Haussdorff convergence rates for the Tangential Delaunay complex

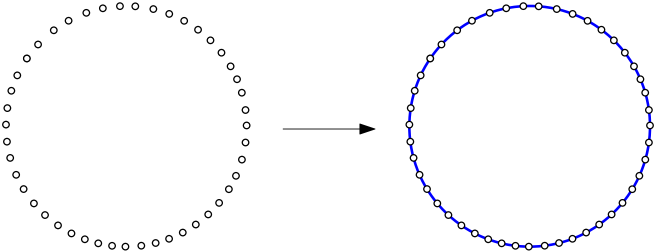
Eddie Aamari[1] Clément Levrard[2]

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

Gudhi/Topdata Workshop

24/10/2015

Manifold reconstruction



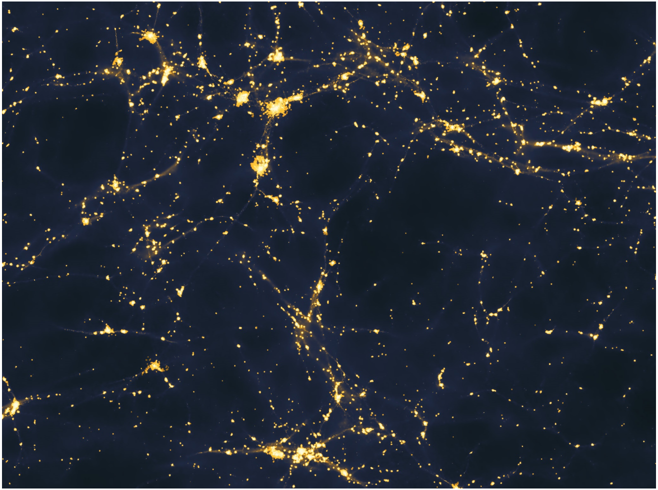
Input: observations $\{X_1, \ldots, X_n\}$ drawn *i.i.d.* on/nearby a manifold $\mathcal{M} \subset \mathbb{R}^D$.

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

- topological guarantees (homeomorphism),

- a good geometric approximation (Haussdorf distance).

Motivation



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

A simplicial complex estimator

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

Figure: Sample points

A simplicial complex estimator

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

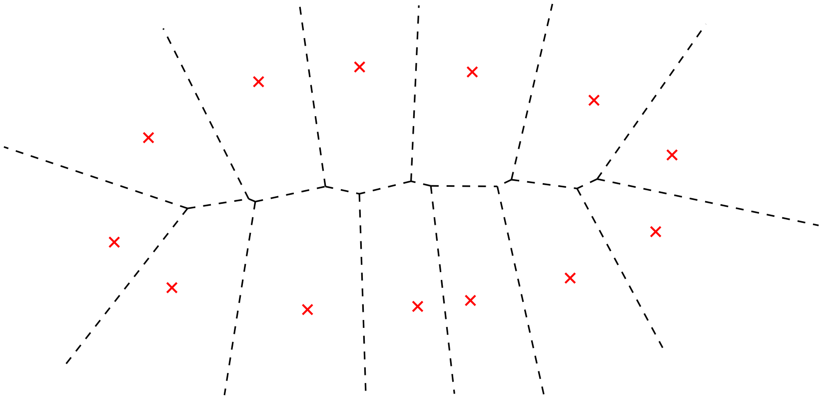


Figure: Voronoi diagram

A simplicial complex estimator

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

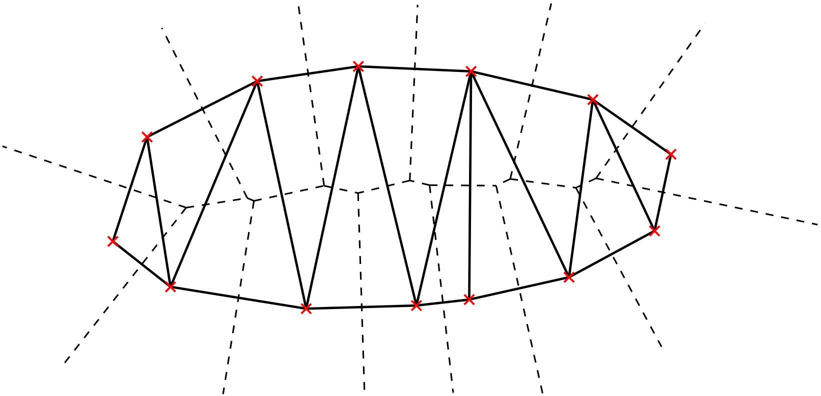


Figure: Delaunay complex

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 $\bigcap_{p\in\tau} \mathrm{Vor}(p) \cap \left(\bigcup_{p\in\tau} T_p\mathcal{M}\right) \neq \emptyset$.

