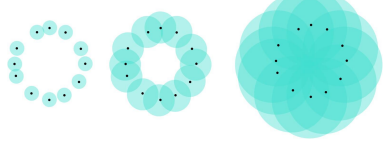


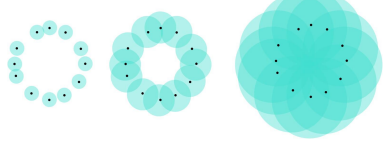
Topological consistency via kernel estimation

Amoukou Salim

Statistics

May 31, 2019

Introduction



The goal is to estimate the topology of a set $D_L = \{x \in \mathbb{R}^d : f(x) \geq L\}$

- Why: Clustering, pattern recognition, econometric...
- With: $D_n = \{(X_1, Y_1), ...(X_n, Y_n)\}$
- Simple idea: Use the plug-in estimators.

Introduction

The goal is to estimate the topology of a set $D_L = \{x \in \mathbb{R}^d : f(x) \geq L\}$

- Why: Clustering, pattern recognition, econometric...
- With: $D_n = \{(X_1, Y_1), ...(X_n, Y_n)\}$
- Simple idea: Use the plug-in estimators.

Introduction

The goal is to estimate the topology of a set $D_L = \{x \in \mathbb{R}^d : f(x) \geq L\}$

- Why: Clustering, pattern recognition, econometric...
- With: $D_n = \{(X_1, Y_1), ...(X_n, Y_n)\}$
- Simple idea: Use the plug-in estimators.

One solution from the course

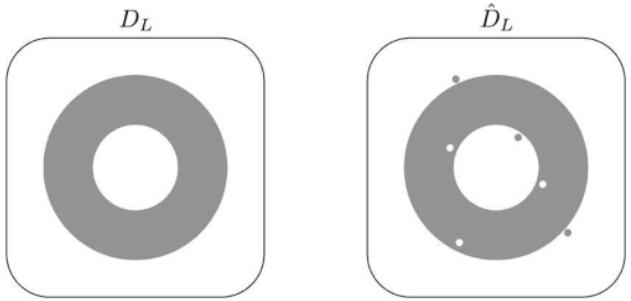
Theorem

Let f be a density with compact support S, f_n a sequence of density estimators and $\hat{S} = \{f_n > \alpha_n\}$ where $\alpha_n \searrow 0$.
If we assume that:

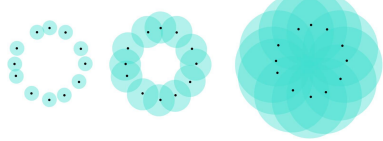
- $\lambda(E_0) = 0$ *where* $E_0 = \{x \in S | f(x) = 0\}$
- $\alpha_n^{-1} \int |f_n - f| d\lambda \to 0$ *(as)*

then $d_\lambda(S, \hat{S}) \to 0$

Goal: Good reconstruction according to Hausdorff distance $\not\Rightarrow$ Good recovery of the topology of X.



How we describe the topology of D_L ?



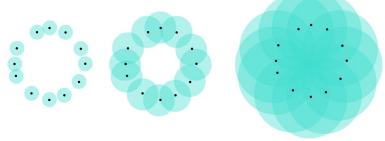
The group of homology of X : A set of vector space $\{H_k(X)\}_{k=0}^{\infty}$
$\rightarrow$Intuitively:

- $rank(H_o(X)) \cong$ connected components
- $rank(H_k(X)) \cong$ k-cycle of X

The persistent homology of X: $PH_*(X)$

- track the evolution of the homology as L decreasing or increasing.
- bottleneck distance to compare different persistent diagram:
  $d_B(PH_*(X), PH_*(Y)) = \inf_\gamma \sup_{p \in Dgm(X)} ||p - \gamma(p)||$

Naive approach



Naive approach to recover the homology of an unknown space S from a random sample $\mathcal{X}$: $U(\mathcal{X}, r) = \bigcup_{x \in \mathcal{X}} B_r(x)$ for some choice of r.

Theorem

Let μ a probability distribution and $K = supp(\mu)$ st:

- *$reach(K) > 0$*
- *μ is (a,b) standard at scale $r_0 > 0$*

Let $X_1, ..., X_n$ be iid sample of μ and $\epsilon < \frac{1}{17} reach(K) \wedge 2r_0$ then $\forall n \geq C_{a,b} log(\frac{1}{\epsilon}) + log(\frac{1}{\epsilon})/\epsilon^b$

$$\mathbb{P}[K^{\frac{1}{2}reach(K)} \text{ and } \mathbb{X}_n^{4\epsilon} \text{ are homotopy equivalent}] \geq 1 - \eta$$

Pb: Strong assumption on the density and the shape of the level set.
Our solution: We will use the same technique but with weak assumption to recover the homology.

Naive approach

Naive approach to recover the homology of an unknown space S from a random sample $\mathcal{X}$: $U(\mathcal{X}, r) = \bigcup_{x \in \mathcal{X}} B_r(x)$ for some choice of r.

Theorem

Let μ a probability distribution and $K = supp(\mu)$ st:

- *$reach(K) > 0$*
- *μ is (a,b) standard at scale $r_0 > 0$*

Let $X_1, ..., X_n$ be iid sample of μ and $\epsilon < \frac{1}{17} reach(K) \wedge 2r_0$ then $\forall n \geq C_{a,b} log(\frac{1}{\epsilon}) + log(\frac{1}{\epsilon})/\epsilon^b$

$$\mathbb{P}[K^{\frac{1}{2} reach(K)} \text{ and } \mathbb{X}_n^{4\epsilon} \text{ are homotopy equivalent}] \geq 1 - \eta$$

Pb: Strong assumption on the density and the shape of the level set.

Our solution: We will use the same technique but with weak assumption to recover the homology.

Naive approach

Naive approach to recover the homology of an unknown space S from a random sample $\mathcal{X}$: $U(\mathcal{X}, r) = \bigcup_{x \in \mathcal{X}} B_r(x)$ for some choice of r.

Theorem

Let μ a probability distribution and $K = supp(\mu)$ st:

- *$reach(K) > 0$*
- *μ is (a,b) standard at scale $r_0 > 0$*

Let $X_1, ..., X_n$ be iid sample of μ and $\epsilon < \frac{1}{17} reach(K) \wedge 2r_0$ then $\forall n \geq C_{a,b} log(\frac{1}{\epsilon}) + log(\frac{1}{\epsilon})/\epsilon^b$

$$\mathbb{P}[K^{\frac{1}{2}reach(K)} \text{ and } \mathbb{X}_n^{4\epsilon} \text{ are homotopy equivalent}] \geq 1 - \eta$$

Pb: Strong assumption on the density and the shape of the level set.

Our solution: We will use the same technique but with weak assumption to recover the homology.

Naive approach

Naive approach to recover the homology of an unknown space S from a random sample $\mathcal{X}$: $U(\mathcal{X}, r) = \bigcup_{x \in \mathcal{X}} B_r(x)$ for some choice of r.

