Diversité de communauté de reads et structure de nuages de points

Alain Franc & *al.*

INRA BioGeCo & INRIA Equipe Pleiade

Toulouse, Séminaire MIAT
le 19 février 2016

Points abordés ...

1. Contexte : concepts et questions biologiques
2. Remarques sur les distances
3. Communautés sur graphes
4. Nuage de points : question de géométrie
5. Metric embedding
6. Topological Data Analysis

Contexte général

1. Contexte de biodiversité : en quoi les objets biologiques sont-ils différents ?
2. Quels sont les patterns et motifs, locaux et globaux ?
3. Domaine de l'évolution et l'écologie des communautés
4. Génomes comme empreinte moléculaire de l'évolution
5. Pas de processus fonctionnels
6. Travaux sur les arbres de la forêts guyanaise, et les communautés de diatomées

Données et questions

1. Taxonomie moléculaire
2. Chaque individu a un *attribut* : une séquence (un mot de 500 lettres $\longleftarrow$ alphabet de 4 lettres $w \in \{A, T, C, G\}^n$)
3. Histoire inférée par des phylogénies moléculaire (Tree of life)
4. Modèles statistiques atteignent leurs limites si $\approx 10^3, 10^4$ individus (ML, bayésien)
5. Travail sur des distances
6. Distance d'édition (Levenstein, 1965)

Tout commence par une distance ...

- On se donne un ensemble de reads $i = 1, \ldots, n$, ou $i \in \{1, n\}$
- et on calcule les distances 2 à 2
- selon le score de l'alignement local (Smith-Waterman)

$$(i,j) \longrightarrow d(i,j) := d_{ij} = \frac{\min\{\ell(i),\ell(j)\} - SW(i,j)}{2}$$

Remarques :

1. la distance de "Smith-Waterman" n'est pas une distance
2. la distance d'édition (Levenstein) est une distance
3. mais ... compromis pour le traitement des bords ...

Distance génétique et distance évolutive

$$d_{\text{evo}}(s, s') = t = \text{âge de l'ancêtre commun le plus proche}$$

C'est une distance ultramétrique

$$d(s, s') \leq \max_{s''}\{d(s, s''), d(s', s'')\}$$

Observation : Elle ne se mesure pas, mais se calcule par un modèle statistique (maximum de vraisemblance, approche bayésienne)

Remarque

Mais quid des organismes diploides ? Un ancêtre commun par gène ?
$\Longrightarrow$ *Un arbre et une distance par gène et non par organsimes chez les diploides ...*

Taxonomie moléculaire et cliques

Une observation

$$d \quad \text{ultramétrique} \quad \Longrightarrow \quad \forall\, \theta, \quad x \sim y \quad \Leftrightarrow d(x,y) \leq \theta \quad \text{transitive}$$

$$d(x,y) \leq \theta,\ d(y,z) \leq \theta \quad \Rightarrow \quad d(x,z) \leq \max\{d(x,y), d(y,z)\} \leq \theta.$$

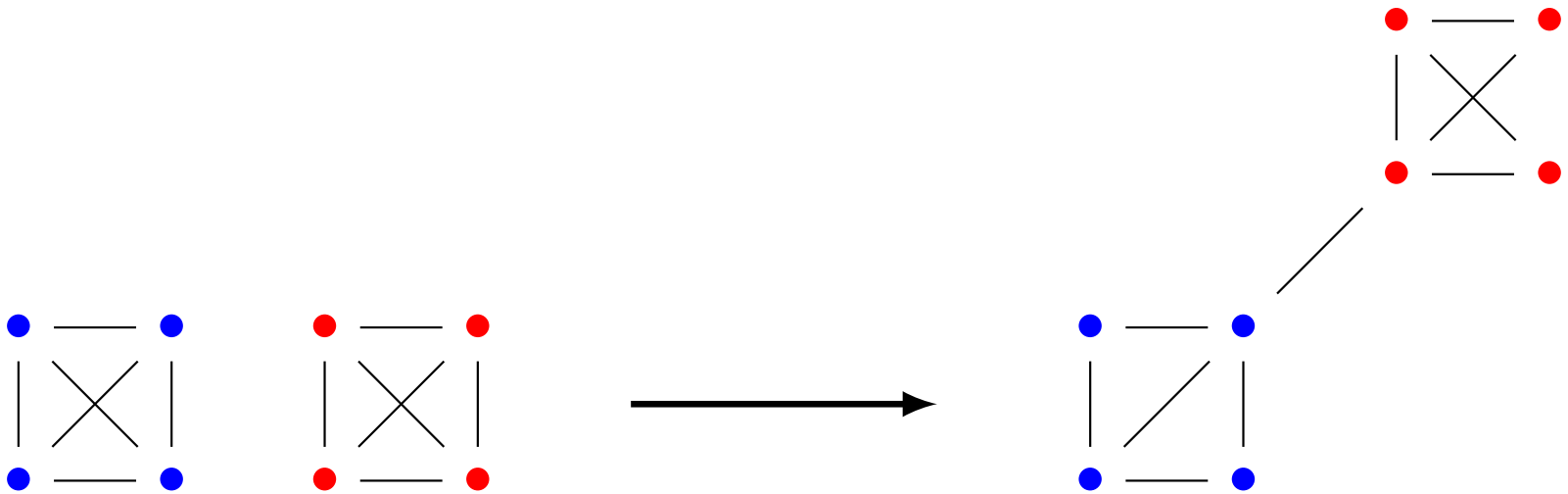
- On en déduit qu'il est possible de partitionner les séquences en classes d'équivalences

