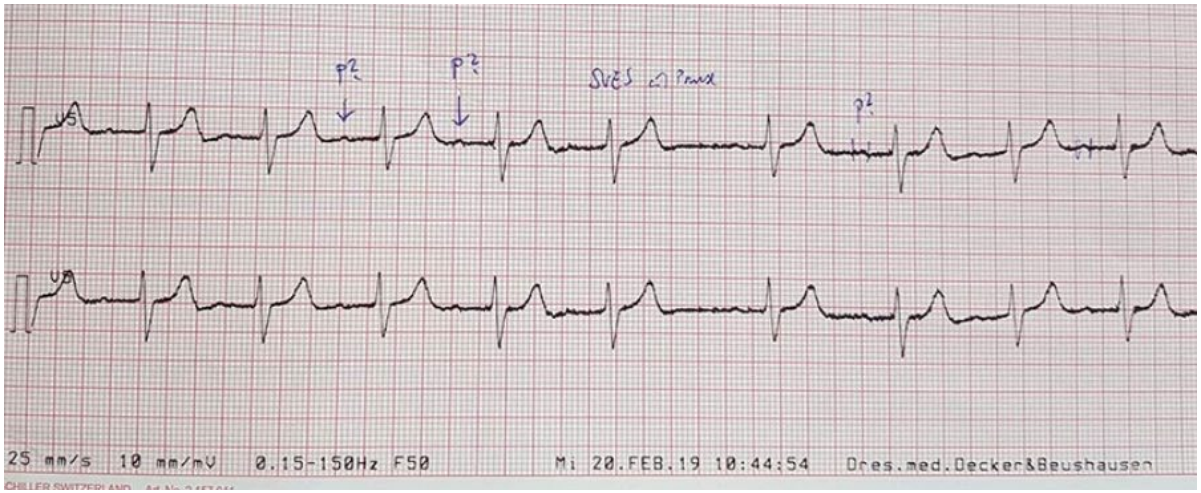
Isopérimétrie, trou spectral et analyse topologique des données

Pierre Pansu, Université Paris-Saclay

15 septembre 2023

Étudier des images de roches poreuses pour décider si la roche est perméable ou non, c'est répondre à une question de topologie.

Dans un électrocardiogramme, ce qui compte, c'est les pics (nombre, hauteur approximative), pas leur localisation temporelle exacte. C'est le relief à reparamétrage près, i.e. avant tout une information topologique.



Version multidimensionnelle : classification guidée par la densité (F. Chazal, L. Guibas, S. Oudot, P. Skrabas)

Donnée : un espace métrique fini.

Première étape. On calcule une densité locale. L'algorithme divise d'abord l'espace en amas qui sont les bassins d'attraction du gradient de la densité. Il y en trop : autant que de maxima locaux.

Deuxième étape. L'algorithme fusionne les amas qui correspondent à des maxima non suffisamment accusés.

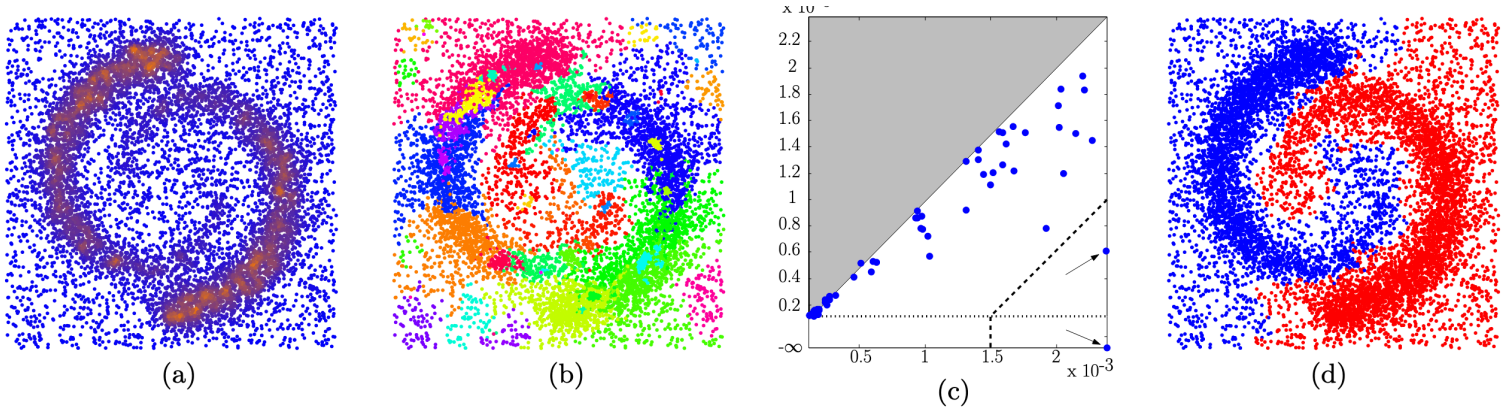


Figure 2: Our approach in a nutshell: (a) estimation of the underlying density function f at the data points; (b) result of the basic graph-based hill-climbing step; (c) approximate PD showing 2 points far off the diagonal corresponding to the 2 prominent peaks of f; (d) final result obtained after merging the clusters of non-prominent peaks.

La méthode a donné d'excellents résultats sur des données de cytologie (classification de cellules à partir de photos).