Theorem

Let μ a probability distribution and $K = supp(\mu)$ st:

- *$reach(K) > 0$*
- *μ is (a,b) standard at scale $r_0 > 0$*

Let $X_1, ..., X_n$ be iid sample of μ and $\epsilon < \frac{1}{17} reach(K) \wedge 2r_0$ then $\forall n \geq C_{a,b} log(\frac{1}{\epsilon}) + log(\frac{1}{\epsilon})/\epsilon^b$

$$\mathbb{P}[K^{\frac{1}{2}reach(K)} \text{ and } \mathbb{X}_n^{4\epsilon} \text{ are homotopy equivalent}] \geq 1 - \eta$$

Pb: Strong assumption on the density and the shape of the level set.

Our solution: We will use the same technique but with weak assumption to recover the homology.

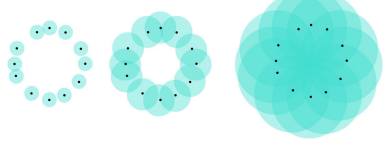
"Tameless"

The main assumption on f is "tameless". Let D_L define like before.

Definition

- L is call homological regular value if $\exists \epsilon > 0$ st $\forall v_1 \leq v_2$ in $[L - 2\epsilon, L + 2\epsilon]$, the inclusion map $H_*(D_{v1}) \hookrightarrow H_*(D_{v2})$ is an isomorphism. Otherwise, we say that L is **homological critical value**
- A function is called tame if it has a **finite** homological critival value and $rank(H_*(D_L))$ is finite $\forall L$

To sum up



Our goal is to present a consistent method for recovering the homology of a given set D_L. We will examine two cases :

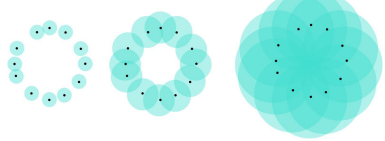
- Density function p with kernel estimator $f_n(x) = \frac{1}{C_k n r^d} \sum K_r(x - X_i)$
  - $p_{max} = \sup_{x \in \mathbb{R}^d} p(x) < +\infty$
  - D_L is bounded for all $L > 0$
- Regression function $f(x) = \mathbb{E}(Y|X = x)$ with kernel estimator $f_n(x) = \frac{Y_i \sum K_r(x - X_i)}{\sum K_r(x - X_i)}$, if p is the marginal density of X then
  - $supp(p)$ compact and $p_{min} = \inf_{x \in \mathbb{R}^d} p(x) > 0$
  - $|Y_i| < Y_{max}$ almost surely
- Kernel estimator satisfies the following:
  - $supp(K) \subset B_1(0)$
  - $K(x) \in [0, 1]$, and $K(0) = 1$
  - $\int K(x)dx = C_K$ for $C_k \in (0, 1)$

To sum up

Our goal is to present a consistent method for recovering the homology of a given set D_L. We will examine two cases :

- Density function p with kernel estimator $f_n(x) = \frac{1}{C_k n r^d} \sum K_r(x - X_i)$
  - $p_{max} = \sup_{x \in \mathbb{R}^d} p(x) < +\infty$
  - D_L is bounded for all $L > 0$
- Regression function $f(x) = \mathbb{E}(Y|X = x)$ with kernel estimator $f_n(x) = \frac{Y_i \sum K_r(x - X_i)}{\sum K_r(x - X_i)}$, if p is the marginal density of X then
  - $supp(p)$ compact and $p_{min} = \inf_{x \in \mathbb{R}^d} p(x) > 0$
  - $|Y_i| < Y_{max}$ almost surely
- Kernel estimator satisfies the following:
  - $supp(K) \subset B_1(0)$
  - $K(x) \in [0, 1]$, and $K(0) = 1$
  - $\int K(x)dx = C_K$ for $C_k \in (0, 1)$

To sum up



Our goal is to present a consistent method for recovering the homology of a given set D_L. We will examine two cases :

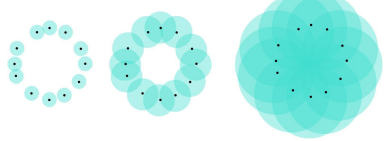
- Density function p with kernel estimator $f_n(x) = \frac{1}{C_k n r^d} \sum K_r(x - X_i)$
  - $p_{max} = \sup_{x \in \mathbb{R}^d} p(x) < +\infty$
  - D_L is bounded for all $L > 0$
- Regression function $f(x) = \mathbb{E}(Y|X = x)$ with kernel estimator $f_n(x) = \frac{Y_i \sum K_r(x - X_i)}{\sum K_r(x - X_i)}$, if p is the marginal density of X then
  - $supp(p)$ compact and $p_{min} = \inf_{x \in \mathbb{R}^d} p(x) > 0$
  - $|Y_i| < Y_{max}$ almost surely
- Kernel estimator satisfies the following:
  - $supp(K) \subset B_1(0)$
  - $K(x) \in [0, 1]$, and $K(0) = 1$
  - $\int K(x)dx = C_K$ for $C_k \in (0, 1)$

To sum up

Our goal is to present a consistent method for recovering the homology of a given set D_L. We will examine two cases :

- Density function p with kernel estimator $f_n(x) = \frac{1}{C_k n r^d} \sum K_r(x - X_i)$
  - $p_{max} = \sup_{x \in \mathbb{R}^d} p(x) < +\infty$
  - D_L is bounded for all $L > 0$
- Regression function $f(x) = \mathbb{E}(Y|X = x)$ with kernel estimator $f_n(x) = \frac{Y_i \sum K_r(x - X_i)}{\sum K_r(x - X_i)}$, if p is the marginal density of X then
  - $supp(p)$ compact and $p_{min} = \inf_{x \in \mathbb{R}^d} p(x) > 0$
  - $|Y_i| < Y_{max}$ almost surely
- Kernel estimator satisfies the following:
  - $supp(K) \subset B_1(0)$
  - $K(x) \in [0, 1]$, and $K(0) = 1$
  - $\int K(x)dx = C_K$ for $C_k \in (0, 1)$

First procedure: Naive estimator

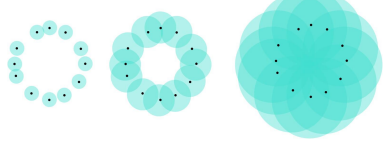


1. Use the dataset to construct an Kernel estimator $\hat{f}$
2. Using the estimator $\hat{f}$, define $\mathcal{X}^L = \{X_i : \hat{f}(X_i) \geq L\}$
3. Consider $U(\mathcal{X}, r)$ as an estimate of D_L, and $\hat{D}_L(n, r) = H_*(U(\mathcal{X}, r))$ as an estimate of $H_*(D_L)$

Pb : $\hat{f}$ may occur error in the step 2

Solution : Use filtering step like in the course to overcome the nosiness of $\hat{D}_L(n, r)$.

First procedure: Naive estimator

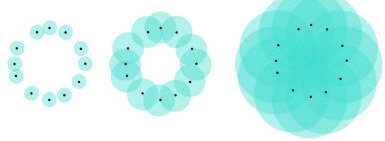


1. Use the dataset to construct an Kernel estimator $\hat{f}$
2. Using the estimator $\hat{f}$, define $\mathcal{X}^L = \{X_i : \hat{f}(X_i) \geq L\}$
3. Consider $U(\mathcal{X}, r)$ as an estimate of D_L, and $\hat{D}_L(n, r) = H_*(U(\mathcal{X}, r))$ as an estimate of $H_*(D_L)$

Pb : $\hat{f}$ may occur error in the step 2

Solution : Use filtering step like in the course to overcome the nosiness of $\hat{D}_L(n, r)$.