- Chaque classe d'équivalence correspond à un taxon au grain θ

Une conséquence

Si $G_\theta = (V, E_\theta)$ avec $x \sim y \Leftrightarrow d(x,y) \leq \theta$, alors G_θ est une réunion de cliques maximales sans intersections.

De graphes de taxons à des communautés sur graphes ...



Quelques notations

- On se donne un graphe $G = (V, E)$
- On définit :

$$\begin{aligned} x \ : \ V &\longrightarrow \Lambda \\ i &\longrightarrow x_i \end{aligned}$$

- On en déduit une partition des nœuds

$$V = \bigsqcup V_\alpha, \qquad V_\alpha = x^{-1}(\alpha) = \{i \in V \ : \ x_i = \alpha\}$$

Qualité d'une partition

- On se donne une partition des nœuds de $G = (V, E)$

$$V = \bigsqcup_{\alpha} V_\alpha$$

- V_α est une communauté

- On se donne pour objectif
  - valoriser les liens au sein d'une communauté
  - pénaliser les liens absents au sein d'une communautés
  - pénaliser les liens entre communautés
  - valoriser les liens absents entre communautés

Notations

- $n_\alpha = |V_\alpha|$
- $E_{\alpha,\beta} = \{(i,j) \in E \ : \ i \in V_\alpha, j \in V_\beta\}, \qquad E_\alpha := E_{\alpha,\alpha}$
- $m_\alpha = |E_\alpha|$
- $m_{\alpha,\beta} = |E_{\alpha,\beta}|$

Alors

$$\begin{aligned}\phi(\mathscr{V}) &= \sum_\alpha m_\alpha - \left(\frac{n_\alpha(n_\alpha-1)}{2} - m_\alpha\right) + \sum_{\alpha\neq\beta} n_\alpha n_\beta - m_{\alpha,\beta} - m_{\alpha,\beta} \\ &= \sum_\alpha 2m_\alpha - \frac{n_\alpha(n_\alpha-1)}{2} + \sum_{\alpha\neq\beta} n_\alpha n_\beta - 2m_{\alpha,\beta}\end{aligned}$$

Tous calculs faits

$$\phi'(\mathscr{V}) = \left(\sum_{\alpha} m_{\alpha} - \sum_{\alpha \neq \beta} m_{\alpha,\beta} \right) - \sum_{\alpha} \frac{n_{\alpha}(n_{\alpha}-1)}{2}$$

où

$$\phi'(\mathscr{V}) = \frac{1}{2} \left(\phi(\mathscr{V}) - \frac{n(n-1)}{2} \right)$$

D'où le problème

$$\begin{cases} \text{given a graph} & G = (V, E) \\ \text{find a partition } \mathscr{V} \text{ of } V & \mathscr{V} = \{V_{\alpha}\}_{\alpha}, \quad \bigsqcup_{\alpha} V_{\alpha} = V \\ \text{such that} & \phi'(\mathscr{V}) \quad \text{maximum} \end{cases}$$

Formulation en physique statistique

Reichardt, J. & Bornholdt, S. – 2006 - *Physical Review E* **(74)** :016110
Fortunato, S. – 2010 – *Physics Reports* **486** :75–174