Plus généralement, dès que les données sont plus que de simples images (i.e. des fonctions sur un espace), ou des matrices de même taille, un descripteur topologique a l'avantage d'être indépendant de choix arbitraires.

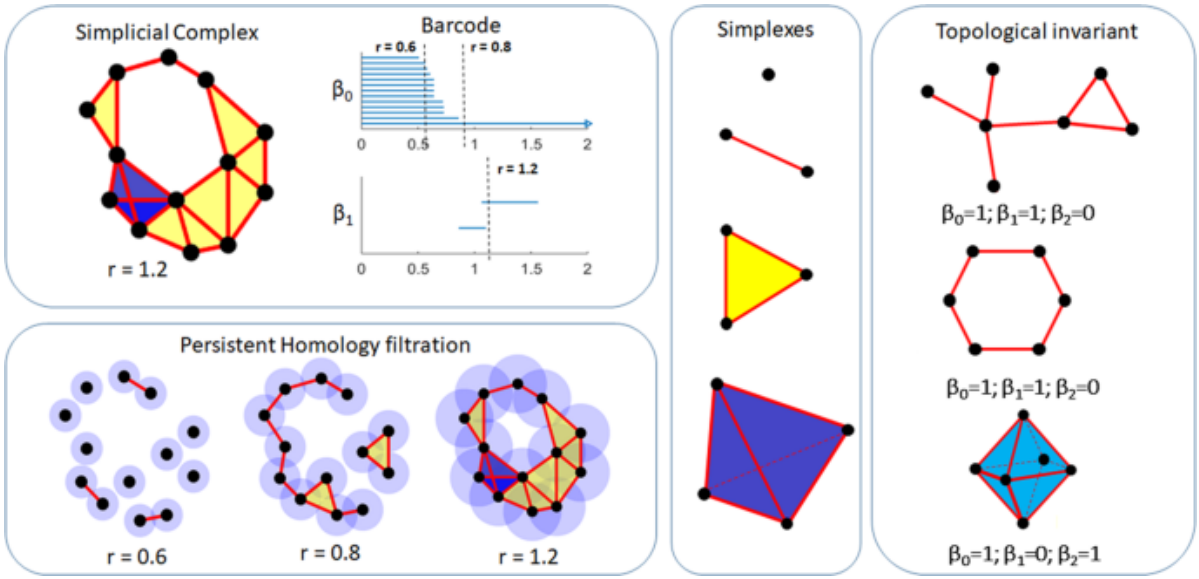
Plus généralement, dès que les données sont plus que de simples images (i.e. des fonctions sur un espace), ou des matrices de même taille, un descripteur topologique a l'avantage d'être indépendant de choix arbitraires.

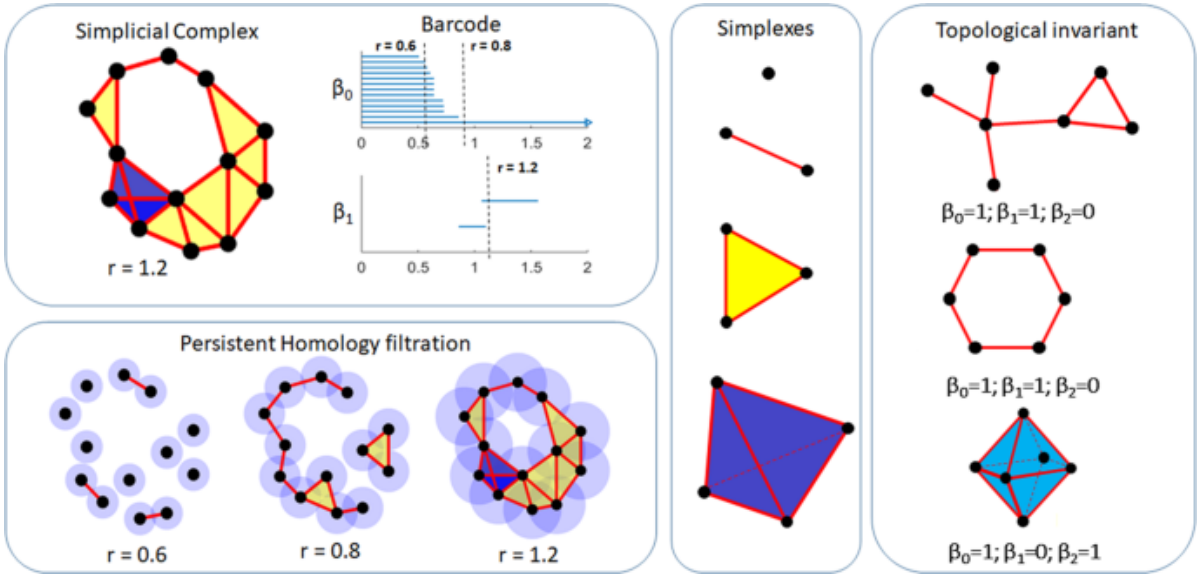
Souvent, les données se présentent comme un nuage fini de points dans $\mathbb{R}^N$.
Comment parler de la topologie d'un nuage ?

Plus généralement, dès que les données sont plus que de simples images (i.e. des fonctions sur un espace), ou des matrices de même taille, un descripteur topologique a l'avantage d'être indépendant de choix arbitraires.

