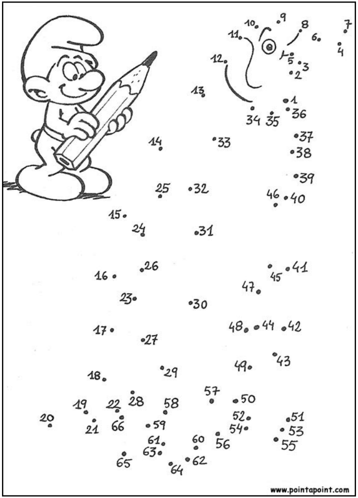
Inférence d'homologie robuste et efficace

Mickaël Buchet, Frédéric Chazal, Steve Oudot, Don Sheehy

9 avril 2013

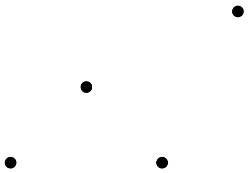
Un petit jeu..



...qui peut être compliqué

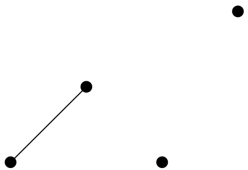
Plan

Principe de l'homologie



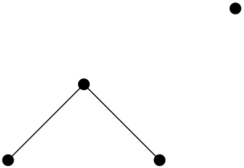
$\beta_0 = 4$ $\beta_1 = 0$

Principe de l'homologie



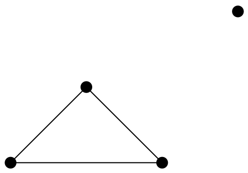
$\beta_0 = 3$ $\beta_1 = 0$

Principe de l'homologie



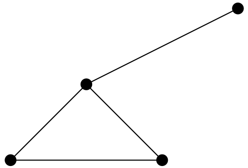
$\beta_0 = 2$ $\beta_1 = 0$

Principe de l'homologie



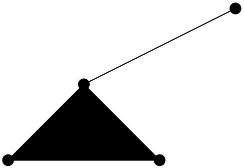
$\beta_0 = 2$ $\beta_1 = 1$

Principe de l'homologie



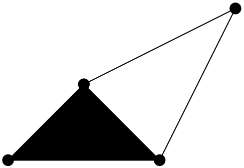
$\beta_0 = 1$ $\beta_1 = 1$

Principe de l'homologie



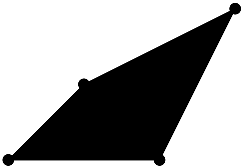
$\beta_0 = 1$ $\beta_1 = 0$

Principe de l'homologie



$\beta_0 = 1$ $\beta_1 = 1$

Principe de l'homologie

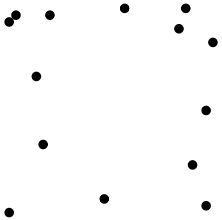


$\beta_0 = 1$ $\beta_1 = 0$

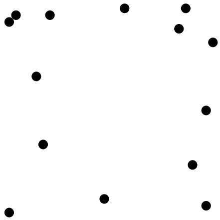
Plan

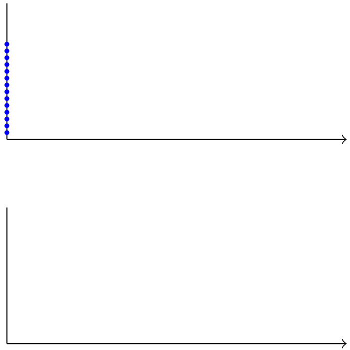
1. Introduction à l'homologie
2. Théorie de la persistance
3. Calcul du diagramme de persistance
4. Distance à la mesure
5. Rips parcimonieux

