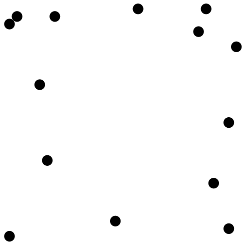
Handling noise and complexity blow-up in topological data analysis.

Mickaël Buchet

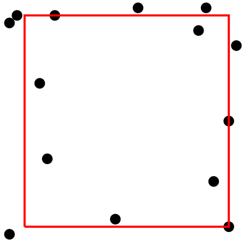
The Ohio State University

June 18, 2015

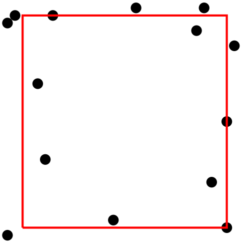
Topological inference

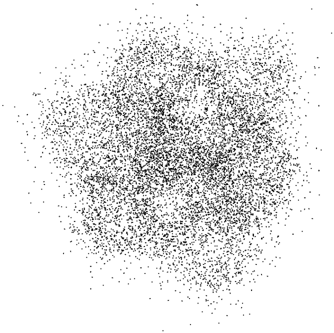


Topological inference



Topological inference





Usual pipeline

Usual pipeline

P		F_α		
Point set	$\longrightarrow$	Discrete representation of continuous objects	$\longrightarrow$	Topological information

Challenges

Challenges

- Accuracy of the representation

Challenges

- Accuracy of the representation
- Size of the data structure

Challenges

- Accuracy of the representation
- Size of the data structure
- Complexity of the process

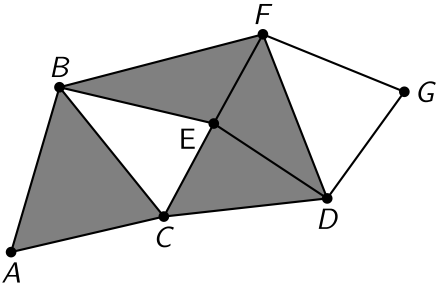
Challenges

- Accuracy of the representation
- Size of the data structure
- Complexity of the process
- Robustness to noise

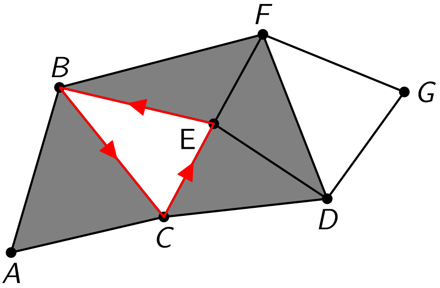
Outline

- Topological information
- Classical data structures
- Sparse data structures
- Handling noise and aberrant values
- Sparsification and parameter free analysis