On définit

$$U(\mathscr{V}) = \sum_{\alpha} m'_{\alpha} + \sum_{\alpha \neq \beta} m_{\alpha,\beta}, \qquad m'_{\alpha} = \#\text{ de liens manquants dans } \quad V_{\alpha}$$

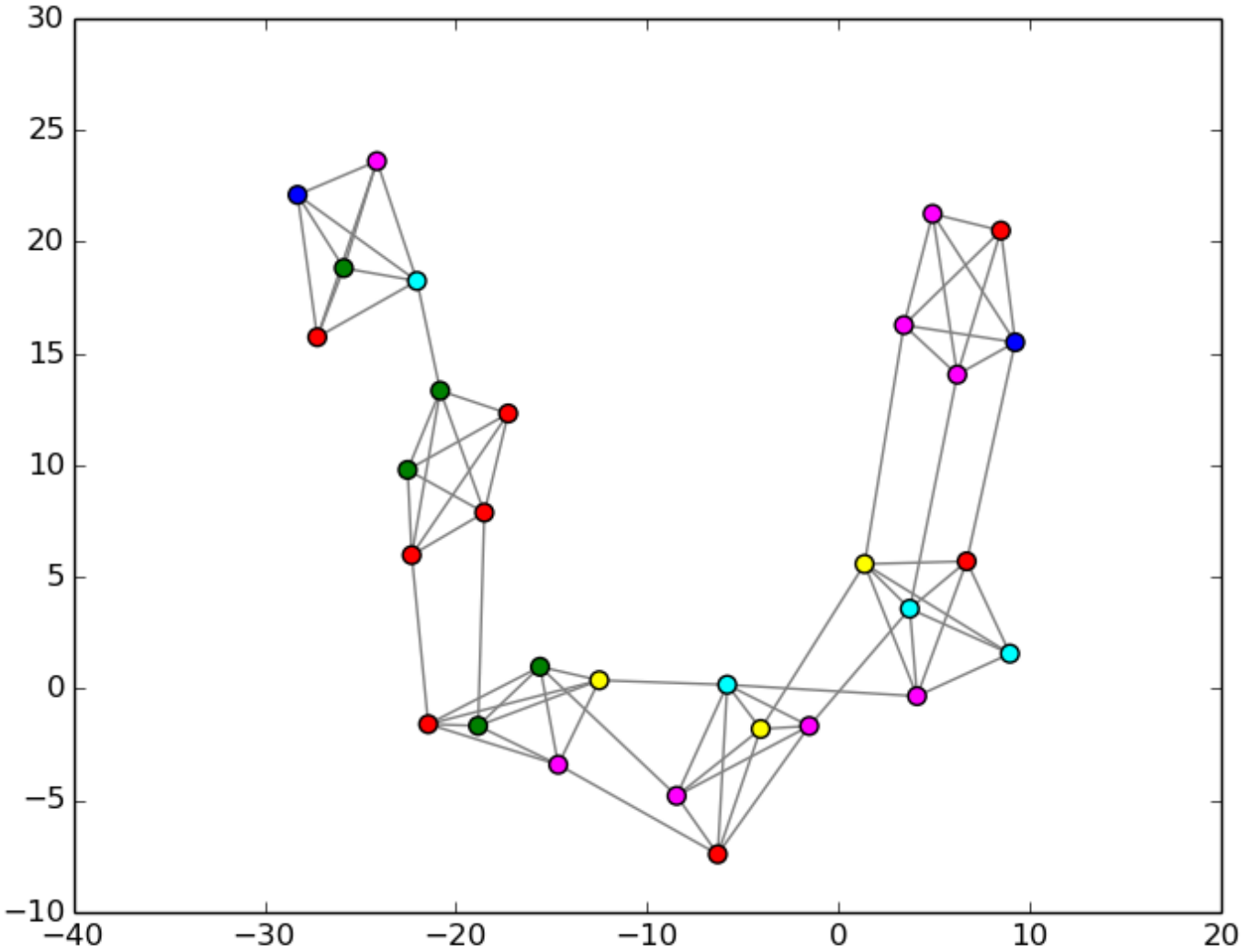
$$U(x) = U(\mathscr{V}), \qquad x \ : \ V \longrightarrow \Lambda, \quad x_i = \alpha \quad \Leftrightarrow \quad i \in V_{\alpha}$$