First procedure: Naive estimator



1. Use the dataset to construct an Kernel estimator $\hat{f}$
2. Using the estimator $\hat{f}$, define $\mathcal{X}^L = \{X_i : \hat{f}(X_i) \geq L\}$
3. Consider $U(\mathcal{X}, r)$ as an estimate of D_L, and $\hat{D}_L(n, r) = H_*(U(\mathcal{X}, r))$ as an estimate of $H_*(D_L)$

Pb : $\hat{f}$ may occur error in the step 2

Solution : Use filtering step like in the course to overcome the nosiness of $\hat{D}_L(n, r)$.

Use image to filter

Since we have $\hat{D}_{L+\epsilon}(n,r) \subset \hat{D}_{L-\epsilon}(n,r)$, we have the homology map :

$i_* : H_*(\hat{D}_{L+\epsilon}(n,r)) \hookrightarrow H_*(\hat{D}_{L+\epsilon}(n,r))$

The new estimator : $\hat{H}(L,\epsilon,n) = Im(i_*)$

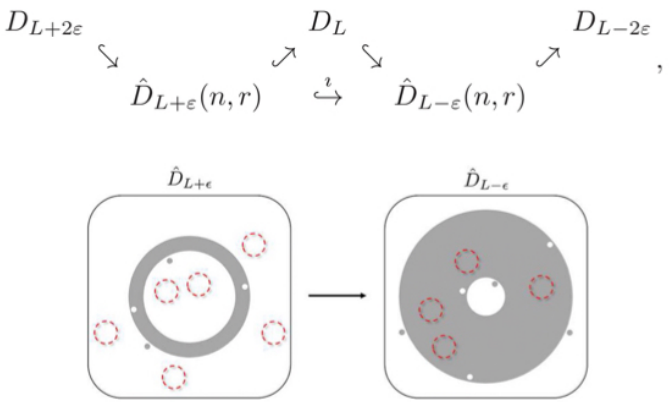


Figure: The intuition behind

ϵ-Regular

To have the precedent inclusion sequence, we require the following regularity condition on L.

Definition

L is ϵ-regular if :

$$\partial D_{L+2\epsilon} \cap \partial D_{L+\frac{3}{2}\epsilon} = D_{L+\frac{1}{2}\epsilon} \cap \partial D_L = \partial D_L \cap D_{L-\frac{1}{2}\epsilon} = \partial D_{L-\frac{3}{2}\epsilon} \cap \partial D_{L-2\epsilon} = \emptyset$$

In particular, if f is continous in $f^{-1}([L-2\epsilon, L+2\epsilon])$ so L is ϵ-regular.

ϵ-Regular

To have the precedent inclusion sequence, we require the following regularity condition on L.

Definition

L is ϵ-regular if :

$$\partial D_{L+2\epsilon} \cap \partial D_{L+\frac{3}{2}\epsilon} = D_{L+\frac{1}{2}\epsilon} \cap \partial D_L = \partial D_L \cap D_{L-\frac{1}{2}\epsilon} = \partial D_{L-\frac{3}{2}\epsilon} \cap \partial D_{L-2\epsilon} = \emptyset$$

In particular, if f is continous in $f^{-1}([L-2\epsilon, L+2\epsilon])$ *so L is* ϵ*-regular.*

The main theorem of the paper

Theorem

Let $L > 0$ and $\epsilon \in (0, \frac{L}{2})$ be such that $f(x)$ has no critical value in the range $[L - 2\epsilon, L + 2\epsilon]$, if $r \to 0$ and $nr^d \to +\infty$ then:

$$\mathbb{P}(H_*(L, \epsilon, n)) \cong H_*(D_L)) \geq 1 - 6ne^{-C^*_{\epsilon/2} nr^d}$$

In particular, if $nr^d \geq Dlog(n)$ with $D > (C_{\epsilon/2})^{-1}$

$$\mathbb{P}(H_*(L, \epsilon, n)) \cong H_*(D_L)) = 1$$

Lemma 1

To prove the theorem, we need 3 lemmas. The first is :

Lemma

For every $L > 0$, and $\epsilon \in (0, L)$, if $r \to 0$ and $nr^d \to \infty$, then there exists a constant C_ϵ such that for n large enough we have :

$$\mathbb{P}(\exists X_i \notin D^{\uparrow}_{L-\epsilon}(r) : f_n(X_i) \geq L) \leq ne^{-C^*_\epsilon nr^d} \tag{1}$$

$$\mathbb{P}(\exists X_i \in D^{\uparrow}_{L-\epsilon}(r) : f_n(X_i) \leq L) \leq ne^{-C^*_\epsilon nr^d} \tag{2}$$

Proof lemma 1

Recall that : $\text{supp}(K) \subset B_1(0)$, $K(x) \in [0,1]$ and $K(0)=1$

the kernel estimator is : $\hat{f}_m(x) = \dfrac{\sum K_h(x - X_i)}{C_K m h^d}$

We have:

$$\mathbb{P}(\exists X_i \notin \hat{\Lambda}^{-1}_{L-\varepsilon}(h) : \hat{f}_m(X_i) \geq L) \leq m\, \mathbb{P}(X_1 \in (\hat{\Lambda}^{-1}_{L-\varepsilon}(h))^c : \hat{f}_m(X_1) \geq L)$$

$$\leq m \int_{(\hat{\Lambda}^{-1}_{L-\varepsilon}(h))^c} f_X(x)\, \mathbb{P}(\hat{f}_m(X_1) \geq L \mid X = x)$$

$$\mathbb{P}(\hat{f}_m(X_1) \geq L \mid X = x) = \mathbb{P}\left(K_h(0) + \sum_{i=2}^{m} K_h(x - X_i) \geq L C_K m h^d\right)$$

$$= \mathbb{P}\left(\sum_{i=2}^{m} Z_i \geq m\left(L C_K h^d - p_h(x)\right) + p_h(x) - 1\right)$$

where $Z_i = K_h(x - X_i) - p_h(x)$ and $p_h(x) = \mathbb{E}[K_h(x - X_i)]$

$$p_h(x) = \mathbb{E}[K_h(x - X_i)] = \int_{B_h(x)} f_X(\xi) K_h(\xi - X_i)\, d\xi \leq (L - \varepsilon) C_K h^d$$

$x \notin \hat{\Lambda}^{-1}_{L-\varepsilon}(h)$

Proof lemma 1

So reassemble the puzzle and we have:

$$\boxed{\mathbb{P}(f_{n}(X_1) \geq L \mid X = x)} \leq \mathbb{P}\left(\sum_{i=2}^{n} Z_i \geq \varepsilon C_K n r^d - 1\right)$$

Note that $|Z_i| \leq 1$ and $\mathrm{Var}(Z_i) = \mathbb{E}[K_r^2(x - X_i)] \leq p_{max} C_K r^d$

Using Bernstein inequality with $t = \varepsilon C_K n r^d - 1$ and the limit we have: $\boxed{\mathbb{P}(f_{n}(X_1) \geq L \mid X = x)} \leq e^{-C_\varepsilon^* n r^d}$

where $C_\varepsilon^* = \dfrac{\varepsilon^2 C_K}{3 p_{max} + \varepsilon}$

- We do same technique for the second and in Regression case.