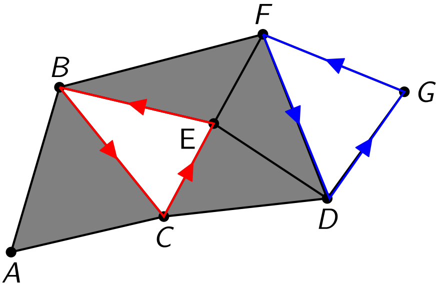
Homology



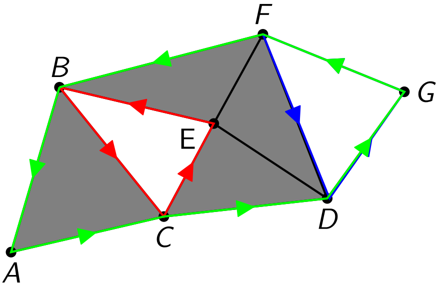
Homology



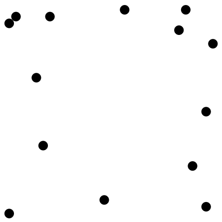
Homology



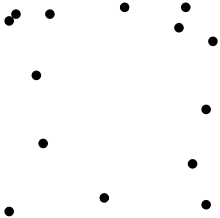
Homology

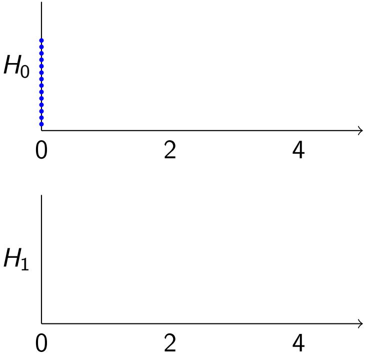


Persistence

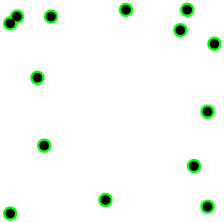


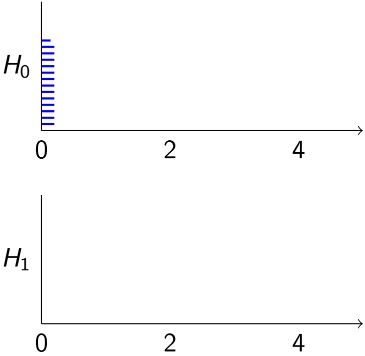
Persistence



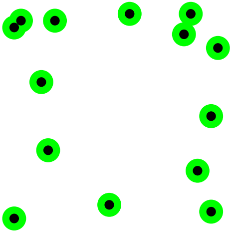


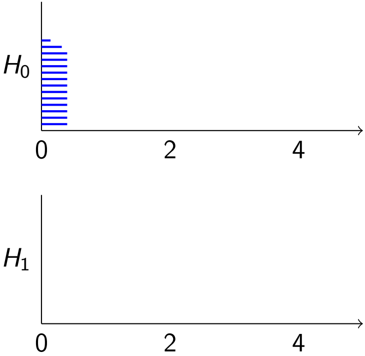
Persistence



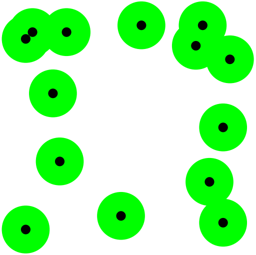


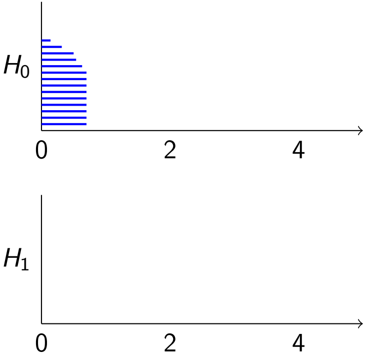
Persistence



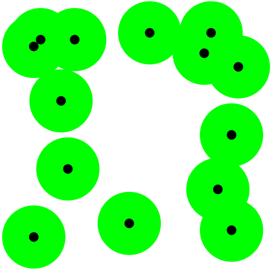


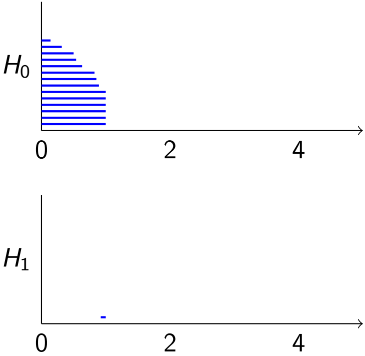
Persistence



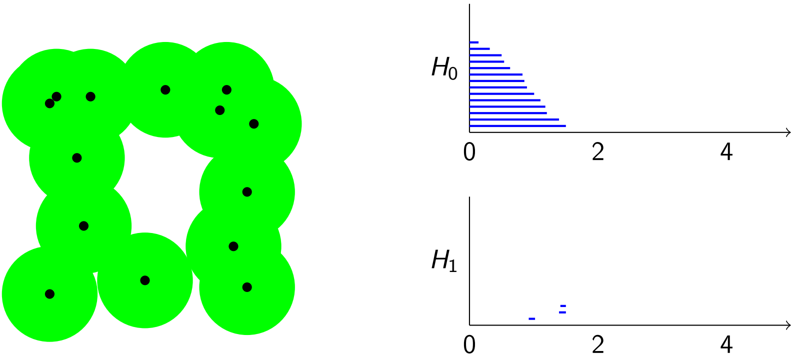


Persistence

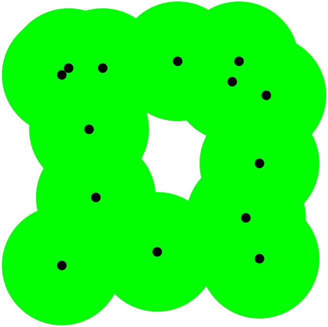


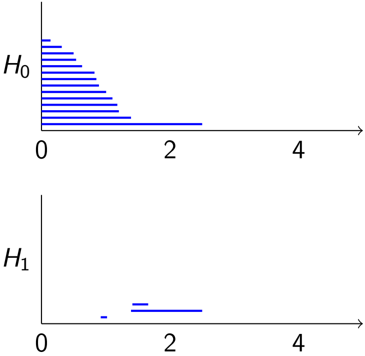


Persistence

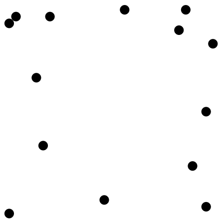


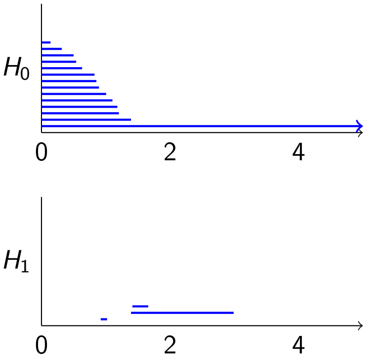
Persistence



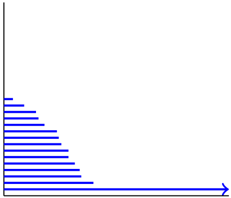


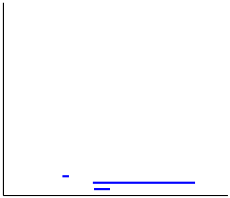
Persistence



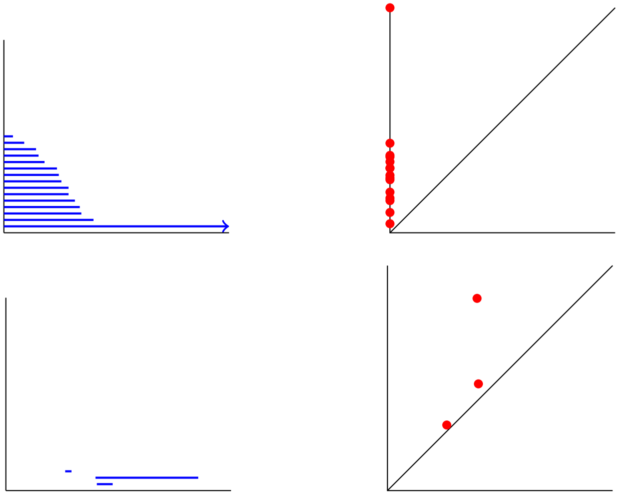


Persistent diagram representation





Persistent diagram representation



Outline

- Topological information
- Classical data structures
- Sparse data structures
- Handling noise and aberrant values
- Sparsification and parameter free analysis

Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

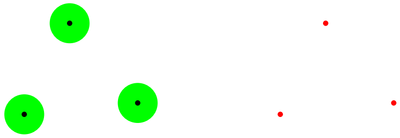
$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

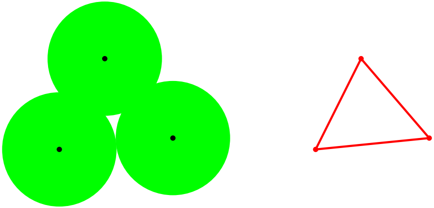


Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

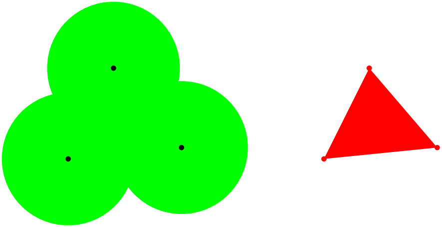


Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

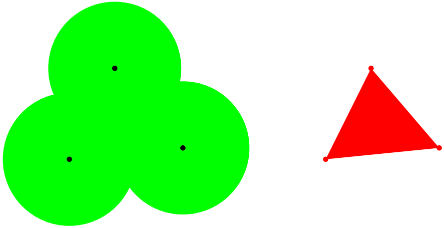


Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$



Theorem (Borsuk, 1948)

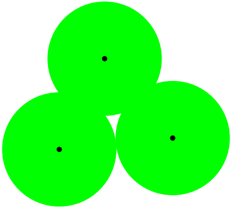
The Čech complex has the same homology as the union of balls if the space has the good cover property.

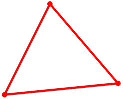
Čech complex

Definition

$(p_1, \ldots, p_l)$ belongs to the Čech for the parameter α, noted C_α, if :

$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$





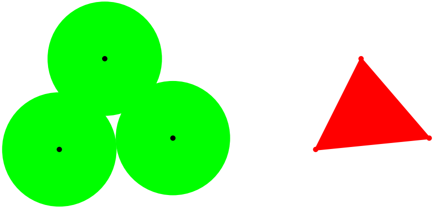
Theorem (Borsuk, 1948)

The Čech complex has the same homology as the union of balls if the space has the good cover property.

Rips complex

Definition

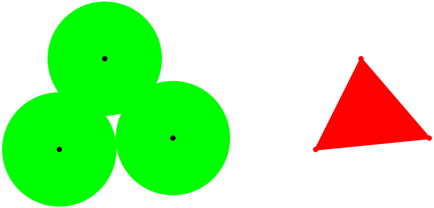
The Rips complex is the maximal complex with the same edges as the Čech complex.



Rips complex

Definition

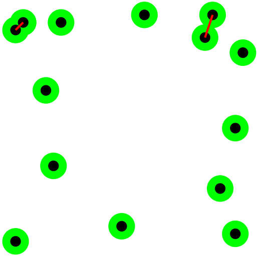
The Rips complex is the maximal complex with the same edges as the Čech complex.



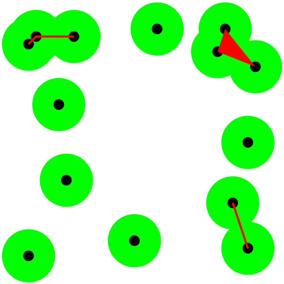
Proposition

$$C_\alpha \subset R_\alpha \subset C_{2\alpha}$$

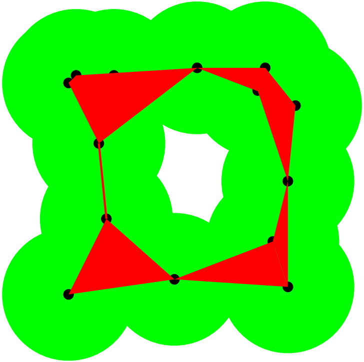
Illustration



Illustration



Illustration



Complexity

- Computing the persistence diagram of a filtered simplicial complex has complexity $O(N^3)$.

Complexity

- Computing the persistence diagram of a filtered simplicial complex has complexity $O(N^3)$.
- In practice, the computation time is linear in N.

Complexity

- Computing the persistence diagram of a filtered simplicial complex has complexity $O(N^3)$.
- In practice, the computation time is linear in N.
- A d-dimensional Rips complex has $\Theta(n^{d+1})$ simplexes.

Complexity

- Computing the persistence diagram of a filtered simplicial complex has complexity $O(N^3)$.
- In practice, the computation time is linear in N.
- A d-dimensional Rips complex has $\Theta(n^{d+1})$ simplexes.
- The computation of the persistence diagram of n points up to dimension d has complexity $O(n^{d+1})$.

Complexity

- Computing the persistence diagram of a filtered simplicial complex has complexity $O(N^3)$.
- In practice, the computation time is linear in N.
- A d-dimensional Rips complex has $\Theta(n^{d+1})$ simplexes.
- The computation of the persistence diagram of n points up to dimension d has complexity $O(n^{d+1})$.

- Unusable in high dimensions.

Nested Rips

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Nested Rips

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Solution: take two different parameters $\delta < \delta'$ and look at the image of $H_*(R_\delta)$ by the morphism induced by the inclusion $R_\delta \hookrightarrow R_{\delta'}$.

Nested Rips

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Solution: take two different parameters $\delta < \delta'$ and look at the image of $H_*(R_\delta)$ by the morphism induced by the inclusion $R_\delta \hookrightarrow R_{\delta'}$.

$$R_\delta \longrightarrow R_{\delta'}$$

Nested Rips

Sometimes, there is no parameter such that the Rips complex capture the topology of M.

Solution: take two different parameters $\delta < \delta'$ and look at the image of $H_*(R_\delta)$ by the morphism induced by the inclusion $R_\delta \hookrightarrow R_{\delta'}$.