Souvent, les données se présentent comme un nuage fini de points dans $\mathbb{R}^N$.
Comment parler de la topologie d'un nuage ?

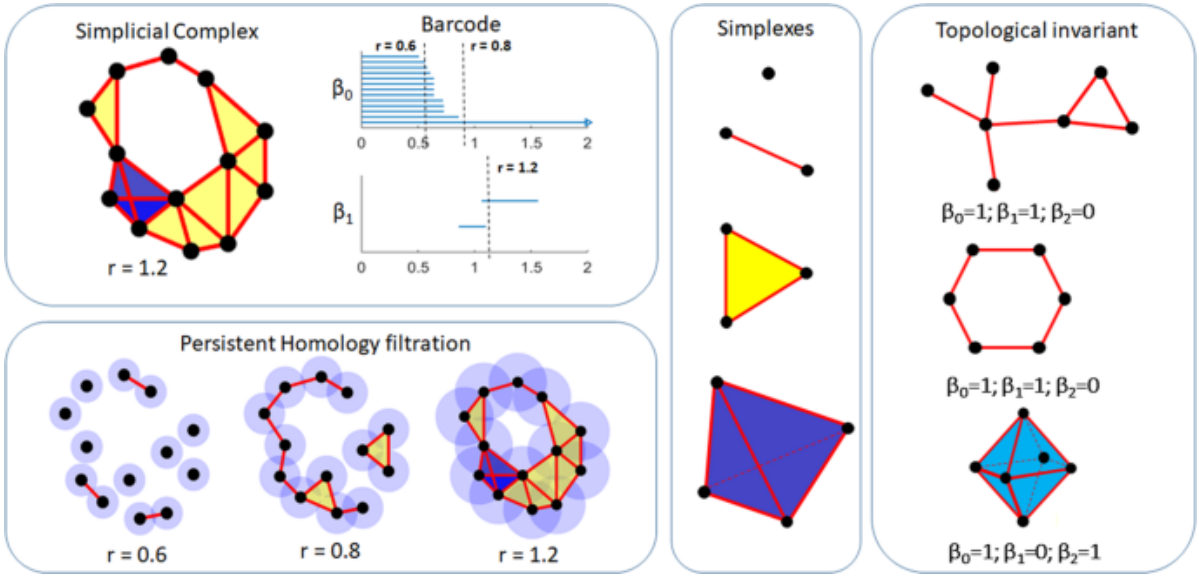
On lui associe une famille croissante de polyèdres simpliciaux : N_r est le nerf du recouvrement par des boules de rayon r.





Il n'y a pas a priori de choix naturel pour r.

On calcule l'homologie à coefficients dans $\mathbb{F}_2$, $H_k(N_r, \mathbb{F}_2)$, pour chaque r.

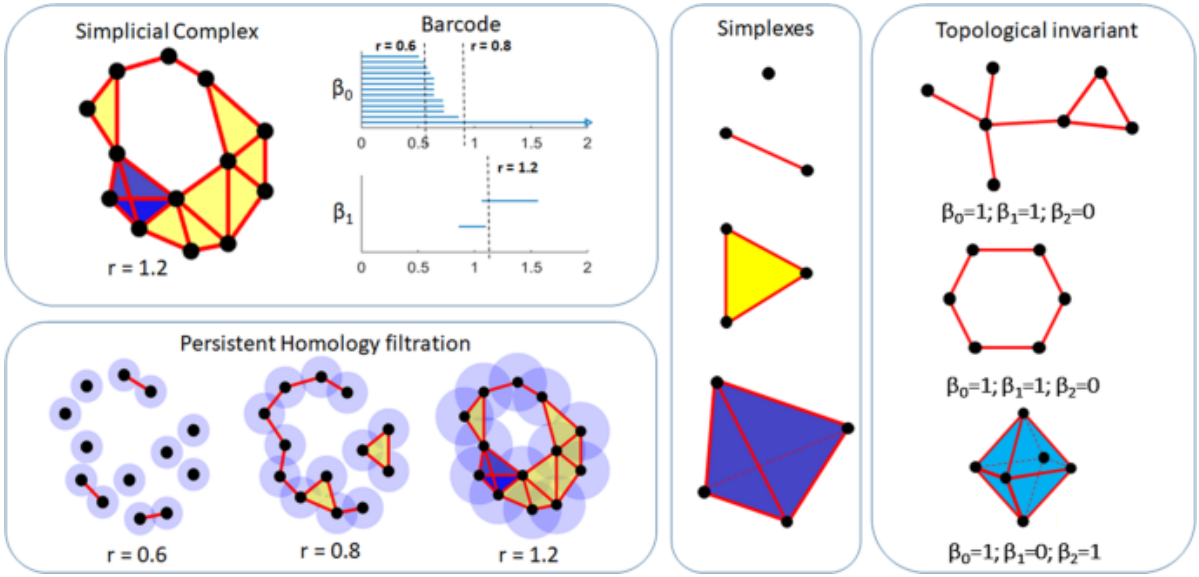


Il n'y a pas a priori de choix naturel pour r.

On calcule l'homologie à coefficients dans $\mathbb{F}_2$, $H_k(N_r, \mathbb{F}_2)$, pour chaque r.

Inconvénient. $H_k(N_r, \mathbb{F}_2)$ est très instable.