Inférence et persistance

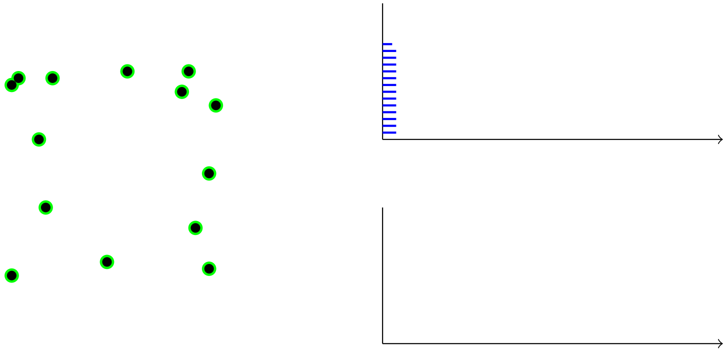


Inférence et persistance

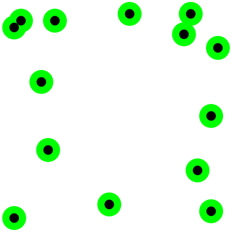


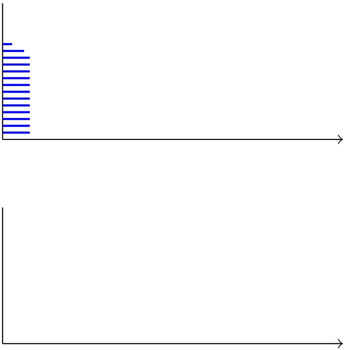


Inférence et persistance

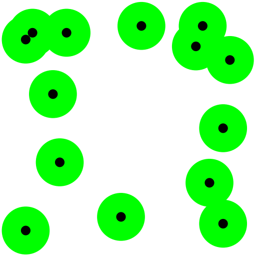


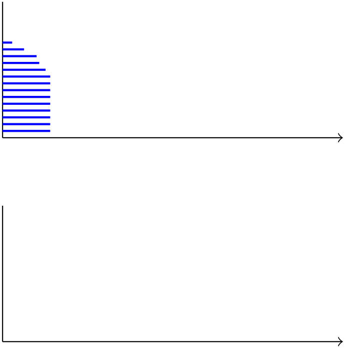
Inférence et persistance



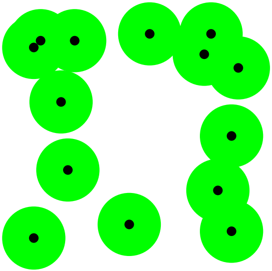


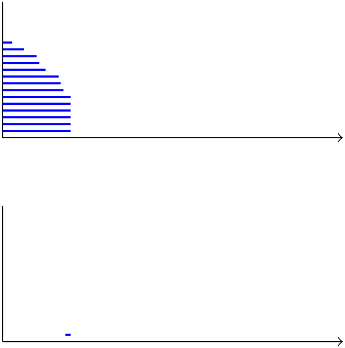
Inférence et persistance



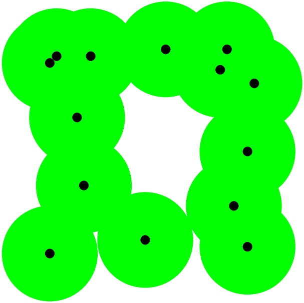


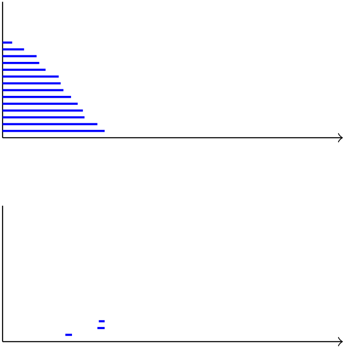
Inférence et persistance



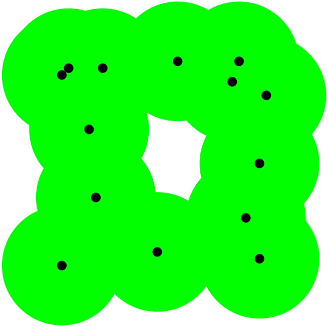


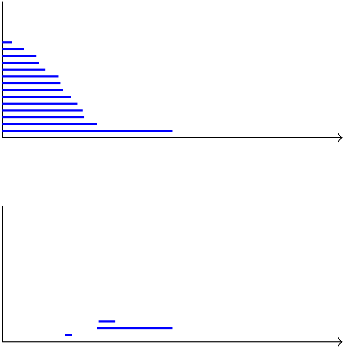
Inférence et persistance





Inférence et persistance





Plan

Complexes simpliciaux

Définition

Un simplexe de dimension l est un ensemble de l points distincts.