$$p(x) = \frac{1}{Z} \exp -\beta U(x)$$

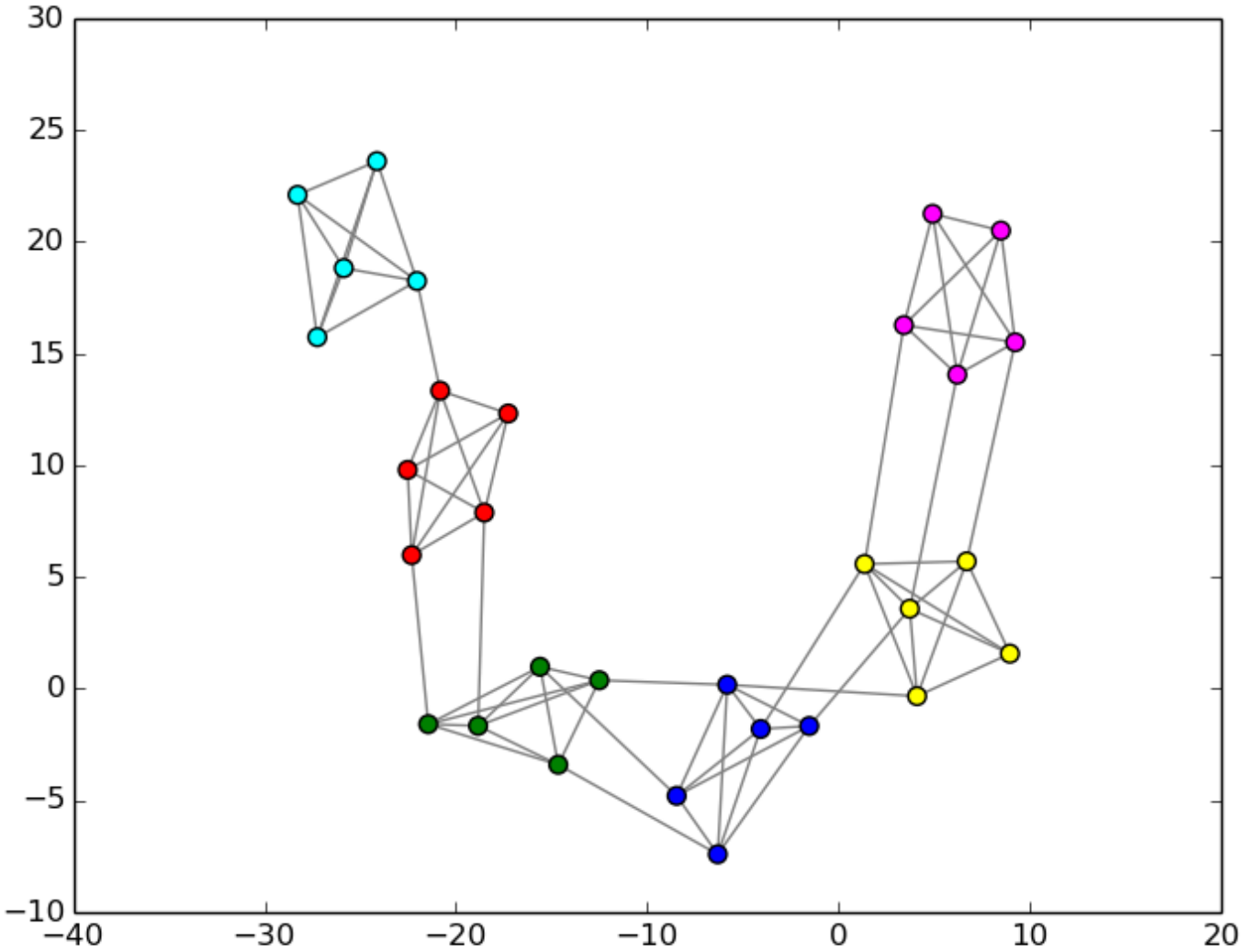
L'objectif est de trouver l'état x^* d'énergie minimale

$$x^* = \underset{x \in \Omega}{\operatorname{argmin}}\ U(x)$$

Recuit simulé



Recuit simulé



Question (recherche pour $n \gg 10^3$)

On se donne

- un ensemble de séquences $\{s_i \ ; \ 1 \leq i \leq n\}$
- un tableau de distances deux à deux $D = [d_{ij}]_{i,j=1,\ldots,n}$; $d_{ij} = d(s_i, s_j)$