Lemma 2

Lemma

For every $L > 0$, and $\epsilon \in (0, L)$, if $r \to 0$ and $nr^d \to \infty$, then there exists a constant C_ϵ such that for n large enough we have :

$$\mathbb{P}(D^{\downarrow}_{L+\epsilon}(2r) \subset \hat{D}_L(n, r) \subset D^{\uparrow}_{L-\epsilon}(2r)) \geq 1 - 3ne^{-C_\epsilon nr^d} \tag{3}$$

Proof lemma 2

- Recall that we want $\mathbb{P}\left(\hat{\mathbb{L}}^{\downarrow}_{L+\varepsilon}(2h) \subset \hat{\mathbb{L}}_{L}(m,h) \subset \hat{\mathbb{L}}^{\uparrow}_{L-\varepsilon}(2h)\right) \geq 1 - 3m\, e^{-C^{*}_{\varepsilon} m h^{d}}$

- $\mathbb{P}\left(\hat{\mathbb{L}}_{L}(m,h) \not\subset \hat{\mathbb{L}}^{\uparrow}_{L-\varepsilon}(2h)\right) \leq \mathbb{P}\left(\exists X_i \notin \mathbb{L}^{\uparrow}_{L-\varepsilon}(h) : \hat{f}_m(X_i) \geq L\right) \leq m\, e^{-C^{*}_{\varepsilon} m h^{d}}$

For the second inclusion, we have to note that:

- Since $\mathbb{L}^{\downarrow}_{L+\varepsilon}(2h)$ is bounded, let $S \subset \mathbb{L}^{\downarrow}_{L+\varepsilon}(2h)$ be a δh-covering.
- If $\exists x \in \mathbb{L}^{\downarrow}_{L+\varepsilon}(2h)$ that is not covered by $\cup(\mathcal{X}_m, h)$ so $\exists s \in S$ s.t. S is not covered by $\cup(\mathcal{X}_m, (1-\delta)h)$.

Lemma: $\forall y \in \mathcal{X}_m^{L}$: $\|x - y\| > h$, and let $s \in S$ such that $\|x - s\| \leq \delta h \Rightarrow$

$$\|y - s\| = \|y - x + x - s\| \geq \|y - x\| - \|x - s\| \geq h - \delta h \geq h(1-\delta)$$

Donc on peut dire:

$$\mathbb{P}\left(\hat{\mathbb{L}}^{\downarrow}_{L+\varepsilon}(2h) \subset \hat{\mathbb{L}}_{L}(m,h)\right) \leq \mathbb{P}\left(\exists s \in S : B_{(1-\delta)h}(s) \cap \mathcal{X}_m^{L} = \emptyset\right)$$

$$\leq \mathbb{P}\left(\exists s \in S : B_{(1-\delta)h}(s) \cap \mathcal{X}_m^{L} = \emptyset ;\ \mathbb{L}^{\downarrow}_{L+\varepsilon}(h) \cap \mathcal{X}_m \subset \mathcal{X}_m^{L}\right) + \mathbb{P}\left(\exists s \in S : B_{(1-\delta)h}(s) \cap \mathcal{X}_m^{L} = \emptyset ;\ \mathbb{L}^{\downarrow}_{L+\varepsilon}(h) \cap \mathcal{X}_m \not\subset \mathcal{X}_m^{L}\right)$$

$$\leq \mathbb{P}\left(\exists s \in S : B_{(1-\delta)h}(s) \cap \mathcal{X}_m = \emptyset\right) + \mathbb{P}\left(\mathbb{L}^{\downarrow}_{L+\varepsilon}(h) \cap \mathcal{X}_m \not\subset \mathcal{X}_m^{L}\right)$$

Original sample does not $\Leftarrow$ $(1-\delta)h$

$\leq C_1 h^{-d} e^{-C_2 m h^{d}}$; $O(m)$

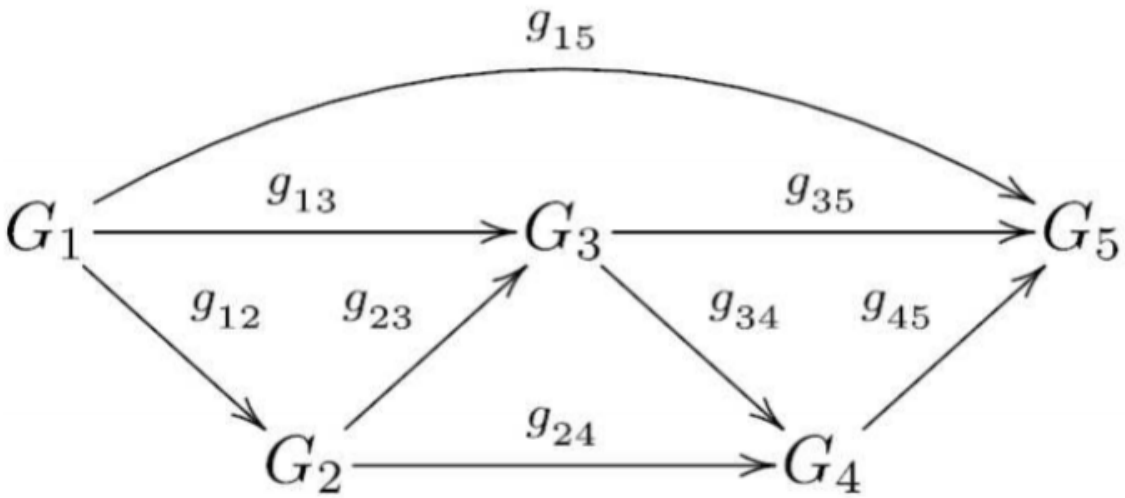
get rid to many point $\Leftarrow$ $\mathbb{L}^{\downarrow}_{L+\varepsilon}(h)$

$\leq m\, e^{-C^{*}_{\varepsilon} m h^{d}}$

$\Rightarrow \leq 2m\, e^{-C^{*}_{\varepsilon} m h^{d}}$

Lemma 3

Consider the following commutative diagram of group :



Define $G_{i,j} = Im(g_{i,j}) \subset G_j$.
If $g_{3,5} : G_3 \to G_5$ is an isomorphism from G_3 to $G_{1,5}$. Then the map $g_{3,4} : G_3 \to G_4$ is an isomorphism from G_3 to $G_{2,4}$. In particular, we have

$\boxed{G_3 \cong G_{2,4}}$

The proof of the theorem

Proof.