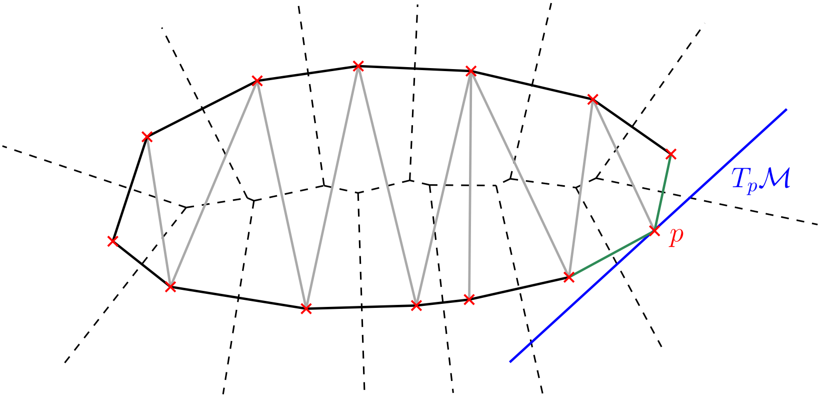
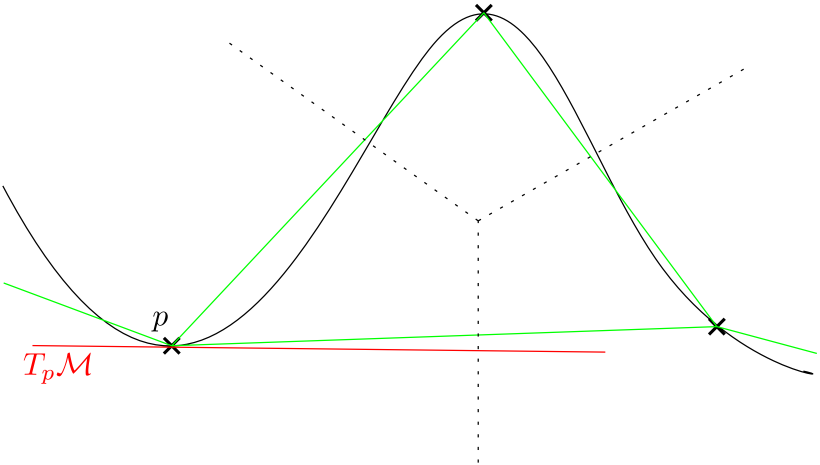


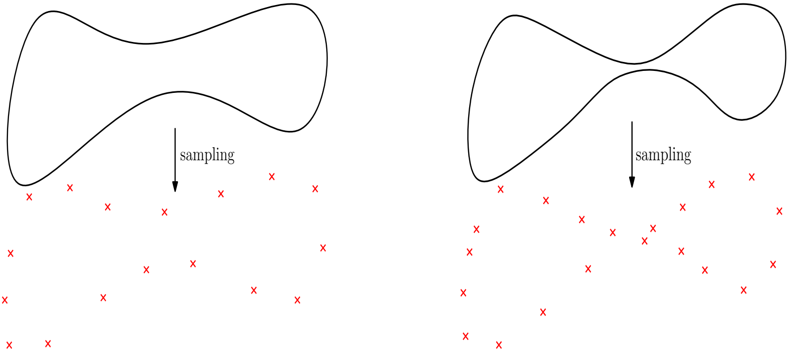
Figure: Tangential Delaunay complex [Boissonnat,Ghosh 2014]

Geometric condition



$\rightarrow$ Bound on curvature.