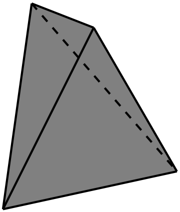
$$\begin{array}{ccc} R_\delta & \longrightarrow & R_{\delta'} \\ H_*(R_\delta) & \longrightarrow & H_*(R_{\delta'}) \end{array}$$

Outline

- Topological information
- Classical data structures
- Sparse data structures
- Handling noise and aberrant values
- Sparsification and parameter free analysis

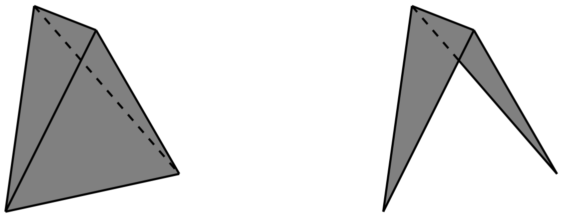
Collapses and contractions (I)

Attali, Lieutier and Salinas 2012, 2013



Collapses and contractions (I)

Attali, Lieutier and Salinas 2012, 2013



Collapses and contractions (II)

- Implicit construction of the simplicial complex.

Collapses and contractions (II)

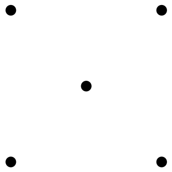
- Implicit construction of the simplicial complex.
- Reduction of the simplicial complex with topological guarantees.

Collapses and contractions (II)

- Implicit construction of the simplicial complex.
- Reduction of the simplicial complex with topological guarantees.
- Adapted to homology, not persistence.

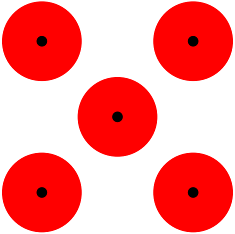
Sparse Rips

Sheehy, 2012



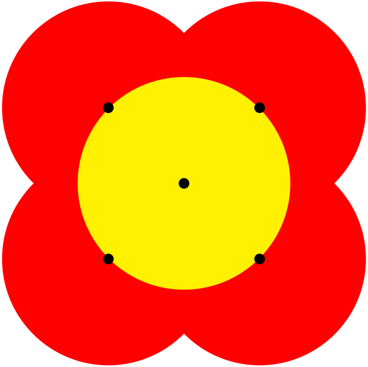
Sparse Rips

Sheehy, 2012



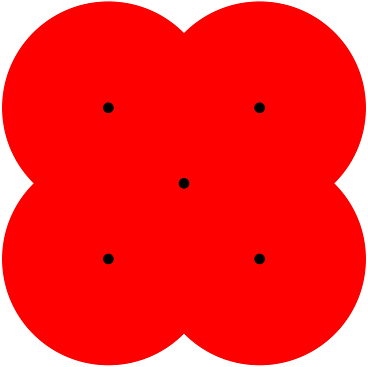
Sparse Rips

Sheehy, 2012



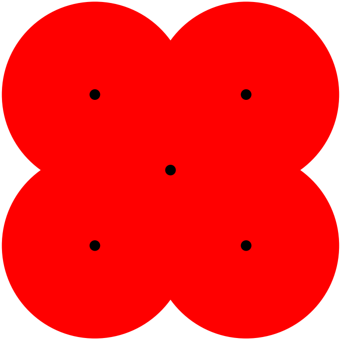
Sparse Rips

Sheehy, 2012



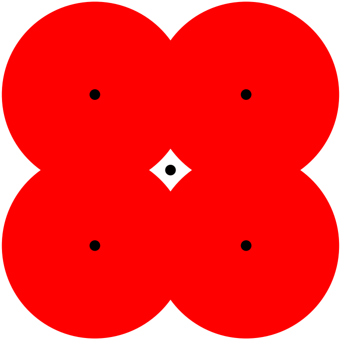
Topological noise

Naively removing points can create topological noise.

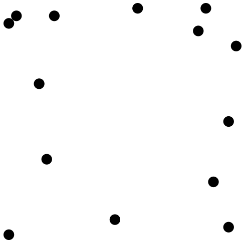


Topological noise

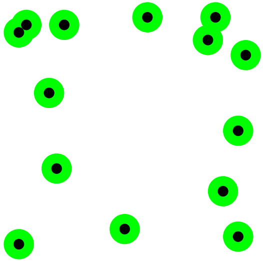
Naively removing points can create topological noise.



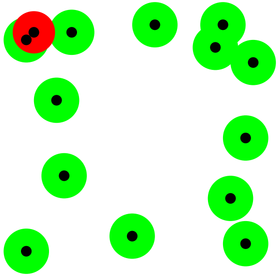
Illustration



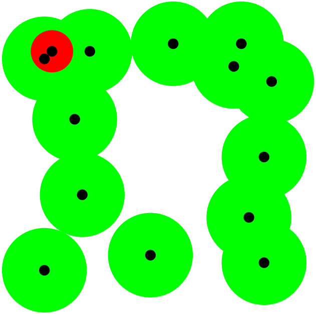
Illustration



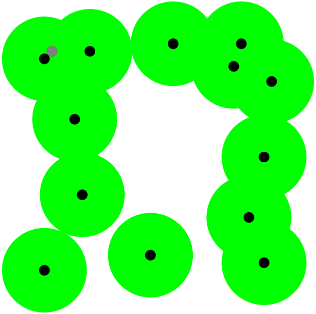
Illustration



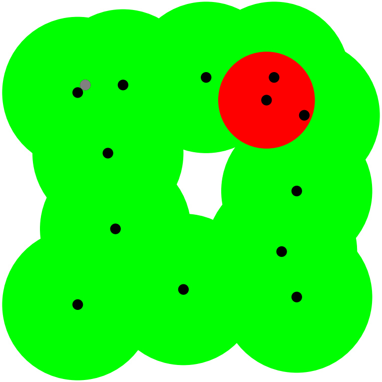
Illustration



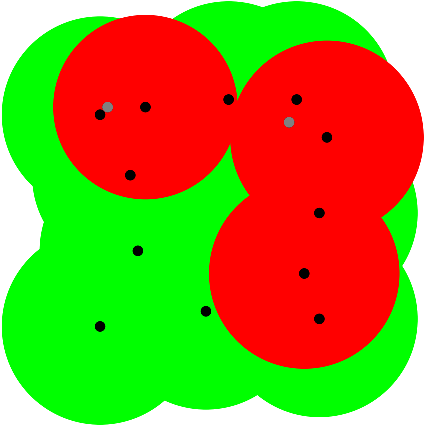
Illustration



Illustration



Illustration



Construction (I)

- Let $(p_1, \ldots, p_n)$ be a futhest point sampling of P.

Construction (I)

- Let $(p_1, \ldots, p_n)$ be a futhest point sampling of P.
  - p_1 is arbitrary and $\lambda_1 = \infty$.

Construction (I)

- Let $(p_1, \ldots, p_n)$ be a futhest point sampling of P.
  - p_1 is arbitrary and $\lambda_1 = \infty$.
  - $p_i = \operatorname{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1},)$ and $P_{i-1} = (p_1, \ldots, p_{i-1})$.

Construction (I)

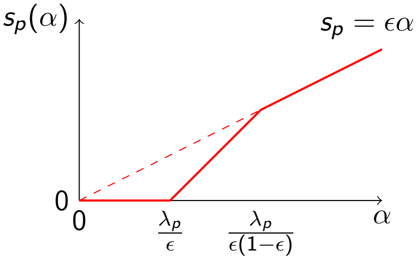
- Let $(p_1, \ldots, p_n)$ be a futhest point sampling of P.
  - p_1 is arbitrary and $\lambda_1 = \infty$.
  - $p_i = \mathrm{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1},)$ and $P_{i-1} = (p_1, \ldots, p_{i-1})$.
  - Fix $\lambda_i = d_{\mathbb{X}}(p_i, P_{i-1})$.

Construction (I)

- Let $(p_1, \ldots, p_n)$ be a futhest point sampling of P.
  - p_1 is arbitrary and $\lambda_1 = \infty$.
  - $p_i = \mathrm{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1},)$ and $P_{i-1} = (p_1, \ldots, p_{i-1})$.
  - Fix $\lambda_i = d_{\mathbb{X}}(p_i, P_{i-1})$.
- Fix $\bar{N}_\gamma = \{p \in P | \lambda_p \geq \gamma\}$.

Construction (I)

- Let $(p_1, \dots, p_n)$ be a futhest point sampling of P.
  - p_1 is arbitrary and $\lambda_1 = \infty$.
  - $p_i = \mathrm{argmax}_{p \in P \setminus P_{i-1}} d_{\mathbb{X}}(p, P_{i-1},)$ and $P_{i-1} = (p_1, \dots, p_{i-1})$.
  - Fix $\lambda_i = d_{\mathbb{X}}(p_i, P_{i-1})$.
- Fix $\bar{N}_\gamma = \{p \in P | \lambda_p \geq \gamma\}$.
- Perturbed metric : $f_\alpha(p,q) = d_{\mathbb{X}}(p,q) + s_p(\alpha) + s_q(\alpha)$.