Recall that we want to prove that: if $r \to 0$, $mh^d \to 0$ then $H^*(L_\varepsilon(m,r)) \cong H^*(\Lambda_L)$

Using Lemma 2 for $\hat{\Lambda}_{L-\varepsilon}(m,r)$ and $\hat{\Lambda}_{L+\varepsilon}(m,r)$, we have:

$$\mathbb{P}\left(\check{\Lambda}_{L+\frac{3}{2}\varepsilon}(2r) \subset \hat{\Lambda}_{L+\varepsilon}(m,r) \subset \hat{\Lambda}^{1}_{L+\frac{1}{2}\varepsilon}(2r)\right) \geq 1 - 3m e^{-C'_{\varepsilon/2} mh^d}$$

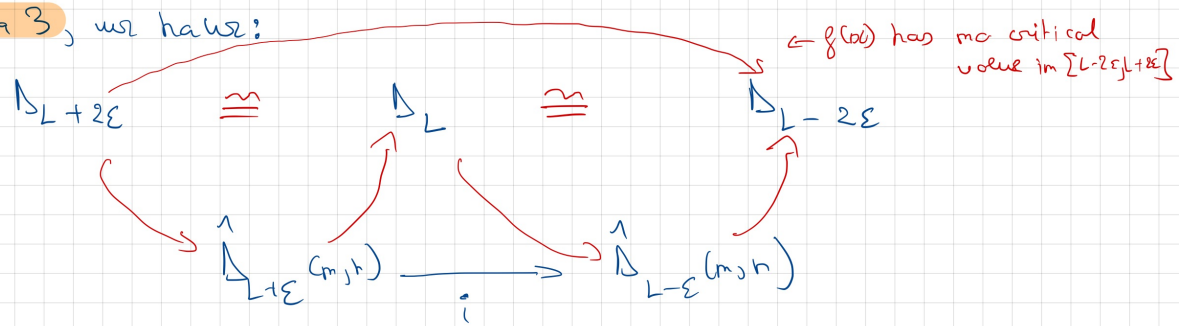
$$\mathbb{P}\left(\check{\Lambda}_{L-\frac{1}{2}\varepsilon}(2r) \subset \hat{\Lambda}_{L-\varepsilon}(m,r) \subset \hat{\Lambda}^{1}_{L-\frac{3}{2}\varepsilon}(2r)\right) \geq 1 - 3m e^{-C'_{\varepsilon/2} mh^d}$$

We want Λ_L not $\hat{\Lambda}^{1}_{L}$ or $\check{\Lambda}_{L}$ but since we assume L is ε-regular, if r small enough:

$$\Lambda_{L+2\varepsilon} \subset \check{\Lambda}_{L+\frac{3}{2}\varepsilon} \subset \hat{\Lambda}^{1}_{L+\frac{1}{2}\varepsilon}(2r) \subset \Lambda_L \subset \check{\Lambda}_{L-\frac{1}{2}\varepsilon} \subset \hat{\Lambda}^{1}_{L-\frac{3}{2}\varepsilon}(2r) \subset \Lambda_{L-2\varepsilon}$$

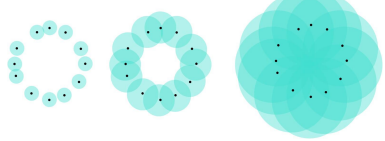
So we have: $\mathbb{P}\left(\Lambda_{L+2\varepsilon} \subset \hat{\Lambda}_{L+\varepsilon}(m,r) \subset \Lambda_L \subset \hat{\Lambda}_{L-\varepsilon}(m,r) \subset \Lambda_{L-2\varepsilon}\right) \geq 1 - 6m e^{-C'_{\varepsilon/2} mh^d}$

Using Lemma 3, we have:



by Lemma 3: $H_*(\Lambda_L) \cong H_*(\mathrm{Im}(i_*))$

Application to manifold learning



Let $\mathcal{M}$ be a smoth m-dimensional closed manifold in $\mathbb{R}^d$. We wish to recover the homology of $\mathcal{M}$ given $\mathcal{X}_n = \{X_1, ..., X_n\}$.

$\rightarrow$ The case $\mathcal{X}_n \subset \mathcal{M}$ have been extensively studied [ref], for example:

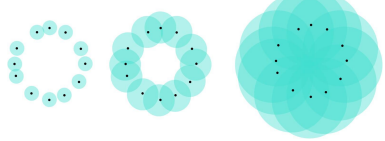
$\rightarrow$ Theorem : If $nr^d \geq Clog(n)$ and $C > (w_d p_{min})^{-1}$ then $H_*(U(\mathcal{X}_n, r)) \cong H_*(\mathcal{M})$ as.

$\rightarrow$ Is there similar result when noise is present ?

$\rightarrow$ Yes in special case or in the course with wasserstein.

Goal : We want to resolve this problem with a large class of distributions and with as few assumptions as possible.

Application to manifold learning



Let $\mathcal{M}$ be a smoth m-dimensional closed manifold in $\mathbb{R}^d$. We wish to recover the homology of $\mathcal{M}$ given $\mathcal{X}_n = \{X_1, ..., X_n\}$.

$\rightarrow$ The case $\mathcal{X}_n \subset \mathcal{M}$ have been extensively studied [ref], for example:

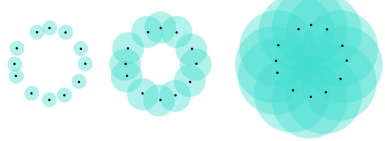
$\rightarrow$ Theorem : If $nr^d \geq Clog(n)$ and $C > (w_d p_{min})^{-1}$ then $H_*(U(\mathcal{X}_n, r)) \cong H_*(\mathcal{M})$ as.

$\rightarrow$ Is there similar result when noise is present ?

$\rightarrow$ Yes in special case or in the course with wasserstein.

Goal : We want to resolve this problem with a large class of distributions and with as few assumptions as possible.

Application to manifold learning



Let $\mathcal{M}$ be a smoth m-dimensional closed manifold in $\mathbb{R}^d$. We wish to recover the homology of $\mathcal{M}$ given $\mathcal{X}_n = \{X_1, ..., X_n\}$.

$\rightarrow$ The case $\mathcal{X}_n \subset \mathcal{M}$ have been extensively studied [ref], for example:

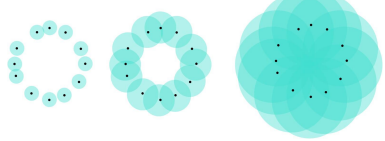
$\rightarrow$ Theorem : If $nr^d \geq Clog(n)$ and $C > (w_d p_{min})^{-1}$ then $H_*(U(\mathcal{X}_n, r)) \cong H_*(\mathcal{M})$ as.

$\rightarrow$ Is there similar result when noise is present ?

$\rightarrow$ Yes in special case or in the course with wasserstein.

Goal : We want to resolve this problem with a large class of distributions and with as few assumptions as possible.

Application to manifold learning



Let $\mathcal{M}$ be a smoth m-dimensional closed manifold in $\mathbb{R}^d$. We wish to recover the homology of $\mathcal{M}$ given $\mathcal{X}_n = \{X_1, ..., X_n\}$.

$\rightarrow$ The case $\mathcal{X}_n \subset \mathcal{M}$ have been extensively studied [ref], for example:

$\rightarrow$ Theorem : If $nr^d \geq Clog(n)$ and $C > (w_d p_{min})^{-1}$ then $H_*(U(\mathcal{X}_n, r)) \cong H_*(\mathcal{M})$ as.

$\rightarrow$ Is there similar result when noise is present ?

$\rightarrow$ Yes in special case or in the course with wasserstein.