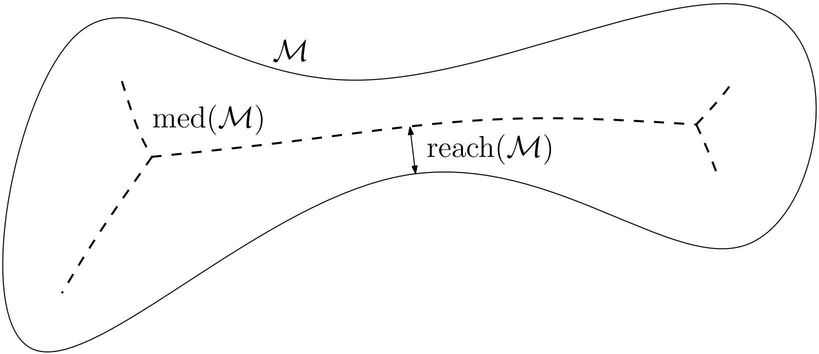
Geometric condition



→ No infinitely small "bottleneck".

Geometric condition

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



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

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 on* $\mathcal{P}$ *and* T *such that:*

- $\mathrm{Del}^{\omega}(\mathcal{P}, T)$ *and* $\mathcal{M}$ *are isotopic,*
- $\mathrm{d_H}\left(\mathrm{Del}^{\omega}(\mathcal{P}, T), \mathcal{M}\right) \leq C\varepsilon^2$*, 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 on* $\mathcal{P}$ *and* T *such that:*

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

Problem: When sampling at random

- ε ?
- $T_p\mathcal{M}$ unknown, $\hat{T}_p\mathcal{M}$?
- What is $\mathrm{Del}^{\omega}(\mathcal{P}, \hat{T})$?

Statistical Model

Geometric assumptions:

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

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

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

Statistical Model

Geometric assumptions:

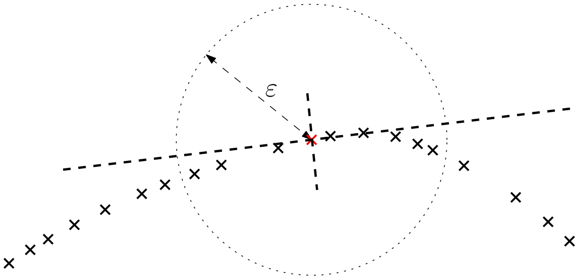
- $\mathcal{M}$ is a closed and connected Riemannian d-submanifold of $\mathbb{R}^D$,
- $\mathrm{reach}(\mathcal{M}) := \rho > 0$.

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

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

$\rightarrow$ $\mathrm{d_H}(\mathcal{P}, \mathcal{M}) \leq 2 \left(\frac{\kappa(d) \log(n)}{n}\right)^{\frac{1}{d}}$, for κ large enough, w.h.p

Tangent Space Estimation: Local PCA

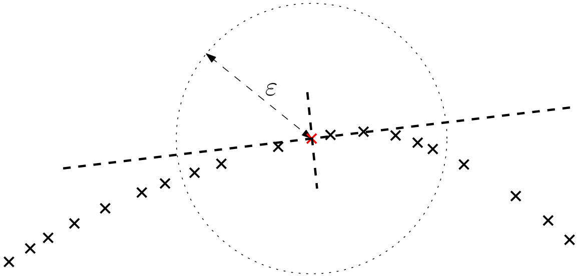


Define $\hat{T}_j$ as the span of the d first eigenvectors of

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

- N_j: number of points in $\mathcal{B}(X_j, \varepsilon)$
- $\bar{X}_j$: local mean

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* $1 - \left(\frac{1}{n}\right)^{2/d}$,

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

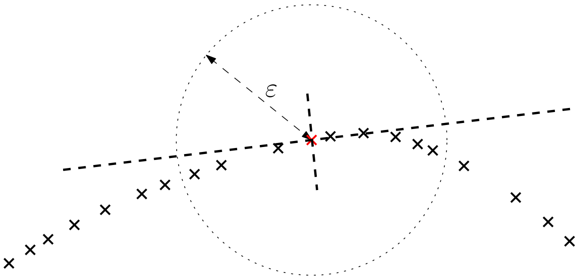
Tangent Space Estimation: Local PCA/Sketch of Proof

$$\Sigma_j = \varepsilon^2 \left[\left(\begin{array}{c|c} I_d & 0 \\ \hline 0 & 0 \end{array} \right) + Bias + \left(\begin{array}{c|c} Dev_{1,1} & Dev_{1,2} \\ \hline Dev_{2,1} & Dev_{2,2} \end{array} \right) \right]$$

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

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

Tangent Space Estimation: Local PCA/Sketch of Proof



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

→ $Bias \lesssim \varepsilon/\rho$,

→ $Dev_{2,1} \lesssim \frac{\varepsilon/\rho}{\sqrt{N_j}}$,

→ $N_j \gtrsim (n-1)\varepsilon^d$.