Construction (II)

Definition

The sparse Rips complex is given by:

$$Q_\alpha = \{\sigma \subset \bar{N}_{\epsilon(1-\epsilon)\alpha} | \forall p, q \in \sigma,\ f_\alpha(p, q) < 2\alpha\}$$

Construction (II)

Definition

The sparse Rips complex is given by:

$$Q_\alpha = \{\sigma \subset \bar{N}_{\epsilon(1-\epsilon)\alpha} | \forall p, q \in \sigma,\ f_\alpha(p,q) < 2\alpha\}$$

Definition

The sparse Rips filtration is given by:

$$S_\beta = \bigcup_{\alpha \leq \beta} Q_\alpha.$$

Properties of sparse Rips

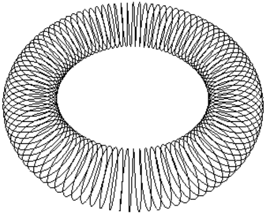
Theorem (Sheehy, 2012)

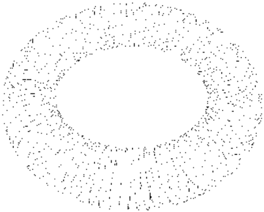
$\{S_\alpha\}$ contains $O(C^l n)$ simplexes where l is the intrinsic dimension of the underlying object.

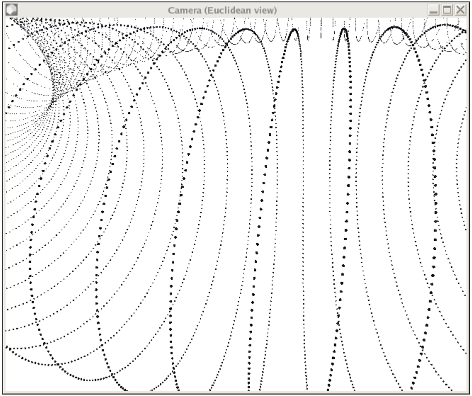
Theorem (Sheehy, 2012)

$\{S_\alpha\}$ is $\frac{1}{1-\epsilon}$-interleaved with the Rips filtration $\{R_\alpha\}$.

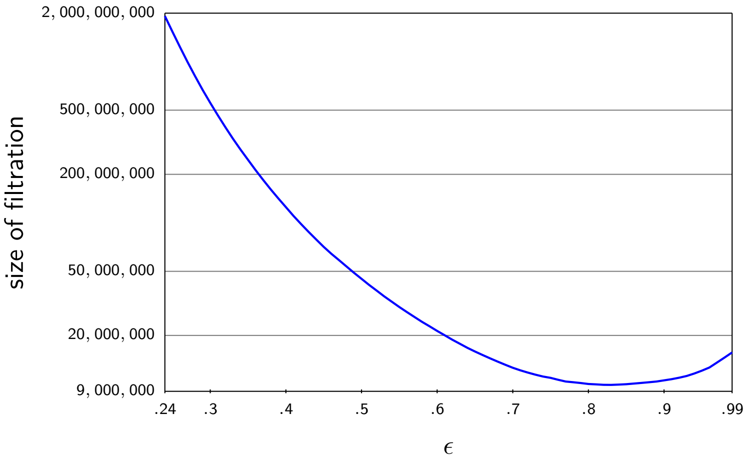
Spiral



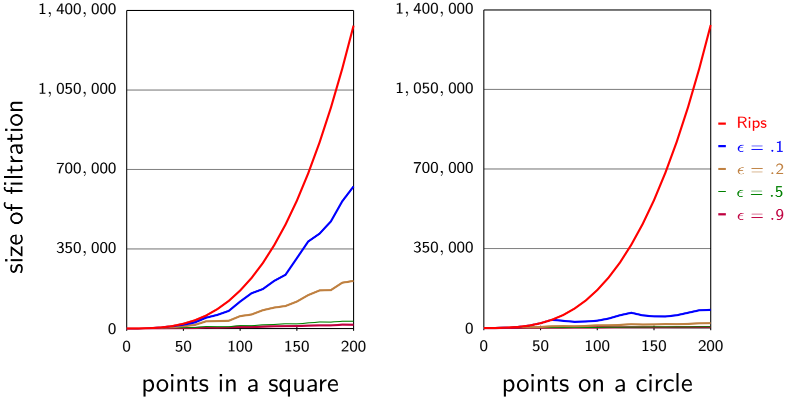




Size of the filtration depending on ϵ

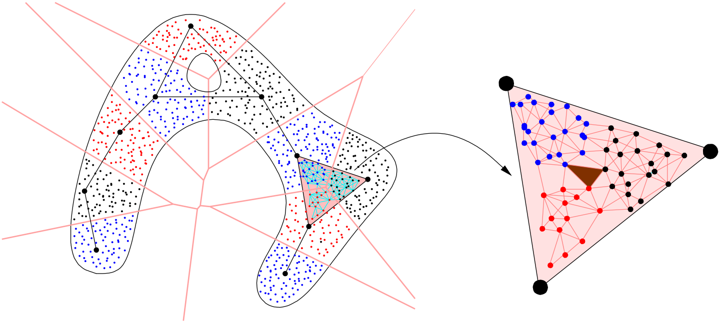


Intrinsic dimension influence



Graph induced complex (I)

Dey, Fan and Wang, 2013



Graph induced complex (II)

- Small construction with good guarantees and complexity for dimension 1.

Graph induced complex (II)

- Small construction with good guarantees and complexity for dimension 1.
- Extensions to higher dimesion needs more complex computations.

Pipeline

Point cloud

Pipeline

$$\text{Point cloud} \xrightarrow{\rotatebox{90}{d_K^{-1}}} \quad \text{Balls}$$

Pipeline

$$\text{Point cloud} \xrightarrow{d_K^{-1}} \text{Balls} \xrightarrow{\text{Nerve}} \text{Čech}$$

Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "d_K^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{Čech} \arrow[d, "\text{2-int}"] \\
& & \text{Rips}
\end{tikzcd}
$$

Pipeline

$$\begin{tikzcd}
\text{Point cloud} \arrow[r, "d_K^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{Čech} \arrow[d, "\text{2-int}"] \\
& \text{Sparse Rips} & \text{Rips} \arrow[l, "\frac{1}{1-\epsilon}\text{-int}"']
\end{tikzcd}$$

Pipeline

$$\begin{tikzcd}
\text{Point cloud} \arrow[r, "d_K^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{Čech} \arrow[d, "2\text{-int}"] \\
\text{Information} & \text{Sparse Rips} \arrow[l, "\text{Algorithm}"'] & \text{Rips} \arrow[l, "\frac{1}{1-\epsilon}\text{-int}"']
\end{tikzcd}$$

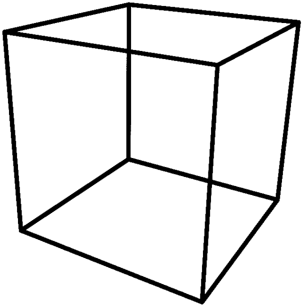
Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "d_K^{-1}"] \arrow[d, red] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{Čech} \arrow[d, "2\text{-int}"] \\
\text{Information} & \text{Sparse Rips} \arrow[l, "\text{Algorithm}"'] & \text{Rips} \arrow[l, "\frac{1}{1-\epsilon}\text{-int}"']
\end{tikzcd}
$$

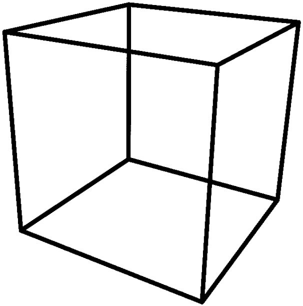
Outline

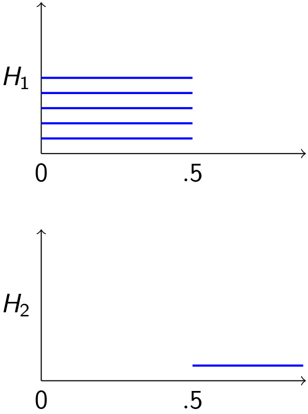
- Topological information
- Classical data structures
- Sparse data structures
- Handling noise and aberrant values
- Sparsification and parameter free analysis

Outliers (I)