Comment en extraire une information stable ?



Il n'y a pas a priori de choix naturel pour r.

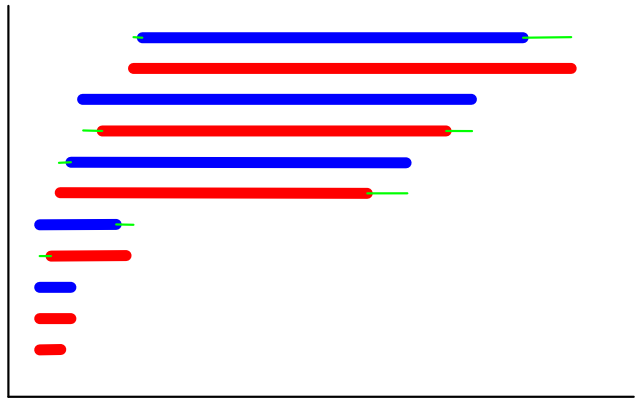
On calcule l'homologie à coefficients dans $\mathbb{F}_2$, $H_k(N_r, \mathbb{F}_2)$, pour chaque r.

On sait détecter quand une classe d'homologie naît (elle n'appartient pas à $im(H_k(N_{r-\epsilon}, \mathbb{F}_2) \to H_k(N_r, \mathbb{F}_2))\ \forall \epsilon$) et quand elle meurt, d'où une collection d'intervalles $[r, s]$, le *code-barre*.

Idée : les grandes barres sont robustes (stables par perturbation du nuage en distance de Hausdorff), les petites sont du bruit.

Idée : les grandes barres sont robustes (stables par perturbation du nuage en distance de Hausdorff), les petites sont du bruit.

La *distance bottleneck* entre deux codes-barres s'obtient en appariant les barres au mieux, quitte à jeter les barres trop courtes.



Théorème (de stabilité : Cohen-Steiner, Edelsbrunner, Harer 2005)

dist. bottleneck(codes-barres) $\leq$ *dist. Hausdorff (nuages).*

Le théorème de stabilité invite à négliger les petites barres. Que signifie ce qui reste ?

Théorème (Signification des grandes barres : Chazal, Lieutier 2005)

Si un ensemble $X \subset \mathbb{R}^N$ a un Weak feature size 4ϵ, et si

$$\textit{dist. Hausdorff (nuage,}X\textit{)} < \epsilon,$$

alors l'homologie des grandes barres coincide avec l'homologie de X.

Le théorème de stabilité invite à négliger les petites barres. Que signifie ce qui reste ?

Théorème (Signification des grandes barres : Chazal, Lieutier 2005)

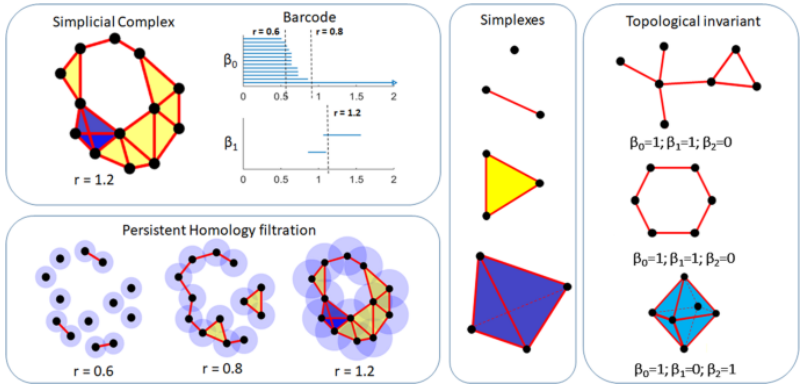
Si un ensemble $X \subset \mathbb{R}^N$ a un Weak feature size 4ϵ, et si

$$\textit{dist. Hausdorff (nuage,}X\textit{)} < \epsilon,$$

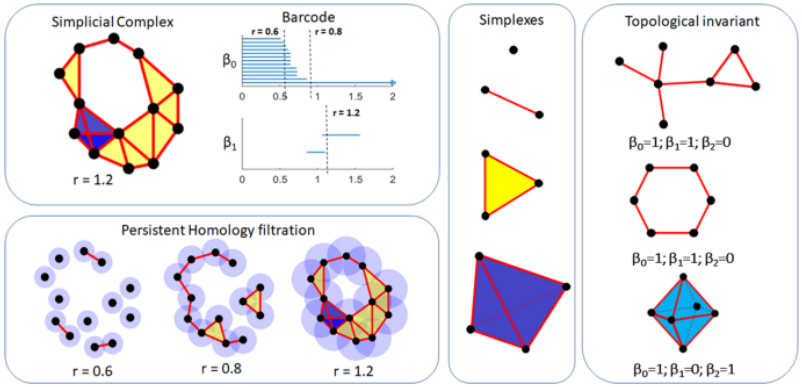
alors l'homologie des grandes barres coincide avec l'homologie de X.

Moralité : étant donné un objet un peu régulier, on peut calculer sa topologie (au moins, son homologie) à partir d'un échantillon assez dense, avec des garanties théoriques.