Question 1

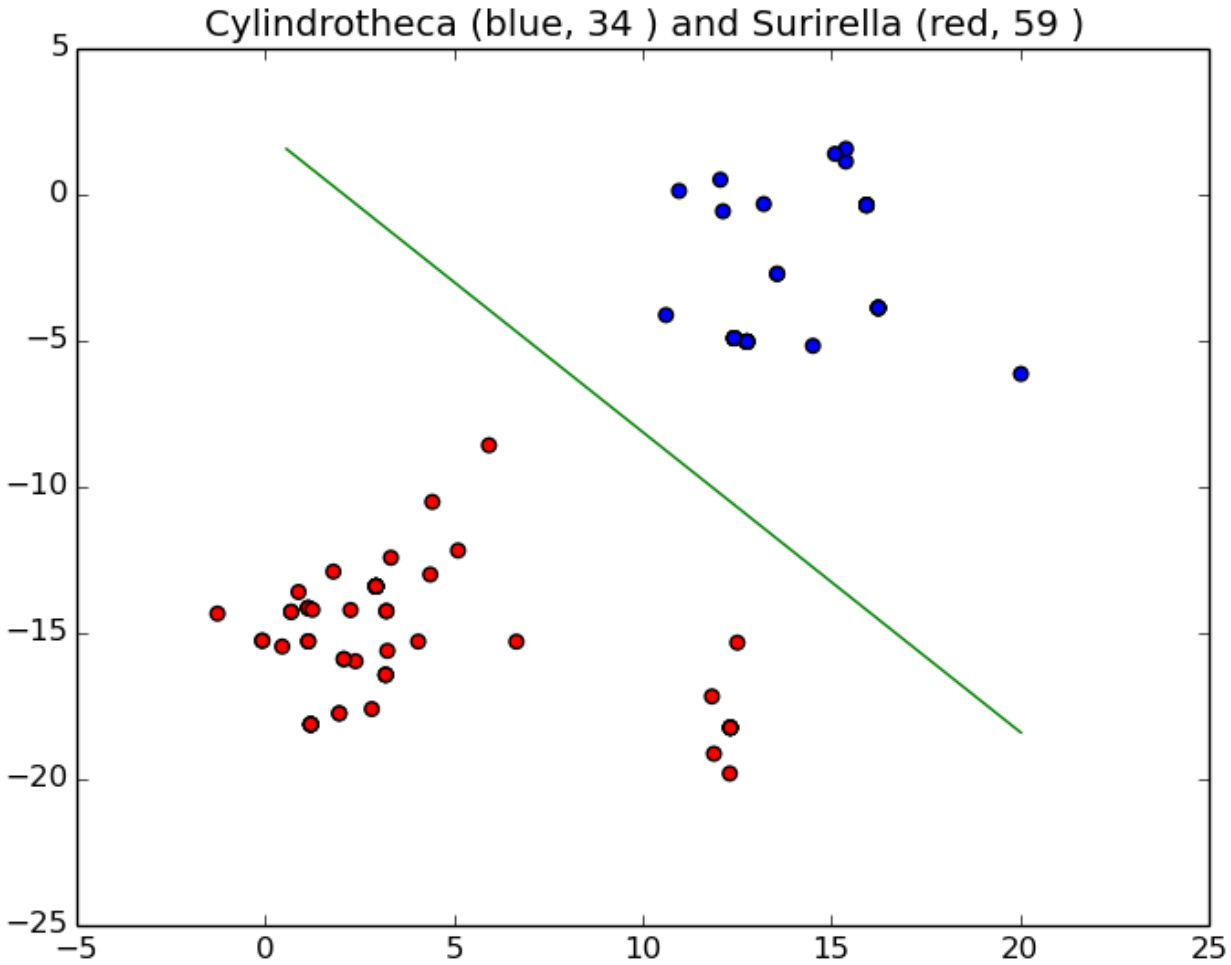
Existe-t-il une dimension $d \in \mathbb{N}$ et un nuage de points
$\mathscr{X} = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
tel que $\quad \forall\, i, j, \quad d(x_i, x_j) = d_{ij}$?

Question 2

Si oui, quelle est la forme de ce nuage ?

Pourquoi un nuage de points ?