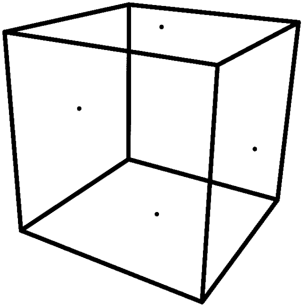


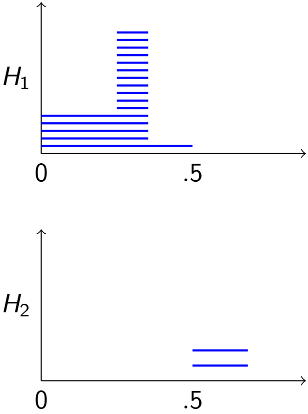
Outliers (I)





Outliers (II)





Distance to a measure

Definition (Chazal, Cohen-Steiner, Mérigot, 2011)

Let μ be a measure and $m \in]0, 1[$, then

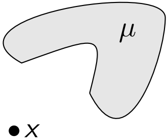
$$d_{\mu,m}(x) = \frac{1}{\sqrt{m}} \inf_{\nu \in \mathrm{Sub}_m(\mu)} W_2(m\delta_x, \nu)$$

Distance to a measure

Definition (Chazal, Cohen-Steiner, Mérigot, 2011)

Let μ be a measure and $m \in]0,1[$, then

$$d_{\mu,m}(x) = \frac{1}{\sqrt{m}} \inf_{\nu \in \mathrm{Sub}_m(\mu)} W_2(m\delta_x, \nu)$$

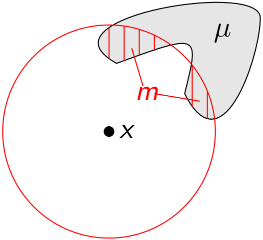


Distance to a measure

Definition (Chazal, Cohen-Steiner, Mérigot, 2011)

Let μ be a measure and $m \in]0,1[$, then

$$d_{\mu,m}(x) = \frac{1}{\sqrt{m}} \inf_{\nu \in \mathrm{Sub}_m(\mu)} W_2(m\delta_x, \nu)$$



Case of an empirical measure

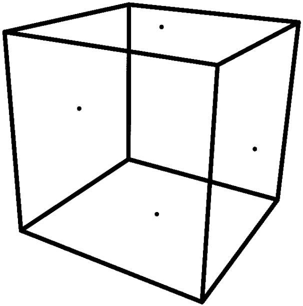
Proposition

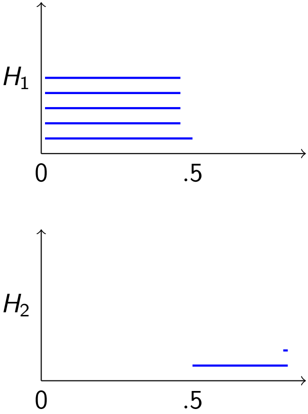
Let μ be the empirical on P and $k = m|P|$ is an integer then:

$$d_{\mu,m}(x) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} d_{\mathbb{X}}(x, p_i(x))^2}$$

where $p_i(x)$ is the i^{th}-neighbour of x in P.

Results





Barycentric decomposition

Proposition

In Euclidean spaces, the distance to an empirical measure is a power distance.

Barycentric decomposition

Proposition

In Euclidean spaces, the distance to an empirical measure is a power distance.

$$d_{\mu,m}(x)^2 = \frac{1}{k}\sum_{i=1}^{k} ||x - p_i(x)||^2$$

Barycentric decomposition

Proposition

In Euclidean spaces, the distance to an empirical measure is a power distance.

$$
\begin{aligned}
d_{\mu,m}(x)^2 &= \frac{1}{k}\sum_{i=1}^{k} ||x - p_i(x)||^2 \\
&= ||x - bar(x)||^2 + \frac{1}{k}\sum_{i=1}^{k} ||p_i(x) - bar(x)||^2
\end{aligned}
$$

where $bar(x) = \sum_{i=1}^{k} p_i(x)$

Barycentric decomposition

Proposition

In Euclidean spaces, the distance to an empirical measure is a power distance.

$$
\begin{aligned}
d_{\mu,m}(x)^2 &= \frac{1}{k}\sum_{i=1}^{k} ||x - p_i(x)||^2 \\
&= ||x - bar(x)||^2 + \frac{1}{k}\sum_{i=1}^{k} ||p_i(x) - bar(x)||^2 \\
&= ||x - bar(x)||^2 + w_{bar(x)}^2
\end{aligned}
$$

where $bar(x) = \sum_{i=1}^{k} p_i(x)$

Barycentric decomposition

Proposition

In Euclidean spaces, the distance to an empirical measure is a power distance.

$$
\begin{aligned}
d_{\mu,m}(x)^2 &= \frac{1}{k}\sum_{i=1}^{k} ||x - p_i(x)||^2 \\
&= ||x - bar(x)||^2 + \frac{1}{k}\sum_{i=1}^{k} ||p_i(x) - bar(x)||^2 \\
&= ||x - bar(x)||^2 + w_{bar(x)}^2 \\
&= \min_{b\in B} \left(||x-b||^2 + w_b^2\right)
\end{aligned}
$$

where $bar(x) = \sum_{i=1}^{k} p_i(x)$ and B is the set of all barycentres of k points.

Power distances

Definition

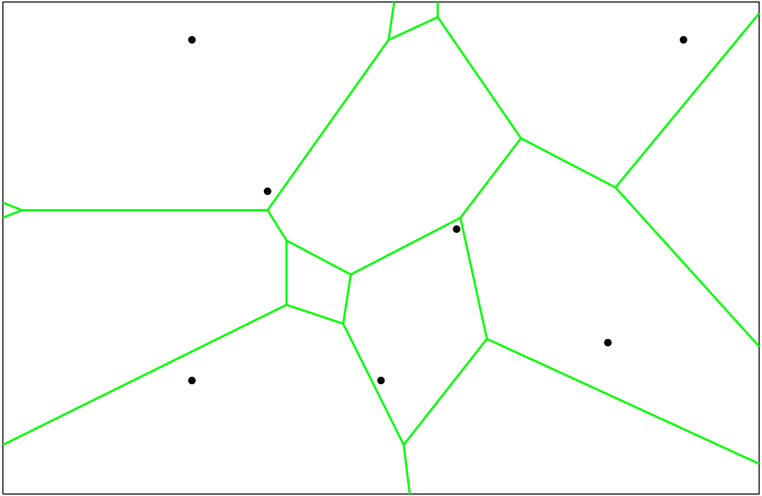
Let P be a point set and $w : P \to \mathbb{R}$ a weight function. The power distance associated with (P, w) is defined by:

$$f(x) = \sqrt{\min_{p \in P} ||x - p||^2 + w(p)^2}$$

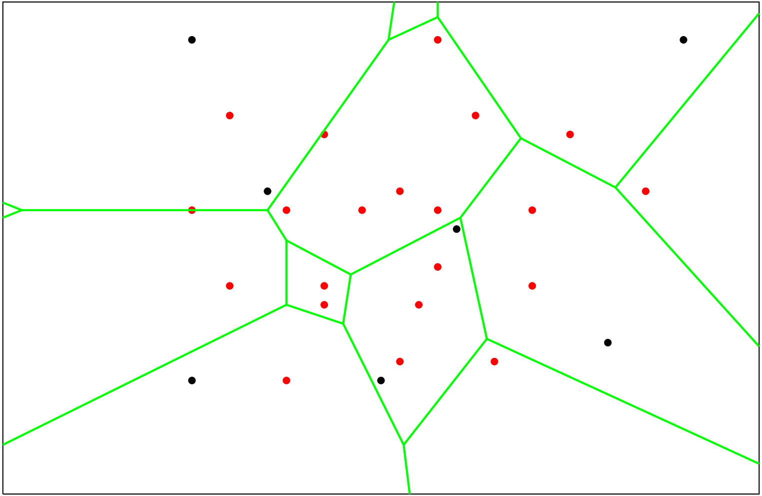
Sub-level sets of a power distance f are unions of balls.

$$f^{-1}(]-\infty, \alpha]) = \bigcup_{p \in P} \bar{B}(p, \sqrt{\alpha^2 - w(p)^2})$$