L'homologie des complexes simpliciaux, c'est de l'algèbre linéaire. On considère l'espace vectoriel (sur $\mathbb{F}_2$ ou $\mathbb{R}$) dont une base est l'ensemble des simplexes (sommets, arêtes, faces,...).
Ici, de dimension 48 (13 sommets, 23 arêtes, 11 faces, 1 tétraèdre).



L'homologie des complexes simpliciaux, c'est de l'algèbre linéaire. On considère l'espace vectoriel (sur $\mathbb{F}_2$ ou $\mathbb{R}$) dont une base est l'ensemble des simplexes (sommets, arêtes, faces,...).
Ici, de dimension 48 (13 sommets, 23 arêtes, 11 faces, 1 tétraèdre).

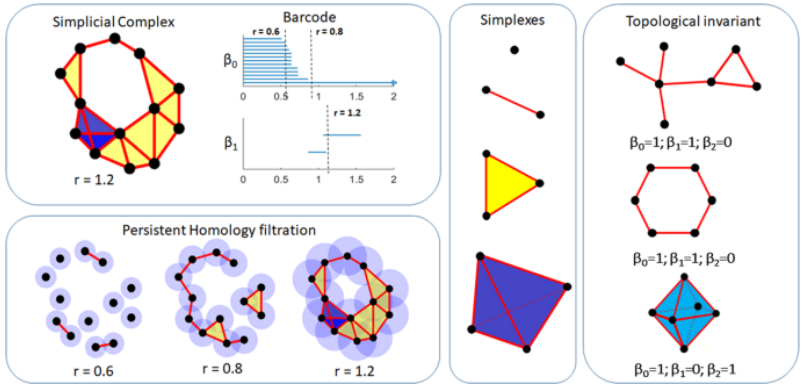


Le bord d'un simplexe est une combinaison linéaire d'autres simplexes, d'où une application linéaire ∂, et son adjoint d. Le bord du bord vaut zéro : $\partial \circ \partial = 0$, d'où $d \circ d = 0$. En particulier, $\mathrm{Im}(\partial) \subset \mathrm{Ker}(\partial)$, $\mathrm{Im}(d) \subset \mathrm{Ker}(d)$. On définit

$$\text{homologie } = \mathrm{Ker}(\partial)/\mathrm{Im}(\partial), \quad \text{cohomologie } = \mathrm{Ker}(d)/\mathrm{Im}(d).$$

Les deux comptent les trous, i.e. les cycles qui ne sont pas des bords.

L'homologie des complexes simpliciaux, c'est de l'algèbre linéaire. On considère l'espace vectoriel (sur $\mathbb{F}_2$ ou $\mathbb{R}$) dont une base est l'ensemble des simplexes (sommets, arêtes, faces,...).
Ici, de dimension 48 (13 sommets, 23 arêtes, 11 faces, 1 tétraèdre).



Le bord d'un simplexe est une combinaison linéaire d'autres simplexes, d'où une application linéaire ∂, et son adjoint d. Le bord du bord vaut zéro : $\partial \circ \partial = 0$, d'où $d \circ d = 0$. En particulier, $\mathrm{Im}(\partial) \subset \mathrm{Ker}(\partial)$, $\mathrm{Im}(d) \subset \mathrm{Ker}(d)$. On définit

$$\text{homologie } = \mathrm{Ker}(\partial)/\mathrm{Im}(\partial), \quad \text{cohomologie } = \mathrm{Ker}(d)/\mathrm{Im}(d).$$

Les deux comptent les trous, i.e. les cycles qui ne sont pas des bords.

Plus généralement, un *complexe de chaines* est un espace vectoriel B muni d'une application linéaire d telle que $d \circ d = 0$.

Question. Quelles garanties de robustesse pour le calcul de l'homologie d'un complexe de chaînes ?

On a vu comment l'homologie persistante apportait de la robustesse. En analyse numérique, la robustesse est obtenue par un autre moyen.

On a vu comment l'homologie persistante apportait de la robustesse. En analyse numérique, la robustesse est obtenue par un autre moyen.

Si $F : B \to B'$ est un opérateur inversible entre espaces normés, son *conditionnement* est

$$\kappa(F) = |F||F^{-1}|.$$

On a vu comment l'homologie persistante apportait de la robustesse. En analyse numérique, la robustesse est obtenue par un autre moyen.

Si $F : B \to B'$ est un opérateur inversible entre espaces normés, son *conditionnement* est

$$\kappa(F) = |F||F^{-1}|.$$

Le conditionnement assure la stabilité de la résolution de l'équation $F(x) = b$.

On a vu comment l'homologie persistante apportait de la robustesse. En analyse numérique, la robustesse est obtenue par un autre moyen.

Si $F : B \to B'$ est un opérateur inversible entre espaces normés, son *conditionnement* est

$$\kappa(F) = |F||F^{-1}|.$$

Le conditionnement assure la stabilité de la résolution de l'équation $F(x) = b$.

Si on change b en $b' = b + \Delta b$, la solution change en $x' = x + \Delta x$, telle que

$$\frac{|\Delta x|}{|x|} \leq \kappa(F) \frac{|\Delta b|}{|b|}.$$

Si on change F en $F' = F + \Delta F$, la solution change en $x' = x + \Delta x$, telle que

$$\frac{|\Delta x|}{|x'|} \leq \kappa(F) \frac{|\Delta F|}{|F|}.$$

On a vu comment l'homologie persistante apportait de la robustesse. En analyse numérique, la robustesse est obtenue par un autre moyen.

