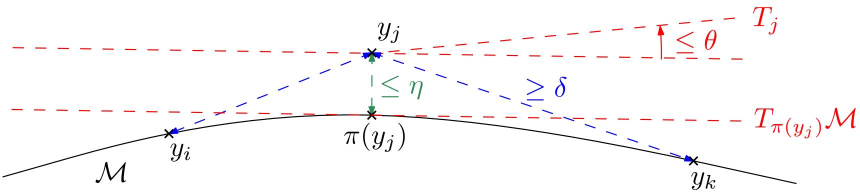
Convergence Result

1. Partition the sample into noise/data with slab counting,
2. Take as estimator $\hat{\mathcal{M}}$, the Delaunay triangulation of $\mathbb{Y}_n$ restricted to the estimated tangent spaces $\hat{T}_j$'s.

With $\varepsilon \asymp \left(\frac{\log(n)}{\beta n}\right)^{\frac{1}{d+1}}$, all remaining X_j's satisfy

- $d(X_j, \mathcal{M}) \leq \varepsilon^2$,
- $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c\varepsilon$,
- $\mathrm{d_H}(\{X_j's\}, \mathcal{M}) \leq \varepsilon$.

Convergence Result



With $\varepsilon \asymp \left(\frac{\log(n)}{\beta n}\right)^{\frac{1}{d+1}}$, all remaining X_j's satisfy

- $d(X_j, \mathcal{M}) \leq \varepsilon^2$,
- $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c\varepsilon$,
- $\mathrm{d_H}(\{X_j's\}, \mathcal{M}) \leq \varepsilon$.

Convergence Result

1. Partition the sample into noise/data with slab counting,
2. Take as estimator $\hat{\mathcal{M}}$, the Delaunay triangulation of $\mathbb{Y}_n$ restricted to the estimated tangent spaces $\hat{T}_j$'s.

Theorem (Aamari, L. 2015)

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

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

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

Current work: almost the true rate

Step 0 : Take $\varepsilon^{(0)} \asymp \left(\alpha_0 \frac{\log(n)}{\beta n}\right)^{\frac{1}{d+1}}$, then w.p $\geq 1 - p_0$, TSE + SD gives $\mathcal{D}^{(1)}$ such that

- $\rightarrow$ $X_j \in \mathcal{M} \Rightarrow X_j \in \mathcal{D}^{(1)}$,
- $\rightarrow$ $X_j \in \mathcal{D}^{(1)} \Rightarrow d(X_j, \mathcal{M}) \leq \varepsilon^{(0),2}$.

Current work: almost the true rate

Step 0 : Take $\varepsilon^{(0)} \asymp \left(\alpha_0 \frac{\log(n)}{\beta n}\right)^{\frac{1}{d+1}}$, then w.p $\geq 1 - p_0$, TSE + SD gives $\mathcal{D}^{(1)}$ such that

- $\rightarrow$ $X_j \in \mathcal{M} \Rightarrow X_j \in \mathcal{D}^{(1)}$,
- $\rightarrow$ $X_j \in \mathcal{D}^{(1)} \Rightarrow d(X_j, \mathcal{M}) \leq \varepsilon^{(0),2}$.

Step 1 : Take $\varepsilon^{(1)} \asymp \left(\alpha_1 \frac{\log(n)}{\beta n}\right)^{\gamma_1}$, with $(d+2)\gamma_1 = 2\gamma_0 + 1$, then w.p $\geq 1 - p_0 - p_1$, TPE *on* $\mathcal{D}^{(1)}$ gives

- $\rightarrow$ $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c\varepsilon^{(1)}$,

and TSE + SD gives $\mathcal{D}^{(2)}$ such that

- $\rightarrow$ $X_j \in \mathcal{M} \Rightarrow X_j \in \mathcal{D}^{(2)}$,
- $\rightarrow$ $X_j \in \mathcal{D}^{(2)} \Rightarrow d(X_j, \mathcal{M}) \leq \varepsilon^{(1),2}$.

Current work: almost the true rate

Step 0 : Take $\varepsilon^{(0)} \asymp \left(\alpha_0 \frac{\log(n)}{\beta n}\right)^{\frac{1}{d+1}}$, then w.p $\geq 1 - p_0$, TSE + SD gives $\mathcal{D}^{(1)}$ such that

- $\rightarrow$ $X_j \in \mathcal{M} \Rightarrow X_j \in \mathcal{D}^{(1)}$,
- $\rightarrow$ $X_j \in \mathcal{D}^{(1)} \Rightarrow d(X_j, \mathcal{M}) \leq \varepsilon^{(0),2}$.

Step 1 : Take $\varepsilon^{(1)} \asymp \left(\alpha_1 \frac{\log(n)}{\beta n}\right)^{\gamma_1}$, with $(d+2)\gamma_1 = 2\gamma_0 + 1$, then w.p $\geq 1 - p_0 - p_1$, TPE *on* $\mathcal{D}^{(1)}$ gives

- $\rightarrow$ $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c\varepsilon^{(1)}$,

and TSE + SD gives $\mathcal{D}^{(2)}$ such that

- $\rightarrow$ $X_j \in \mathcal{M} \Rightarrow X_j \in \mathcal{D}^{(2)}$,
- $\rightarrow$ $X_j \in \mathcal{D}^{(2)} \Rightarrow d(X_j, \mathcal{M}) \leq \varepsilon^{(1),2}$.

$$\vdots$$

Step m : Same result with exponent $\gamma_m \rightarrow 1/d$, w.p $\geq 1 - p_0 - \ldots - p_m$

Current work: almost the true rate

Denoising setting: Fix $0 < \delta < 1/d(d+1)$, and set $m = \lceil \log(1/\delta) - \log(d(d+1)) \rceil$. Iterate TSE + SD with windows $\varepsilon^{(r)} = \left(\alpha_\delta \frac{\log(n)}{\beta n}\right)^{\gamma_r}$, $r = 0, \ldots, m$.

Estimation: Let $\hat{\mathcal{M}}$ denote the Delaunay Tangential complex built on $\mathcal{D}^{(m+1)}$.

Proposition (Aamari, L., 2016)

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

Furthermore, this rate of convergence holds and we have ambient isotopy w.h.p.

Conclusion

Some advances:

- A feasible manifold reconstruction procedure achieving (almost) the minimax convergence rate,
- with topological guarantees,
- and limited dependency on the ambiant dimension.

Some new questions:

- True rates for tangent space estimation (current work)?
- Adaptive thresholds in the denoising procedure?