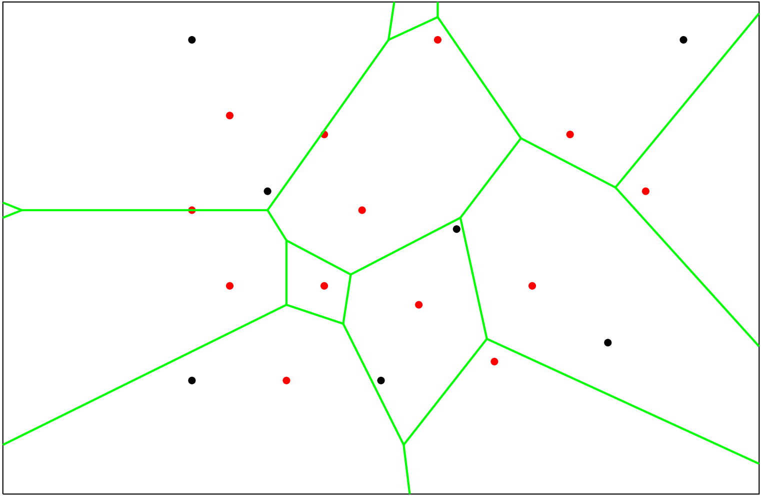
k^{th}-order Voronoi diagram



k^{th}-order Voronoi diagram



k^{th}-order Voronoi diagram



Size of k^{th}-order Voronoi diagram

Sub-level sets of $d_{\mu,m}$ are unions of balls.

$$d_{\mu,m}^{-1}(]-\infty,\alpha]) = \bigcup_{b\in B} \bar{B}(b, \sqrt{\alpha^2 - w_b^2})$$

Theorem (Clarkson, Shor, 1989)

The number of non-empty cells in Voronoi diagrams from order 1 to k is

$$O\left(n^{\lfloor \frac{d+1}{2} \rfloor} k^{\lceil \frac{d+1}{2} \rceil}\right).$$

Using the weighted metric

We used a weighted Rips filtration to compute the persistence diagram:

$$R_\alpha = \left\{ \sigma \subset P | \forall p, q \in P,\ d_{\mathbb{X}}(p,q) \leq \sqrt{\alpha^2 - w_p^2} + \sqrt{\alpha^2 - w_q^2} \right\}$$

Using the weighted metric

We used a weighted Rips filtration to compute the persistence diagram:

$$R_\alpha = \left\{\sigma \subset P | \forall p, q \in P,\ d_{\mathbb{X}}(p,q) \leq \sqrt{\alpha^2 - w_p^2} + \sqrt{\alpha^2 - w_q^2}\right\}$$

This structure induces a metric $\tilde{d}$.

Using the weighted metric

We used a weighted Rips filtration to compute the persistence diagram:

$$R_\alpha = \left\{\sigma \subset P | \forall p, q \in P,\ d_{\mathbb{X}}(p,q) \leq \sqrt{\alpha^2 - w_p^2} + \sqrt{\alpha^2 - w_q^2}\right\}$$

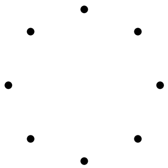
This structure induces a metric $\tilde{d}$.
Replacing $d_{\mathbb{X}}$ by $\tilde{d}$ causes loss of properties on the size of the sparse filtration.

Using the weighted metric

We used a weighted Rips filtration to compute the persistence diagram:

$$R_\alpha = \left\{ \sigma \subset P | \forall p, q \in P,\ d_{\mathbb{X}}(p,q) \leq \sqrt{\alpha^2 - w_p^2} + \sqrt{\alpha^2 - w_q^2} \right\}$$

This structure induces a metric $\tilde{d}$.
Replacing $d_{\mathbb{X}}$ by $\tilde{d}$ causes loss of properties on the size of the sparse filtration.



Adaptation to the weighted Rips

Definition

The sparse weighted Rips is defined by:

$$T_\alpha = R_\alpha \bigcap S_\alpha.$$

Adaptation to the weighted Rips

Definition

The sparse weighted Rips is defined by:

$$T_\alpha = R_\alpha \bigcap S_\alpha.$$

Theorem (Buchet et al., 2015)

R_α and T_α are κ-interleaved where $\kappa = 1 + \frac{\sqrt{1+t^2}\epsilon}{1-\epsilon}$, id est :

$$\begin{tikzcd}
R_\alpha \arrow[r, hook] & R_{\kappa\alpha} \\
T_\alpha \arrow[u, hook] \arrow[ur, hook] \arrow[r, hook] & T_{\kappa\alpha} \arrow[u, hook]
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
R_\alpha \arrow[r, hook] \arrow[dr, "\pi_{\frac{\alpha}{1-\epsilon}}"] & R_{\kappa\alpha} \\
T_\alpha \arrow[u, hook] \arrow[r, hook] & T_{\kappa\alpha} \arrow[u, hook]
\end{tikzcd}$$

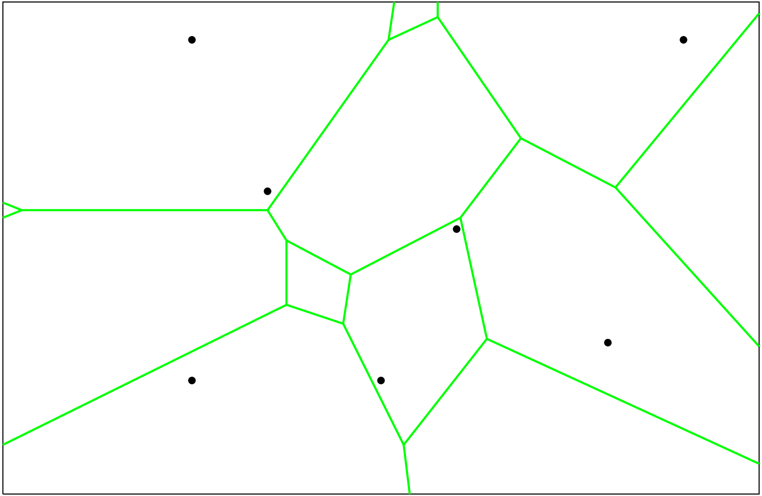
Witnessed k-distance

Approximation by sampling barycentres.

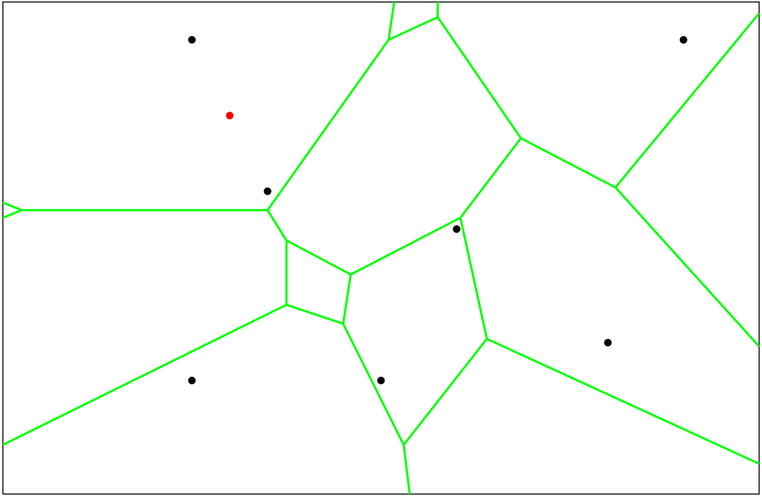
Definition (Guibas, Mérigot, Morozov, 2011)