Analyse discriminante par SVM



ACP sur un tableau de distances (MDS)

- **Problème**

On se donne	$D = [d_{ij}]$
	une dimension $r \ll n$
On construit	$X_r = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^r$
tels que	$\|x_i - x_j\| \approx d_{ij}$

Résultat

Il existe une solution exacte : méthode spectrale

voir Izenman, 2007, Modern Multivariate Statistical Techniques

Multidimensional Scaling : la solution

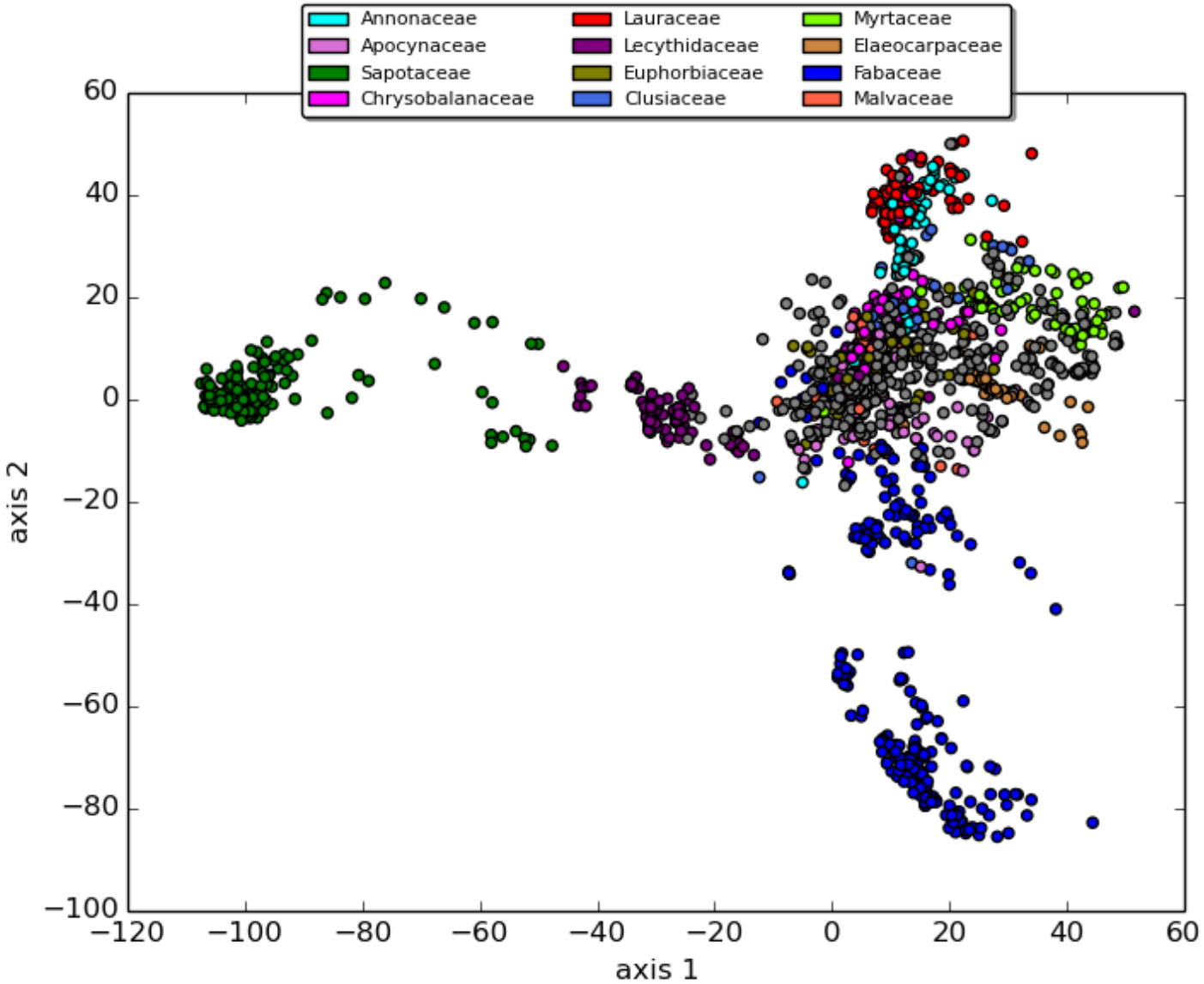
Remarque

Il est possible de reconstruire $\langle x_i, x_j \rangle$ *à partir des* $d_x(i,j) = \|x_i - x_j\|$.