Complexes simpliciaux

Définition

Un simplexe de dimension l est un ensemble de l points distincts.

Définition

Un complexe simplicial est un ensemble de simplexes tel que si un simplexe appartient au complexe alors tous ses sous-ensembles appartiennent au complexe.

Le complexe de Čech

Définition

$(p_1, \ldots, p_l)$ *appartient au complexe de Čech pour le paramètre* α*, noté* C_α*, si :*

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

Le complexe de Čech

Définition

$(p_1, \ldots, p_l)$ appartient au complexe de Čech pour le paramètre α, noté C_α, si :

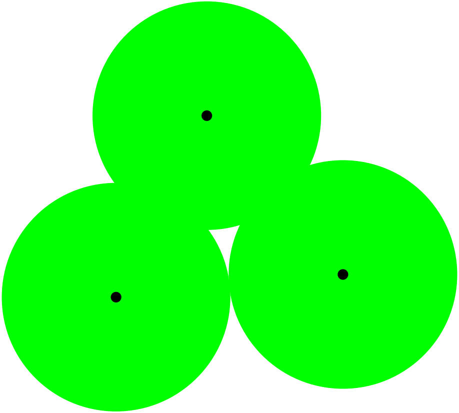
$$\cap_{i=1}^{n} B(p_i, \alpha) \neq \emptyset$$

Théorème

Le complexe de Čech a la même homologie que l'union de boules si l'espace est "gentil".

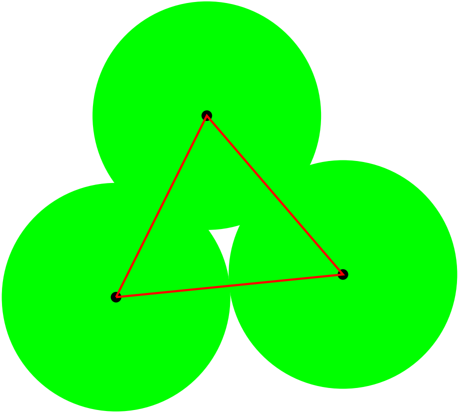
Un problème pas si simple

Quand une intersection de boules est-elle non vide ?



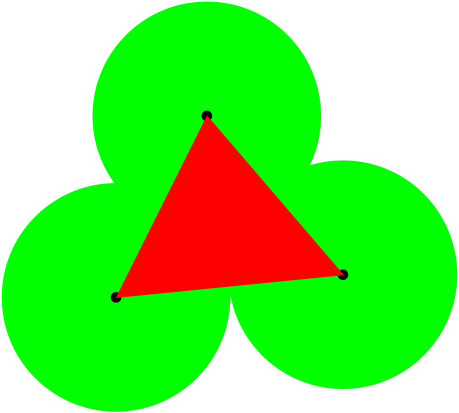
Un problème pas si simple

Quand une intersection de boules est-elle non vide ?

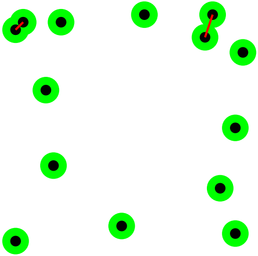


Une solution : le complexe de Rips

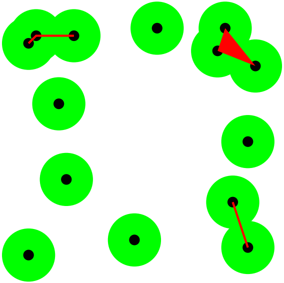
R_α contient tous les simplexes possibles avec les arêtes de C_α.