Si $F : B \to B'$ est un opérateur inversible entre espaces normés, son *conditionnement* est

$$\kappa(F) = |F||F^{-1}|.$$

Le conditionnement assure la stabilité de la résolution de l'équation $F(x) = b$.

Si on change b en $b' = b + \Delta b$, la solution change en $x' = x + \Delta x$, telle que

$$\frac{|\Delta x|}{|x|} \leq \kappa(F) \frac{|\Delta b|}{|b|}.$$

Si on change F en $F' = F + \Delta F$, la solution change en $x' = x + \Delta x$, telle que

$$\frac{|\Delta x|}{|x'|} \leq \kappa(F) \frac{|\Delta F|}{|F|}.$$

Lorsque $d : B \to B$ est un complexe de chaînes (i.e. $d \circ d = 0$), on s'intéresse au conditionnement de $\bar{d} : B/\mathrm{Ker}(d) \to \mathrm{Im}(d)$. En dimension infinie, il peut être infini.

On munit les chaînes (cochaînes) d'un complexe simplicial des normes ℓ^p.
Exemple. Le graphe rectiligne de longueur n satisfait $H^1 = 0$. La 1-cochaîne g égale à $\bar{1}$ sur l'arête centrale et 0 ailleurs s'écrit df où

$$\|g\|_p = 1, \quad \|f\|_p \sim n^{1/p}.$$

On munit les chaînes (cochaînes) d'un complexe simplicial des normes ℓ^p.
Exemple. Le graphe rectiligne de longueur n satisfait $H^1 = 0$. La 1-cochaîne g égale à $\bar{1}$ sur l'arête centrale et 0 ailleurs s'écrit df où

$$\|g\|_p = 1, \quad \|f\|_p \sim n^{1/p}.$$

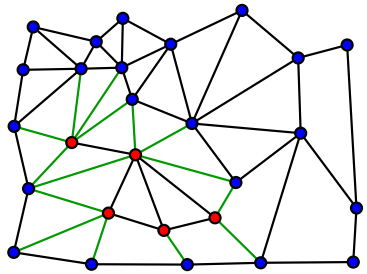
Quand n est grand, résoudre $df = g$ est instable. Le calcul de l'homologie est mal conditionné.

Définition

Le **conditionnement** *d'un graphe X est $\kappa(X, p, \mathbf{k}) = |\bar{d}||\bar{d}^{-1}|$ où $\bar{d} : C(X, \mathbf{k})/\mathrm{Ker}(d) \to dC(X, \mathbf{k})$. (Il dépend de p et du corps $\mathbf{k}$).*

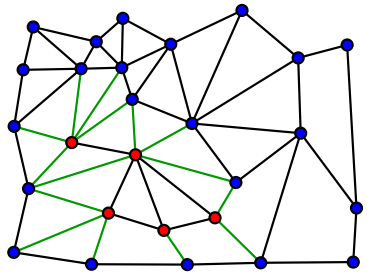
Isopérimétrie = art de couper l'espace en deux.

$|A| = 5$, $|\partial A| = 15$.



Isopérimétrie = art de couper l'espace en deux.

$|A| = 5$, $|\partial A| = 15$.



Définition

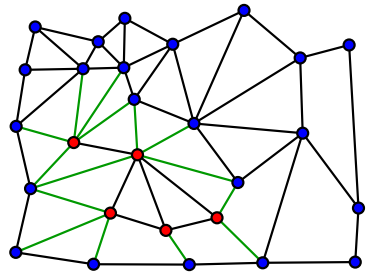
La **constante de Cheeger** $h(X)$ *d'un graphe* X *est le plus grand* h *tel que pour tout ensemble* A *de sommets tel que* $|A| \leq \frac{1}{2}|X|$,

$$|\partial A| \geq h\,|A|.$$

Ici, ∂A *est l'ensemble des arêtes reliant* A *à son complémentaire.*

Isopérimétrie = art de couper l'espace en deux.

$|A| = 5$, $|\partial A| = 15$.



Définition

La **constante de Cheeger** $h(X)$ *d'un graphe* X *est le plus grand* h *tel que pour tout ensemble* A *de sommets tel que* $|A| \leq \frac{1}{2}|X|$,

$$|\partial A| \geq h\,|A|.$$

Ici, ∂A *est l'ensemble des arêtes reliant* A *à son complémentaire.*

Proposition

$$h(X) = \frac{2}{\kappa(X, 1, \mathbb{F}_2)} = 2(\|\bar{d}\|_{1\to 1}\|\bar{d}^{-1}\|_{1\to 1})^{-1} \text{ sur } \mathbb{F}_2.$$