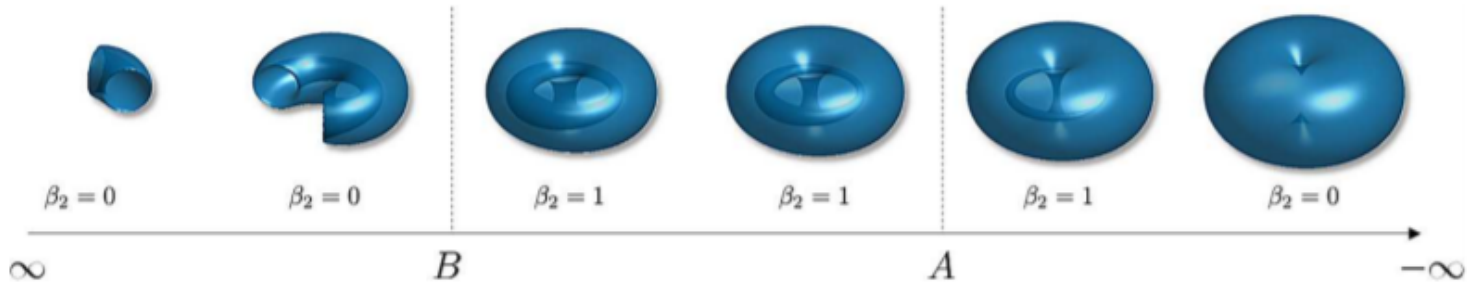
Goal : We want to resolve this problem with a large class of distributions and with as few assumptions as possible.

The general class of distribution

Let $p : \mathbb{R}^d \to \mathbb{R}^+$ be a probability distribution. We say that p represents a noisy version of $\mathcal{M}$, if there exists $0 < A < B < \infty$ st :

- $\forall L \in [A, B]$, $D_L \cong \mathcal{M}$
- $\forall L > B$, $D_L \cong \mathcal{M}'$ where $\mathcal{M}' \subset \mathcal{M}$ is a compact locally contractible proper subset of $\mathcal{M}$

Exemple:



The ideal case

If we know A, B then the recovery will be simple.

$\rightarrow$ Choose L st $[L - 2\epsilon, L + 2\epsilon] \subset [A, B]$

$\rightarrow$ Theorem 1 implies that $H_*(L, \epsilon, n) \cong H_*(D_L) \cong H_*(\mathcal{M})$

We don't known A,B.

Solution :

- M is a connected and orientable.
- $B - A > 8\epsilon$

The ideal case

If we know A, B then the recovery will be simple.

$\rightarrow$ Choose L st $[L - 2\epsilon, L + 2\epsilon] \subset [A, B]$

$\rightarrow$ Theorem 1 implies that $H_*(L, \epsilon, n) \cong H_*(D_L) \cong H_*(\mathcal{M})$

We don't known A,B.

Solution :

- M is a connected and orientable.
- $B - A > 8\epsilon$

The procedure

Define: $N_\epsilon = \sup\lceil f(x)/2\epsilon \rceil \qquad L_{max} = 2\epsilon N_\epsilon \qquad L_i = L_{max} - 2i\epsilon$

1. Compute $H_*(L_i, \epsilon, n)$ for all $i = 1, ...N_\epsilon$
2. Define $i_* = 1 + min\{i \in \{1, ..., N_\epsilon\} : rank(H_*(L_i, \epsilon, n)) = 1\}$
3. Return $H_*(L_{i*}, \epsilon, n)$

Remarks : we have to choose r for the procedure, the next theorem tell us that if we choose r correctly, we have consistence.

The procedure

Define: $N_\epsilon = \sup\lceil f(x)/2\epsilon \rceil \qquad L_{max} = 2\epsilon N_\epsilon \qquad L_i = L_{max} - 2i\epsilon$

1. Compute $H_*(L_i, \epsilon, n)$ for all $i = 1, ...N_\epsilon$
2. Define $i_* = 1 + min\{i \in \{1, ..., N_\epsilon\} : rank(H_*(L_i, \epsilon, n)) = 1\}$
3. Return $H_*(L_{i*}, \epsilon, n)$

Remarks : we have to choose r for the procedure, the next theorem tell us that if we choose r correctly, we have consistence.

Theorem 2

Theorem

Let $\mathcal{M}$ be a m-dimensional closed, connected, orientable manifold in $\mathbb{R}^d$. Let $X_1, ..., X_n$ be sample from noisy density of $\mathcal{M}$. if we choose $r \to 0$ such that $nr^d \geq Dlog(n)$ with $D > (C^*_{\epsilon/2})^{-1}$. Applying the procedure then:

$$\mathbb{P}(H_*(L_{i*}, \epsilon, n) \cong H_*(\mathcal{M})) = 1$$

Remarks:

- Need to know $m, \epsilon, L_{max}, N_\epsilon$.

The proof of the theorem 2

Proof.

Recall that $N_\varepsilon = \lceil f(a)/2\varepsilon \rceil$ and $L_{max} = 2\varepsilon N_\varepsilon$.

Let $E = \bigcup_{i=1}^{N_\varepsilon} \{ \hat{\Lambda}_{L_{i-1}} = \Lambda_{L_i + 2\varepsilon} \hookrightarrow \hat{\Lambda}_{L_i + \varepsilon}^{(m,r)} \hookrightarrow \Lambda_{L_i} \hookrightarrow \hat{\Lambda}_{L_i - \varepsilon}^{(m,r)} \hookrightarrow \Lambda_{L_i - 2\varepsilon} = \Lambda_{L_{i+1}} \}$

Applying Lemma 2

$$\mathbb{P}(E) \geq 1 - 3 N_\varepsilon m e^{-C_\varepsilon^* m r^d}$$

Now, we have to show that $[L_{i_*} - 2\varepsilon, L_{i_*} + 2\varepsilon] \subset (A, B)$ Pcarré

Rq: Since we assume that $\mathcal{M}$ is connected we have $\beta_0(\mathcal{M}) = 1 = \beta_m(\mathcal{M})$ and if $L > B$ $\beta_m(\mathcal{M}') = 0$

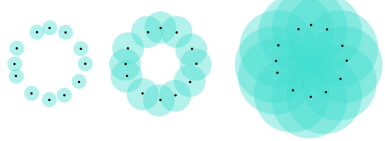
Our requirement that $L_{i-1} - L_i = 2\varepsilon$ and $B - A \geq 8\varepsilon$ guarantees that there are at least four consecutive $L_{i_1} > L_{i_2} > L_{i_3} > L_{i_4} \in (A, B)$.

For $k = 2, 3$ we have $[L_{i_k} - 2\varepsilon, L_i + 2\varepsilon] \subset (A, B)$ but it is not true for i_1.

Since $\hat{\beta}_m(L_{i_1}, \varepsilon, m) = 1$ or 0 then i_* is i_2 or i_3 but in both cases we have

$$[L_{i_*} - 2\varepsilon, L_{i_*} + 2\varepsilon] \subset (A, B)$$

Recover the persistent homology



$PH_*(f)$: tracks the evolution of the homology of D_L for $L = +\infty$ *to* $-\infty$

$\rightarrow$ We have show that we can recover D_L for all $L > 0$

$\rightarrow$ But L is continous

However, we propose this procedure : Let N_ϵ, L_{max}, L_i like previously Define $\hat{D}_\epsilon = \{\hat{D}_{L_i}(n, r)\}_{i \in \mathbb{Z}}$ and $\hat{PH}^\epsilon(f)$ the persistent homology of $\hat{D}_\epsilon$ as the estimator of $PH_*(f)$

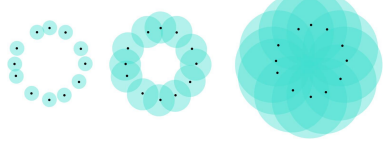
Theorem

if $r \to 0$ *and* $nr^d \to \infty$ *then :*

$$\mathbb{P}(d_B(\hat{PH}^\epsilon(f), PH_*(f)) \leq 5\epsilon) \geq 1 - 3N_\epsilon n e^{-C_{\epsilon/2} nr^d}$$

In particular, if ....