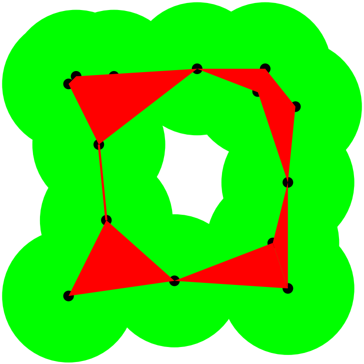
En pratique



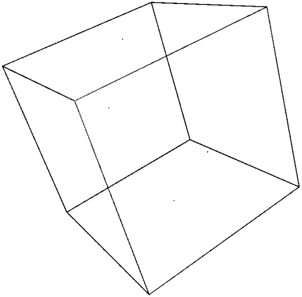
En pratique

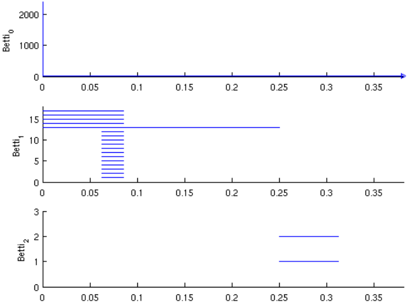


En pratique



Un problème de bruit





Plan

Distance à la mesure

- Idée : moyenner pour éviter les problèmes liés au bruit.

Distance à la mesure

- Idée : moyenner pour éviter les problèmes liés au bruit.
- Soit P un nuage de points dans un espace métrique $\mathbb{X}$.

Distance à la mesure

- Idée : moyenner pour éviter les problèmes liés au bruit.
- Soit P un nuage de points dans un espace métrique $\mathbb{X}$.
- Soit $d_i(x)$ la distance de x à son i^e-plus proche voisin dans P.

Distance à la mesure

- Idée : moyenner pour éviter les problèmes liés au bruit.
- Soit P un nuage de points dans un espace métrique $\mathbb{X}$.
- Soit $d_i(x)$ la distance de x à son i^e-plus proche voisin dans P.

$$d_{\mu_P,m}(x) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} d_i(x)^2}$$

Distance à la mesure

- Idée : moyenner pour éviter les problèmes liés au bruit.
- Soit P un nuage de points dans un espace métrique $\mathbb{X}$.
- Soit $d_i(x)$ la distance de x à son i^e-plus proche voisin dans P.

$$d_{\mu_P,m}(x) = \sqrt{\frac{1}{k}\sum_{i=1}^{k} d_i(x)^2}$$

- Introduite par Chazal, Cohen-Steiner et Mérigot en 2011

Approximation des sous-niveaux par des unions de boules

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

Approximation des sous-niveaux par des unions de boules

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

- Dans $\mathbb{R}^d$ muni de la norme euclidienne :

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

Approximation des sous-niveaux par des unions de boules

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

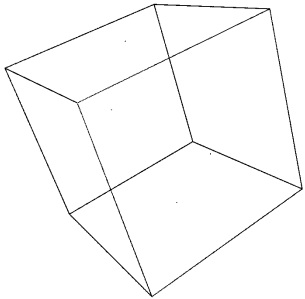
- Dans $\mathbb{R}^d$ muni de la norme euclidienne :

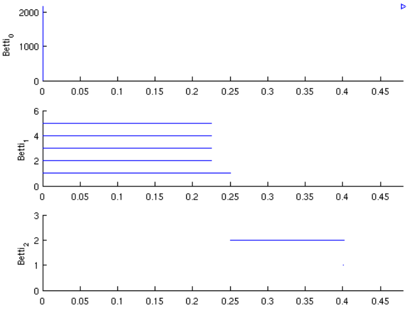
$$\frac{1}{\sqrt{2}} d_{\mu_P,m} \leq d^P_{\mu_P,m} \leq \sqrt{3}\ d_{\mu_P,m}$$