Proposition

Soit Δ l'opérateur auto-adjoint correspondant à la forme quadratique $f \mapsto \|df\|_2^2 = \langle f, \Delta f \rangle$. Soient $\lambda_1 \leq \lambda_2 \leq \cdots$ ses valeurs propres. Si X est connexe, alors $\lambda_1 = 0$ et

$$\lambda_2 = (\|\bar{d}^{-1}\|_{2\to 2})^{-2}.$$

Proposition

Soit Δ l'opérateur auto-adjoint correspondant à la forme quadratique $f \mapsto \|df\|_2^2 = \langle f, \Delta f \rangle$. Soient $\lambda_1 \leq \lambda_2 \leq \cdots$ ses valeurs propres. Si X est connexe, alors $\lambda_1 = 0$ et

$$\lambda_2 = (\|\bar{d}^{-1}\|_{2\to 2})^{-2}.$$

En particulier,

$$2\lambda_2^{-1} \leq \kappa_0(X, 2, \mathbb{R})^2 \leq 4\lambda_2^{-1}.$$

Proposition

Soit Δ l'opérateur auto-adjoint correspondant à la forme quadratique $f \mapsto \|df\|_2^2 = \langle f, \Delta f \rangle$. Soient $\lambda_1 \leq \lambda_2 \leq \cdots$ ses valeurs propres. Si X est connexe, alors $\lambda_1 = 0$ et

$$\lambda_2 = (\|\bar{d}^{-1}\|_{2\to 2})^{-2}.$$

En particulier,

$$2\lambda_2^{-1} \leq \kappa_0(X, 2, \mathbb{R})^2 \leq 4\lambda_2^{-1}.$$

On appelle λ_2 le *trou spectral* du graphe.
Il contrôle la vitesse à laquelle une marche au hasard sur le graphe est mélangeante.
En particulier, la facilité de tirer un sommet au hasard.

Proposition

Soit Δ l'opérateur auto-adjoint correspondant à la forme quadratique $f \mapsto \|df\|_2^2 = \langle f, \Delta f \rangle$. Soient $\lambda_1 \leq \lambda_2 \leq \cdots$ ses valeurs propres. Si X est connexe, alors $\lambda_1 = 0$ et

$$\lambda_2 = (\|\bar{d}^{-1}\|_{2\to 2})^{-2}.$$

En particulier,

$$2\lambda_2^{-1} \leq \kappa_0(X, 2, \mathbb{R})^2 \leq 4\lambda_2^{-1}.$$

On appelle λ_2 le *trou spectral* du graphe.
Il contrôle la vitesse à laquelle une marche au hasard sur le graphe est mélangeante.
En particulier, la facilité de tirer un sommet au hasard.

Moralité. Le complexe de chaînes normé contient des informations intéressantes, au-delà de sa seule homologie.

Question. Peut-on interpréter le conditionnement $\kappa = |\bar{d}||\bar{d}^{-1}|$ comme la distance entre certains complexes de cochaînes normés ?

Question. Peut-on interpréter le conditionnement $\kappa = |\bar{d}||\bar{d}^{-1}|$ comme la distance entre certains complexes de cochaînes normés ?

Pour définir une distance entre complexes de chaines normés, la première idée est de mesurer les conditionnements des isomorphismes.

Définition

Soient $B_1 \xrightarrow{d_1} B_1$ *et* $B_2 \xrightarrow{d_2} B_2$ *deux complexes de chaînes normés. La* **distance de Banach-Mazur** $\mathrm{BMDist}(B_1, B_2)$ *est l'inf des* $\log(|F||F^{-1}|)$ *sur tous les isomorphismes* $F : B_1 \to B_2$ *tels que* $Fd_1 = d_2F$.

Question. Peut-on interpréter le conditionnement $\kappa = |\bar{d}||\bar{d}^{-1}|$ comme la distance entre certains complexes de cochaînes normés ?

Pour définir une distance entre complexes de chaines normés, la première idée est de mesurer les conditionnements des isomorphismes.

Définition

Soient $B_1 \xrightarrow{d_1} B_1$ *et* $B_2 \xrightarrow{d_2} B_2$ *deux complexes de chaînes normés. La* **distance de Banach-Mazur** $\mathrm{BMDist}(B_1, B_2)$ *est l'inf des* $\log(|F||F^{-1}|)$ *sur tous les isomorphismes* $F : B_1 \to B_2$ *tels que* $Fd_1 = d_2F$.

C'est trop restrictif : cela impose que $\dim(B_1) = \dim(B_2)$.

La seconde idée est de mesurer la taille des homotopies.

Définition

Soient $B_1 \xrightarrow{d_1} B_1$ *et* $B_2 \xrightarrow{d_2} B_2$ *deux complexes de chaînes normés. On considère toutes les homotopies bornées, i.e.*

- *des morphismes bornés* $F_1 : B_1 \to B_2$ *et* $F_2 : B_2 \to B_1$ *tels que*

$$d_2 F_1 = F_1 d_1, \quad d_1 F_2 = F_2 d_2,$$

- *des opérateurs bornés* $Q_1 : B_1 \to B_1$ *and* $Q_2 : B_2 \to B_2$ *tels que*

$$1 - F_2 F_1 = d_1 Q_1 + Q_1 d_1, \quad 1 - F_1 F_2 = d_2 Q_2 + Q_2 d_2.$$

On note $q = \max\{|Q_1|, |Q_2|\}$, $f = \max\{1, |F_1||F_2|\}$. *La* **distance homotopique** $\mathrm{HomDist}(B_1, B_2)$ *est l'inf sur toutes les homotopies de* $\min\{\frac{q}{f} + \log f, \frac{f}{q} + \log q\}$.

La seconde idée est de mesurer la taille des homotopies.