In practice



How we compute $H_*(U(\mathcal{X}, r))$?

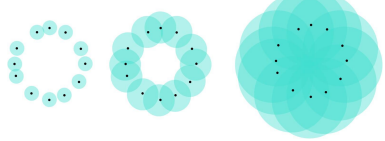
$\rightarrow$ nerf theorem $\Rightarrow H_*(Cech(\mathcal{X}), r) \cong H_*(U(\mathcal{X}), r) \Rightarrow$ Theorems holds for $Cech(\mathcal{X}), r)$

$\rightarrow$ But $H_*(Cech(\mathcal{X}), r)$ is too cost computationally

Like in the course, we want to use an approximation of the Cech: Rips.

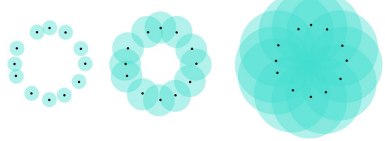
But nerf theorem don't holds for Rips, howerver we can show that the theorems still holds for Rips.

In practice



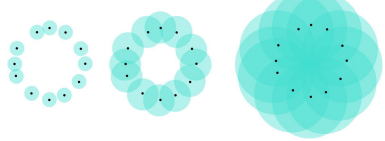
How we compute $H_*(U(\mathcal{X}, r))$?
$\rightarrow$ nerf theorem $\Rightarrow H_*(Cech(\mathcal{X}), r) \cong H_*(U(\mathcal{X}), r) \Rightarrow$ Theorems holds for $Cech(\mathcal{X}), r)$
$\rightarrow$ But $H_*(Cech(\mathcal{X}), r)$ is too cost computationally
Like in the course, we want to use an approximation of the Cech: Rips.
But nerf theorem don't holds for Rips, howerver we can show that the theorems still holds for Rips.

In practice



How we compute $H_*(U(\mathcal{X}, r))$?
$\rightarrow$ nerf theorem $\Rightarrow H_*(Cech(\mathcal{X}), r) \cong H_*(U(\mathcal{X}), r) \Rightarrow$ Theorems holds for $Cech(\mathcal{X}), r)$
$\rightarrow$ But $H_*(Cech(\mathcal{X}), r)$ is too cost computationally
Like in the course, we want to use an approximation of the Cech: Rips.
But nerf theorem don't holds for Rips, howerver we can show that the theorems still holds for Rips.

In practice



How we compute $H_*(U(\mathcal{X}, r))$?
$\rightarrow$ nerf theorem $\Rightarrow H_*(Cech(\mathcal{X}), r) \cong H_*(U(\mathcal{X}), r) \Rightarrow$ Theorems holds for $Cech(\mathcal{X}), r)$
$\rightarrow$ But $H_*(Cech(\mathcal{X}), r)$ is too cost computationally
Like in the course, we want to use an approximation of the Cech: Rips.
But nerf theorem don't holds for Rips, howerver we can show that the theorems still holds for Rips.

Compute the homology of the image map

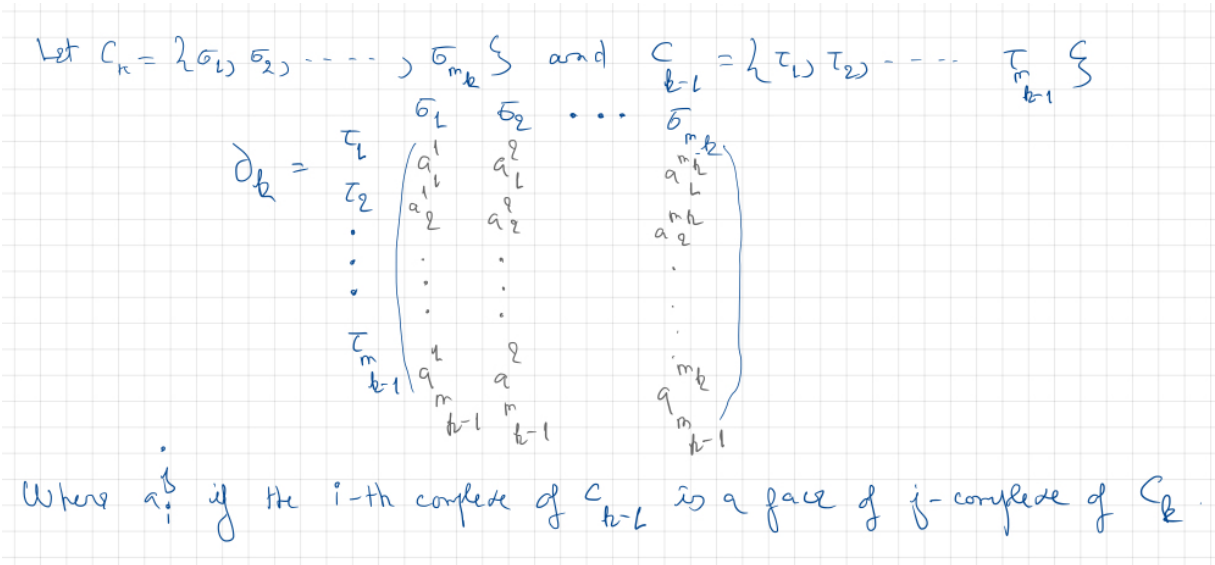
First let compute $\beta_k(\triangle)$ for $\triangle$ a simplicial complex.

- $\triangle_k = \{\sigma_1, ..., \sigma_2\}$
- The map $T_k : \triangle_k^\pi \to \mathbb{R}^k \quad \forall \sigma_i \in \triangle_k \, T_k(\sigma_i) = sgn(\pi) e_i$

We can define the boundary matrix ∂_k be $n_{k-1} \times n_k$ st :
$(\partial_k)_i = \sum_{\sigma \in \triangle_{k-1}} T_k(\sigma)(\text{image})$

So $L_k = \partial_{k+1}\partial_{k+1}^T + \partial_k \partial_k^T$ Then $\beta_k(\triangle) = rank(ker(L_k))$

Example:



Compute the homology of the image map

Let $\triangle^1$, $\triangle^2$ and the map $i_k : H_k(\triangle^1) \to H_k(\triangle^2)$, our goal is to compute the homology of $Im(i_k)$.