$$d^W_{\mu,m}(x) = \min_{p \in P} \sqrt{||x - bar(p)||^2 + w^2_{bar(p)}}$$

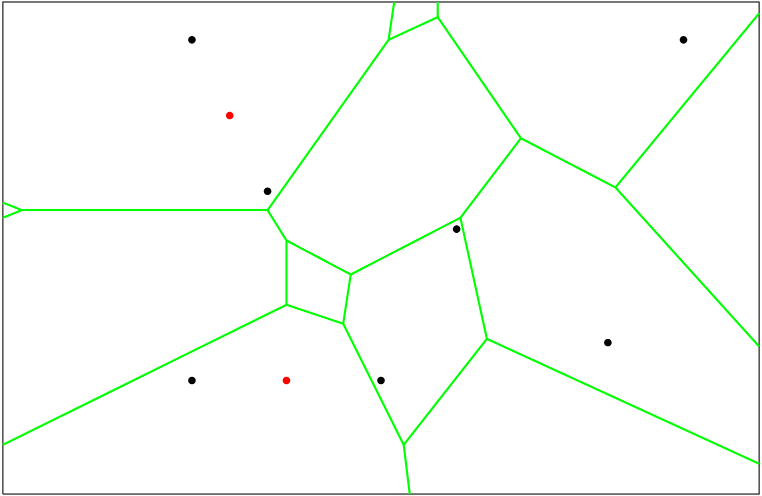
Building the witnessed k-distance



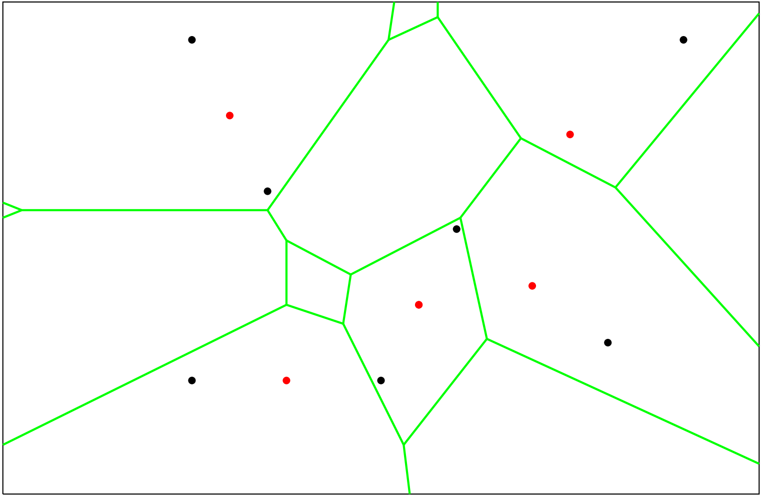
Building the witnessed k-distance



Building the witnessed k-distance



Building the witnessed k-distance



Guarantees of the witnessed k-distance

Approximation by sampling barycentres.

Definition (Guibas, Mérigot, Morozov, 2011)

$$d^W_{\mu,m}(x) = \min_{p \in P} \sqrt{||x - bar(p)||^2 + w^2_{bar(p)}}$$

Theorem (GMM, 2011; Buchet et al., 2015)

$$d_{\mu,m} \leq d^W_{\mu,m} \leq \sqrt{6} d_{\mu,m}$$

Approximation supported by the points

Using a power distance supported by input points.

Definition (Buchet et al., 2015)

$$d^P_{\mu,m}(x) = \min_{p \in P} \sqrt{d_{\mathbb{X}}(x,p)^2 + d_{\mu,m}(p)^2}$$

Approximation supported by the points

Using a power distance supported by input points.

Definition (Buchet et al., 2015)

$$d_{\mu,m}^{P}(x) = \min_{p \in P} \sqrt{d_{\mathbb{X}}(x,p)^2 + d_{\mu,m}(p)^2}$$

Theorem (Buchet et al., 2015)

In Euclidean space:

$$\frac{1}{\sqrt{2}} d_{\mu,m} \leq d_{\mu,m}^{P} \leq \sqrt{3} d_{\mu,m}$$

Approximation supported by the points

Using a power distance supported by input points.

Definition (Buchet et al., 2015)

$$d^P_{\mu,m}(x) = \min_{p \in P} \sqrt{d_{\mathbb{X}}(x,p)^2 + d_{\mu,m}(p)^2}$$

Theorem (Buchet et al., 2015)

In Euclidean space:

$$\frac{1}{\sqrt{2}} d_{\mu,m} \leq d^P_{\mu,m} \leq \sqrt{3} d_{\mu,m}$$

In any metric space:

$$\frac{1}{\sqrt{2}} d_{\mu,m} \leq d^P_{\mu,m} \leq \sqrt{5} d_{\mu,m}$$

All these bounds are tight.

Pipeline

Point cloud

Pipeline

Point cloud $\xrightarrow{{d^{P}_{\mu,m}}^{-1}}$ Balls

Pipeline

$$\text{Point cloud} \xrightarrow{{d^P_{\mu,m}}^{-1}} \text{Balls} \xrightarrow{\text{Nerve}} \text{w. Čech}$$

Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "{d_{\mu,m}^{P}}^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{w. Čech} \arrow[d, "\text{2-int}"] \\
 & & \text{w. Rips}
\end{tikzcd}
$$

Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "{d^P_{\mu,m}}^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{w. Čech} \arrow[d, "\text{2-int}"] \\
 & \text{s. w. Rips} & \text{w. Rips} \arrow[l, "\kappa\text{-int}"']
\end{tikzcd}
$$

Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "{d^{P}_{\mu,m}}^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{w. Čech} \arrow[d, "\text{2-int}"] \\
\text{Information} & \text{s. w. Rips} \arrow[l, "\text{Algorithm}"'] & \text{w. Rips} \arrow[l, "\kappa\text{-int}"']
\end{tikzcd}
$$

Pipeline

$$
\begin{tikzcd}
\text{Point cloud} \arrow[r, "{d_{\mu,m}^{P}}^{-1}"] & \text{Balls} \arrow[r, "\text{Nerve}"] & \text{w. Čech} \arrow[d, "\text{2-int}"] \\
\text{Information} & \text{s. w. Rips} \arrow[l, "\text{Algorithm}"'] & \text{w. Rips} \arrow[l, "\kappa\text{-int}"']
\end{tikzcd}
$$

Outline

- Topological information
- Classical data structures
- Sparse data structures
- Handling noise and aberrant values
- Sparsification and parameter free analysis

Parameter free analysis

Is it possible to obtain a good analysis in an (almost) parameter-free mehod?

Parameter free analysis

Is it possible to obtain a good analysis in an (almost) parameter-free mehod?

Yes, if we have a good sampling.

Good samplings for the distance to a measure

We assume that we have a point cloud P describing an underlying compact set K in a metric space $\mathbb{X}$. Given a mass parameter m, what is a "good" sampling to use the distance to a measure?

Good samplings for the distance to a measure

We assume that we have a point cloud P describing an underlying compact set K in a metric space $\mathbb{X}$. Given a mass parameter m, what is a "good" sampling to use the distance to a measure?

The (ϵ, r) sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon.$

Good samplings for the distance to a measure

We assume that we have a point cloud P describing an underlying compact set K in a metric space $\mathbb{X}$. Given a mass parameter m, what is a "good" sampling to use the distance to a measure?

The (ϵ, r) sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon.$
2. $\forall y \in \mathbb{X},\ d_K(y) \leq r \implies d_K(y) \leq d_{\mu,m}(y) + \epsilon.$

Good samplings for the distance to a measure

We assume that we have a point cloud P describing an underlying compact set K in a metric space $\mathbb{X}$. Given a mass parameter m, what is a "good" sampling to use the distance to a measure?

The (ϵ, ∞) sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon$.
2. $\forall y \in \mathbb{X},\ d_K(y) \leq d_{\mu,m}(y) + \epsilon$.

Good samplings for the distance to a measure

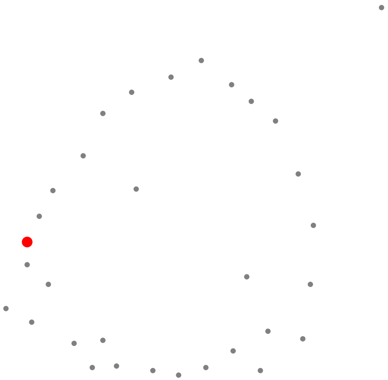
We assume that we have a point cloud P describing an underlying compact set K in a metric space $\mathbb{X}$. Given a mass parameter m, what is a "good" sampling to use the distance to a measure?

The (ϵ, r, c) uniform sampling.

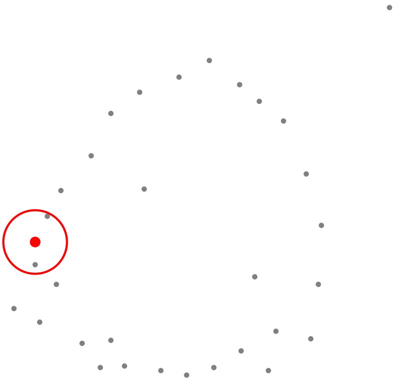
1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon.$
2. $\forall y \in \mathbb{X},\ d_K(y) \leq d_{\mu,m}(y) + \epsilon.$
3. $\forall p \in P,\ d_{\mu,m}(p) \geq \frac{\epsilon}{c}.$