Définition

Soient $B_1 \xrightarrow{d_1} B_1$ *et* $B_2 \xrightarrow{d_2} B_2$ *deux complexes de chaînes normés. On considère toutes les homotopies bornées, i.e.*

- *des morphismes bornés* $F_1 : B_1 \to B_2$ *et* $F_2 : B_2 \to B_1$ *tels que*

$$d_2 F_1 = F_1 d_1, \quad d_1 F_2 = F_2 d_2,$$

- *des opérateurs bornés* $Q_1 : B_1 \to B_1$ *and* $Q_2 : B_2 \to B_2$ *tels que*

$$1 - F_2 F_1 = d_1 Q_1 + Q_1 d_1, \quad 1 - F_1 F_2 = d_2 Q_2 + Q_2 d_2.$$

On note $q = \max\{|Q_1|, |Q_2|\}$, $f = \max\{1, |F_1||F_2|\}$. *La* **distance homotopique** $\mathrm{HomDist}(B_1, B_2)$ *est l'inf sur toutes les homotopies de* $\min\{\frac{q}{f} + \log f, \frac{f}{q} + \log q\}$.

L'expression bizarre est là pour garantir l'inégalité triangulaire. ▸ Détails

Analogie entre complexes de chaînes normés et espaces métriques.

Espace métrique	Complexe de chaînes normé
Borné	Homotope à un complexe de dimension finie
Précompact	Dans l'adhérence des complexes de dim. finie
Distance de Hausdorff-Gromov	Distance homotopique
Critère de compacité (Gromov)	?

Analogie entre complexes de chaînes normés et espaces métriques.

Espace métrique	Complexe de chaînes normé
Borné	Homotope à un complexe de dimension finie
Précompact	Dans l'adhérence des complexes de dim. finie
Distance de Hausdorff-Gromov	Distance homotopique
Critère de compacité (Gromov)	?

Définition

Soit (B, d) un complexe de chaînes normé. Son **profil** *est la plus petite fonction $\pi = (\pi_d, \pi_c) : (0, +\infty) \to (0, +\infty)^2$ ayant la propriété suivante. Pour tout $\epsilon > 0$, il existe un complexe de dimension finie (B', d') tel que*

$$HomDist(B, B') < \epsilon, \quad \dim(B') \leq \pi_d(\epsilon), \quad \kappa(B', d') \leq \pi_c(\epsilon).$$

Analogie entre complexes de chaînes normés et espaces métriques.

Espace métrique	Complexe de chaînes normé
Borné	Homotope à un complexe de dimension finie
Précompact	Dans l'adhérence des complexes de dim. finie
Distance de Hausdorff-Gromov	Distance homotopique
Critère de compacité (Gromov)	?

Définition

Soit (B, d) un complexe de chaînes normé. Son **profil** *est la plus petite fonction $\pi = (\pi_d, \pi_c) : (0, +\infty) \to (0, +\infty)^2$ ayant la propriété suivante. Pour tout $\epsilon > 0$, il existe un complexe de dimension finie (B', d') tel que*

$$HomDist(B, B') < \epsilon, \quad \dim(B') \leq \pi_d(\epsilon), \quad \kappa(B', d') \leq \pi_c(\epsilon).$$

Théorème

Une famille de complexes de chaînes normés est précompacte si et seulement si les profils sont uniformément majorés.

Exemple

Soit M une variété riemannienne compacte lisse. Soit E le complexe des formes différentielles lisses sur M muni de la norme L^2. Alors le profil π de (E, d) est majoré par l'asymptotique des valeurs propres du laplacien de Hodge,

$$\pi_d(\epsilon) \leq \mathrm{Card}\{\lambda \in \mathrm{spectre}(\Delta)\,;\, \lambda < \frac{1}{\epsilon^2}\}, \quad \pi_c(\epsilon) \leq \frac{1}{\epsilon\sqrt{\lambda_2}}.$$

Cela donne une estimée du type $\pi_d(\epsilon) \leq C\,\epsilon^{-N}$, où $N = \dim(M)$.

Exemple

Soit M une variété riemannienne compacte lisse. Soit E le complexe des formes différentielles lisses sur M muni de la norme L^2. Alors le profil π de (E, d) est majoré par l'asymptotique des valeurs propres du laplacien de Hodge,

$$\pi_d(\epsilon) \leq \mathrm{Card}\{\lambda \in \mathrm{spectre}(\Delta)\,;\, \lambda < \frac{1}{\epsilon^2}\}, \quad \pi_c(\epsilon) \leq \frac{1}{\epsilon\sqrt{\lambda_2}}.$$

Cela donne une estimée du type $\pi_d(\epsilon) \leq C\,\epsilon^{-N}$, où $N = \dim(M)$.

Conjecture

Pour des nuages de points répartis sur une variété compacte lisse, les complexes de cochaînes ℓ^2 forment une famille précompacte.

Conclusion

- Certains types de données se prêtent bien à une analyse topologique.
- Déterminer l'homologie d'un objet à partir d'un échantillon pose un problème de robustesse, bien résolu par l'homologie persistante et ses codes-barres.
- L'analyse numérique suggère une autre approche : le conditionnement.
- Le conditionnement va plus loin que l'homologie, c'est une propriété d'un complexe de chaînes normé, intéressante dans plusieurs situations.
- On espère associer des complexes de chaînes normés à des objets assez irréguliers, la robustesse s'exprimant par de la précompacité.