- Dans $\mathbb{X}$, espace métrique :

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

Retour au cube pointé





Un point de complexité

Soit n la taille de P,

Un point de complexité

Soit n la taille de P,

Nombre de boules :

Un point de complexité

Soit n la taille de P,

Nombre de boules : n

Un point de complexité

Soit n la taille de P,

Nombre de boules : n

Calcul de $d_{\mu_P,m}(p)$:

Un point de complexité

Soit n la taille de P,

Nombre de boules :	n
Calcul de $d_{\mu_P,m}(p)$:	$O(n^2 kd)$

Un point de complexité

Soit n la taille de P,

Nombre de boules : n

Calcul de $d_{\mu_P,m}(p)$: $O(n^2 kd)$

Simplexes considérés pour la persistance

Un point de complexité

Soit n la taille de P,

Nombre de boules :	n
Calcul de $d_{\mu_P,m}(p)$:	$O(n^2 kd)$
Simplexes considérés pour la persistance	$t = n^d$

Un point de complexité

Soit n la taille de P,

Nombre de boules :	n
Calcul de $d_{\mu_P,m}(p)$:	$O(n^2kd)$
Simplexes considérés pour la persistance	$t = n^d$
Calcul de la persistance :	

Un point de complexité

Soit n la taille de P,

Nombre de boules :	n
Calcul de $d_{\mu_P,m}(p)$:	$O(n^2kd)$
Simplexes considérés pour la persistance	$t = n^d$
Calcul de la persistance :	$O(t^3)$

Un point de complexité

Soit n la taille de P,

Nombre de boules :	n
Calcul de $d_{\mu_P,m}(p)$:	$O(n^2 kd)$
Simplexes considérés pour la persistance	$t = n^d$
Calcul de la persistance :	$O(t^3)$
Total :	$O(n^{3d})$

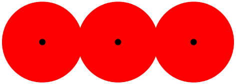
Plan

Rips parcimonieux

On peut essayer de ne pas considérer toute la filtration.

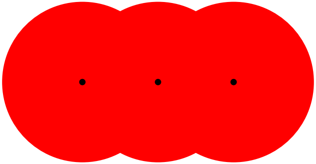
Rips parcimonieux

On peut essayer de ne pas considérer toute la filtration.



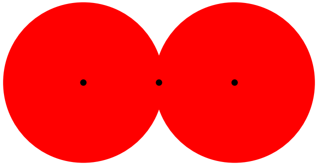
Rips parcimonieux

On peut essayer de ne pas considérer toute la filtration.



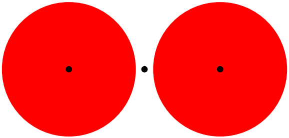
Rips parcimonieux

On peut essayer de ne pas considérer toute la filtration.



Rips parcimonieux

Mais il faut être prudent



Gain en complexité

Nouveau nombre de simplexes : $O(n \ln n)$

Gain en complexité

Nouveau nombre de simplexes : $O(n \ln n)$

Mais le O cache une constante exponentielle en la dimension intrinsèque de l'objet I.

Gain en complexité

Nouveau nombre de simplexes : $O(n \ln n)$

Mais le O cache une constante exponentielle en la dimension intrinsèque de l'objet l.

Calcul approché en utilisant la distance à la mesure :

$$O(1)^l O(n^3 \ln(n)^3)$$

Conclusion

Conclusion

On arrive à s'affranchir de caractère exponentiel en la dimension ambiante d en approximant.

Conclusion

On arrive à s'affranchir de caractère exponentiel en la dimension ambiante d en approximant.

Applications possibles en apprentissage et en analyse de données.

Conclusion

On arrive à s'affranchir de caractère exponentiel en la dimension ambiante d en approximant.

Applications possibles en apprentissage et en analyse de données.

La géométrie algorithmique est cool !