Decimation

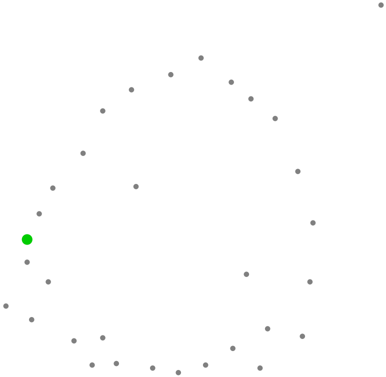
Decimation



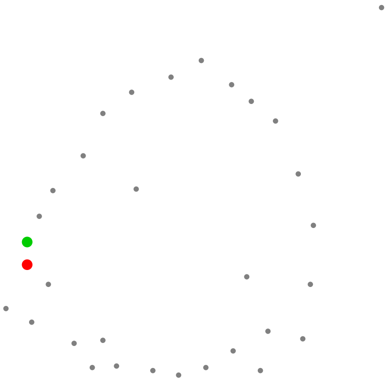
Decimation



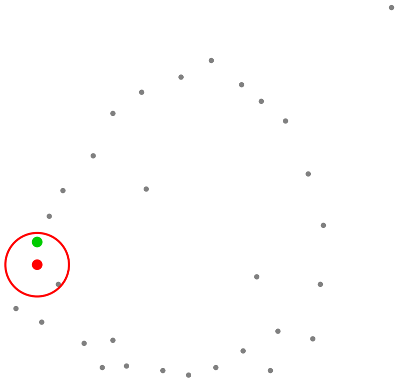
Decimation



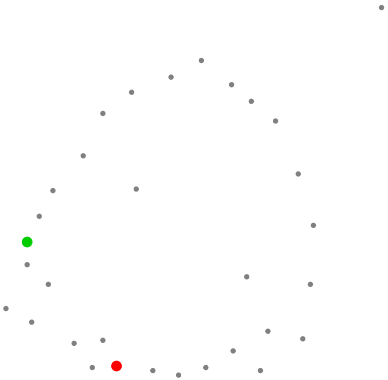
Decimation



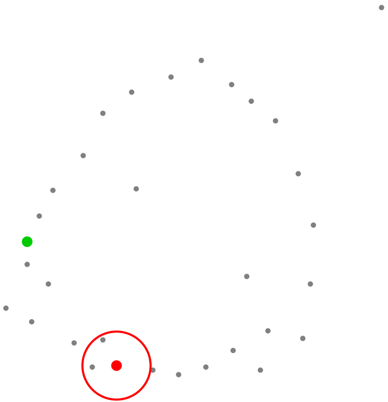
Decimation



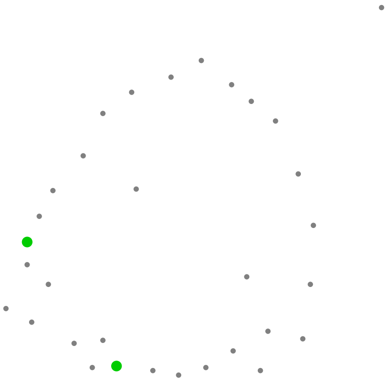
Decimation



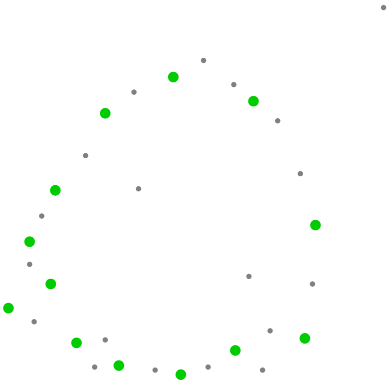
Decimation



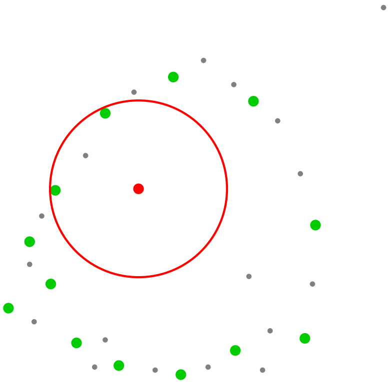
Decimation



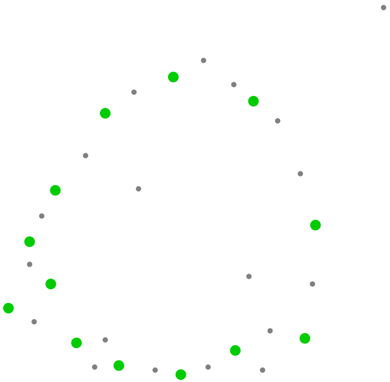
Decimation



Decimation



Decimation



Algorithm

1. $Q_0 = \emptyset$
2. Sort P according to increasing distance to the empirical measure.
3. For i from 1 to $n = |P|$, if $B(p_i, 2d_{\mu,m}(p_i)) \cap Q_{i-1} = \emptyset$:
    - then $Q_i = Q_{i-1} \cup p_i$
    - else $Q_i = Q_{i-1}$.

Guarantees

Theorem

If P is an (ϵ, ∞) sampling of K then:

$$d_H(Q_n, K) \leq 7\epsilon.$$

Theorem

If P is an (ϵ, ∞) uniform sampling of $K \subset \mathbb{R}^d$, with $\epsilon < \frac{1}{28}\mathrm{wfs}(K)$. Then for all α, $\alpha' \in [7\epsilon, \mathrm{wfs}(K) - 7\epsilon]$ such that $\alpha' - \alpha > 14\epsilon$ and for all $\lambda \in (0, \mathrm{wfs}(K))$, we have

$$H_*(X^\lambda) \cong H_*(C_\alpha(Q_n) \hookrightarrow C_{\alpha'}(Q_n)).$$

Adaptive setting

We assume that a feature size function f exists on K which is 1-Lipschitz. The sampling conditions become, for an (ϵ, ∞, c) uniform sampling.

Adaptive setting

We assume that a feature size function f exists on K which is 1-Lipschitz. The sampling conditions become, for an (ϵ, ∞, c) uniform sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon f(x)$.

Adaptive setting

We assume that a feature size function f exists on K which is 1-Lipschitz. The sampling conditions become, for an (ϵ, ∞, c) uniform sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon f(x)$.
2. $\forall y \in \mathbb{X},\ d_K(y) \leq d_{\mu,m}(y) + \epsilon f(\bar{y})$.

Adaptive setting

We assume that a feature size function f exists on K which is 1-Lipschitz. The sampling conditions become, for an (ϵ, ∞, c) uniform sampling.

1. $\forall x \in K,\ d_{\mu,m}(x) \leq \epsilon f(x)$.
2. $\forall y \in \mathbb{X},\ d_K(y) \leq d_{\mu,m}(y) + \epsilon f(\bar{y})$.
3. $\forall p \in P,\ d_{\mu,m}(p) \geq \frac{\epsilon}{c} f(\bar{p})$.

Adaptive guarantees

Theorem

Given an input point P which is an (ϵ, ∞) adaptive sample of a compact K with $\epsilon \leq \frac{1}{2}$, our algorithm returns a 7ϵ Hausdorff adaptive sampling of K.

Id est:

$$\forall x \in K,\ \exists q \in Q_n,\ d_{\mathbb{X}}(x,q) \leq 7\epsilon f(x)$$

$$\forall q \in Q_n,\ \exists x \in K,\ d_{\mathbb{X}}(x,q) \leq 7\epsilon f(\bar{q})$$

Theorem

Given a set L and the feature function $f = d_L$, we consider an (ϵ, ∞, c)-uniform adaptive sample P of K. If $c \leq 2$, $\epsilon \leq \frac{1}{396}$ and $G_{\frac{1}{3}} \cap M_{\frac{\pi}{4}} = \emptyset$ then for any sufficiently small $\beta > 0$, $H_*(d_K^{-1}([0,\beta])) \cong H_*(B_{.032} \hookrightarrow B_{15.6})$.

Take home

- Aberrant noise can be handled using the distance to a measure.

Take home

- Aberrant noise can be handled using the distance to a measure.
- The distance to a measure is compatible with classical data structures.

Take home

- Aberrant noise can be handled using the distance to a measure.
- The distance to a measure is compatible with classical data structures.
- Classical data structures have sparse approximations usable in practice.

Take home

- Aberrant noise can be handled using the distance to a measure.
- The distance to a measure is compatible with classical data structures.
- Classical data structures have sparse approximations usable in practice.
- Noisy data set can be sparsified with guarantees given only one parameter.

References

- Adaptive and robust sparsification of point data,
  ongoing work, Buchet, Dey, Wang and Wang
- Collapsing Rips complexes,
  EuroCG 2013, Atalli, Lieutier and Salinas
- Efficient data structure for representing and simplifying simplicial complexes in high dimension,
  SoCG 2011, Attali, Lieuter and Salinas
- Efficient and robust persistent homology for measures,
  SoDA 2015, Buchet, Chazal, Oudot and Sheehy
- Geometric inference for probability measures,
  JFoCM 2011, Chazal, Cohen-Steiner and Mérigot
- Graph induced complex on point data,
  SoCG 2013, Dey, Fan and Wang
- Topological inference from measures,
  PhD dissertation, Buchet, 2014
- Witmessed k-distance,
  DCG 2013, Guibas, Mérigot and Morozov