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

Find $\mathcal{M}'$ such that

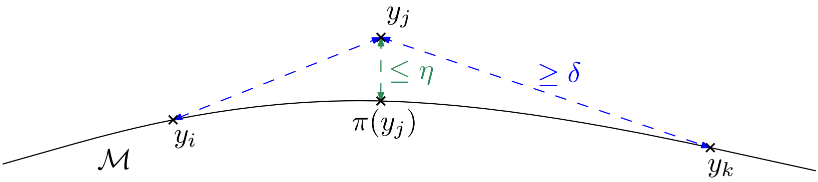
- $\mathcal{M}' \cong \mathcal{M}$
- $\mathrm{d_H}\left(\mathcal{M}', \mathcal{M}\right) \lesssim \varepsilon^2$
- $\mathrm{Del}^\omega(\mathcal{P}, \hat{T}) \cong \mathcal{M}'$
- $\mathrm{d_H}\left(\mathrm{Del}^\omega(\mathcal{P}, \hat{T}), \mathcal{M}'\right) \lesssim \varepsilon^2$,

Interpolation Result

Proposition (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* $\mathcal{M}$*:* $\max_{1 \leq j \leq q} \mathrm{d}(y_j, \mathcal{M}) < \eta$,
- $\max_{1 \leq j \leq q} \angle(T_{\pi(y_j)}\mathcal{M}, T_j) \leq \theta$.

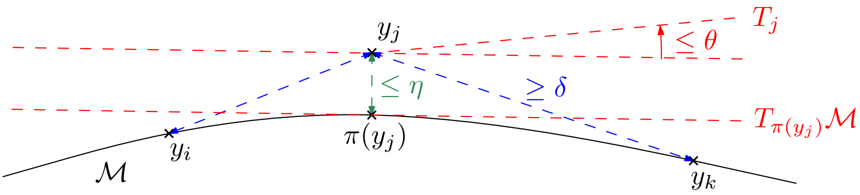


Interpolation Result

Proposition (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 $\mathcal{M}$:* $\max_{1 \leq j \leq q} \mathrm{d}(y_j, \mathcal{M}) < \eta$,
- $\max_{1 \leq j \leq q} \angle(T_{\pi(y_j)}\mathcal{M}, T_j) \leq \theta$.

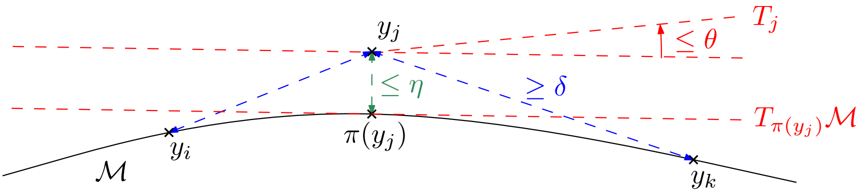


Interpolation Result

Proposition (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)

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

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

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

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

A Noisy Model: Clutter Noise

$$X = (1 - Z)P + Z\mathcal{U}_{\mathcal{B}(0,M)},$$

with $Z \sim \mathcal{B}(\beta) \amalg (P, \mathcal{U}_{\mathcal{B}(0,M)})$, P as previously.