- $\triangle_k^1 = \{\sigma_1, ..., \sigma_{n_k^1}\}$
- $\triangle_k^2 = \{\sigma_1, ..., \sigma_{n_k}, \sigma_{n_k+1}, ..., \sigma_{n_k^2}\}$ -

$$\partial_k^{(2)} = \begin{bmatrix} \partial_k^{(1)} & \cdots \\ 0 & \ddots \end{bmatrix}$$

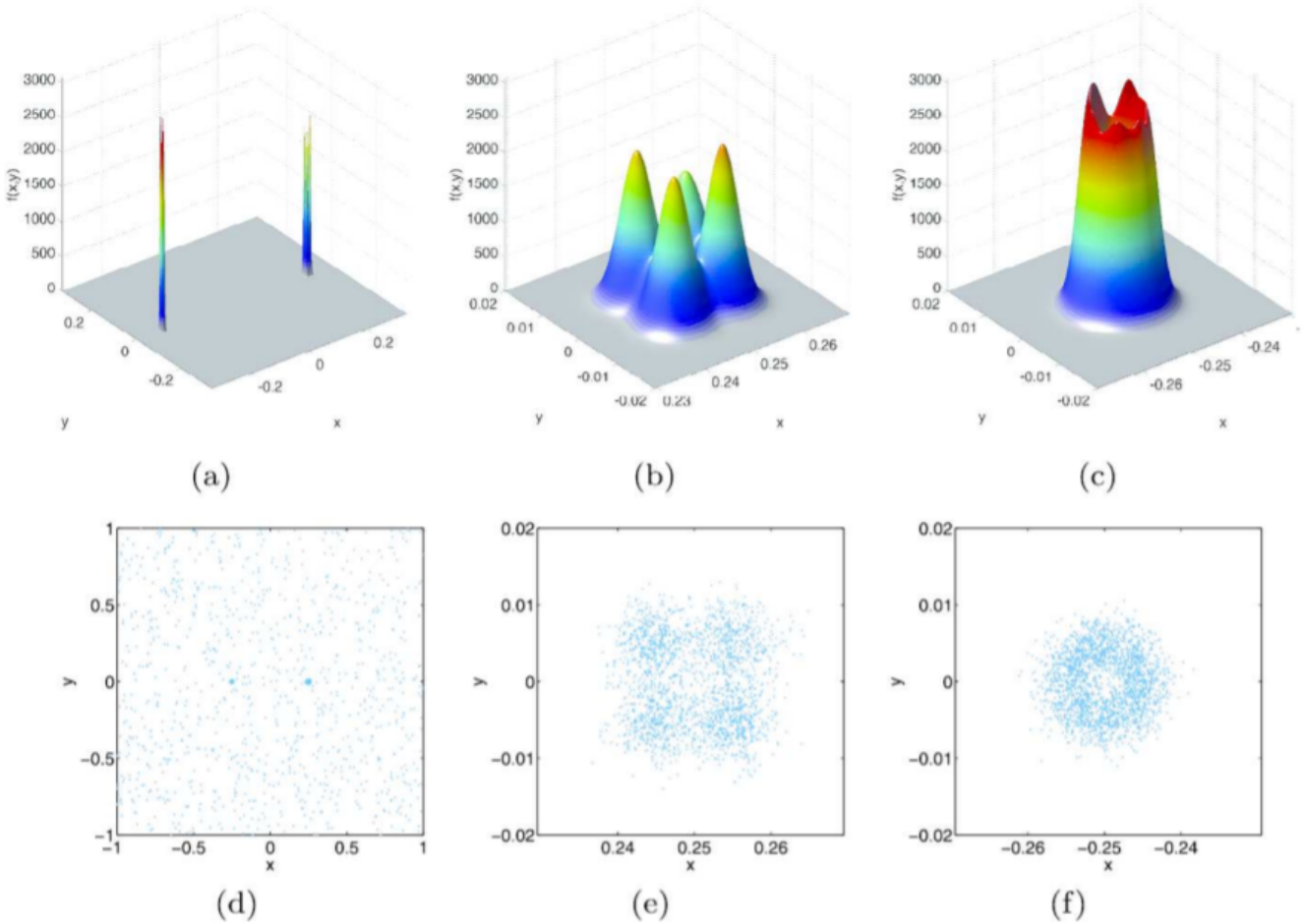
Now, if $\{v_1, ..., v_m\} \subset \mathbb{R}^{n_k^1}$ is a basis for $ker(L_k^1)$. Let $\hat{v}_i \in \mathbb{R}^{n_{k2}}$ be zeros padded version of v_i by linear algebra, we can find that:

$$rank(Im(i_k)) = rank(\hat{\partial}_{k+1}^{(2)}) - rank(\partial_{k+1}^{(2)})$$

where $\hat{\partial}_{k+1}^{(2)} = (\partial_{k+1}^{(2)}, \hat{v_1}, ..., \hat{v_m})$

Results on simuled data

- In the papers, there are results on binary classification, kernel regression and spectral clustering.
- We decide to show an application for Hierarchical clustering :



Results of the estimators

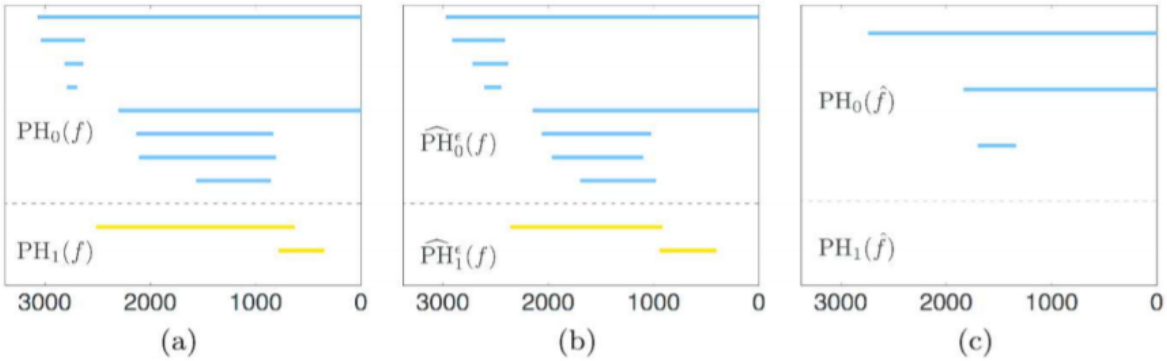
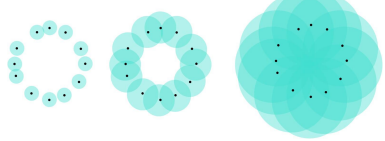


Figure 14. Estimating the persistent homology of the density function f presented in Figure 13. (a) The "true" barcode for the function f, that is, $\mathrm{PH}_*(f)$ (computed by sampling the density function on a fine grid). (b) The barcode computed from the estimator $\widehat{\mathrm{PH}}_*^{\varepsilon}(f)$. The parameters used are $n = 5000$, $r = 0.001$, $\varepsilon = 3.5$. (c) The barcode computed for the kernel density estimator – $\mathrm{PH}_*(\hat{f})$. The kernel parameters are the same as for $\widehat{\mathrm{PH}}_*^{\varepsilon}(f)$, the grid size taken is 500×500. Note that the estimator $\widehat{\mathrm{PH}}_*^{\varepsilon}(f)$ gives a result that is very similar to the true barcode. In both cases there are five significant features in H_0 and two significant features in H_1. The barcode for $\mathrm{PH}_*(\hat{f})$ only recover the coarse features, namely the two clusters, but completely ignores the finer structures. We note that for visualization purposes we filtered out the very small bars before drawing the barcodes here.

Conclusion



Thank you !

Source :

- Topological consistency via kernel estimation [Omer Bobrowski, Sayan Mukherjee, Jonathan E. Taylor]