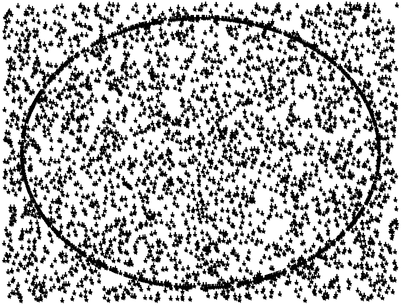
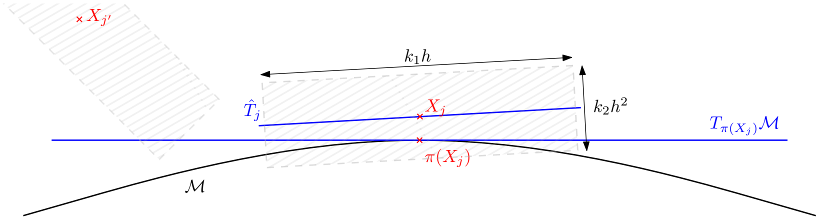


Figure: Clutter noise model

A denoising procedure

Define slabs S_j centered at each X_j:

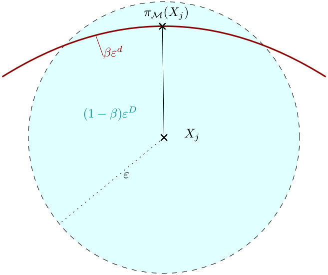


To determine if $X_j \in \mathcal{M}$, consider $P_n(S_j) = |S_j \cap \{X_1, \ldots, X_n\}|$.
As $\varepsilon \to 0$, provided that $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq \varepsilon$, w.h.p,

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

Tangent space estimation again

Locally "no noise"



For $d(X_j, \mathcal{M}) \leq \kappa\varepsilon$, $\kappa < 1$

$$\mathbb{P}(Z = 1 | X \in \mathcal{B}(X_j, \varepsilon)) \leq c(d, D, \beta)\varepsilon^{D-d}$$

$$\begin{aligned}\mathbb{P}(N_{j,n} \neq 0 | \{X_i \in \mathcal{B}(X_j, \varepsilon)\}) \\ \leq C(d, D, \beta) N_j \varepsilon^{D-d}\end{aligned}$$

Tangent space estimation again

$$\Sigma_j = \varepsilon^2 \left[\left(\begin{array}{c|c} I_d & 0 \\ \hline 0 & 0 \end{array} \right) + Bias + Dev \right]_{nonoise},$$

w.h.p - $CN_j \varepsilon^{D-d}$

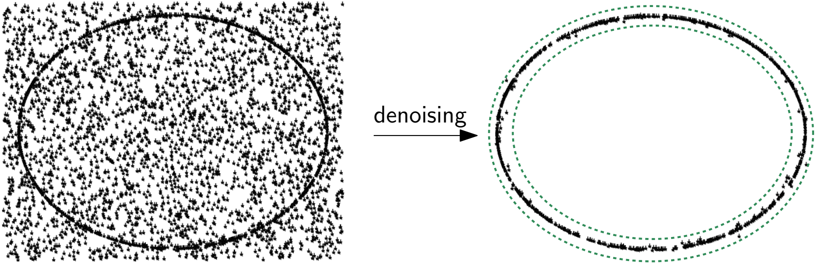
Taking $\varepsilon = \left(\kappa \frac{\log(n)}{\beta n} \right)^{\frac{1}{d}}$ gives

- $\rightarrow$ $Bias \approx Dev_{nonoise,12} \approx \varepsilon/\rho$ w.h.p
- $\rightarrow$ for all $d(X_j, \mathcal{M}) \leq \kappa\varepsilon$, $\angle(\hat{T}_j, T_{\pi(X_j)}\mathcal{M}) \leq c(d, D)\varepsilon/\rho$.

Clustering Result

For $\varepsilon = \left(\kappa \frac{\log(n)}{\beta n}\right)^{\frac{1}{d}}$, 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 ε^2 neighbourhood of $\mathcal{M}$.



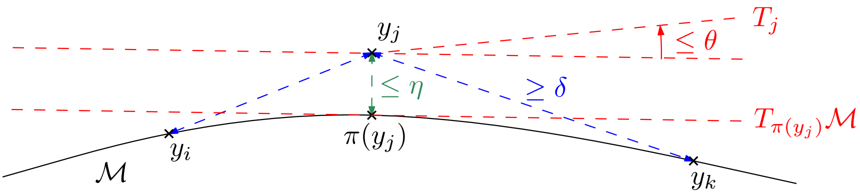
Convergence Result

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

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

where $\cong$ denotes the isotopy equivalence.
Moreover, if $D \geq d+2$, for n large enough,

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

Conclusion

Some advances:

- A feasible manifold reconstruction procedure achieving 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 window? Adaptive threshold in the denoising procedure?