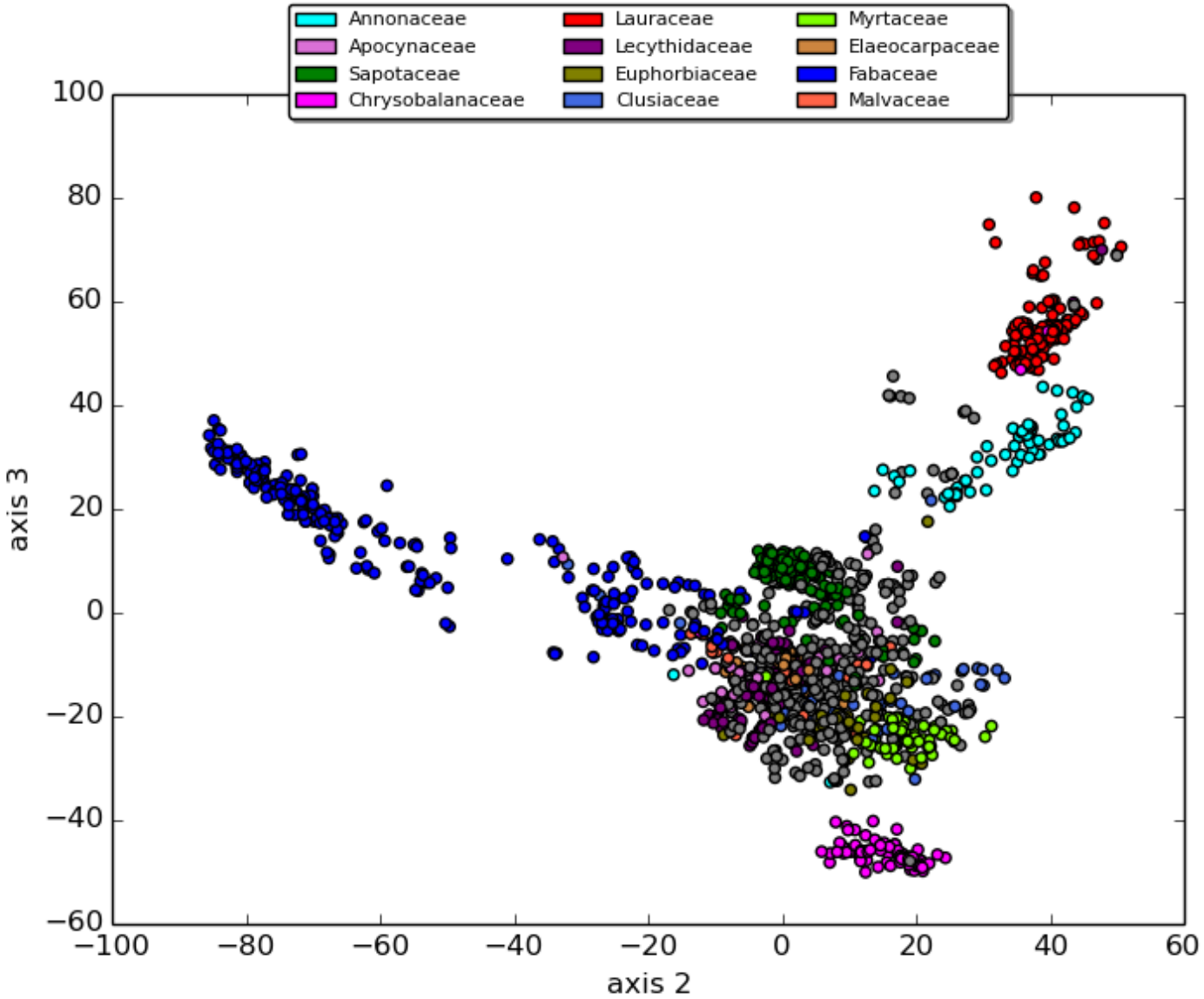
Algorithm 1 pseudocode for Multidimensional Scaling

1: compute $C_{ij} = -\frac{1}{2}(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2), \quad 1 \leq i,j \leq n$
2: compute x_k : $Cx_k = \lambda_k x_k \qquad \{\rightarrow \text{in } O(n^3)\}$
3: $X = (x_k)_k$
4: $L = \text{diag}\,(\sqrt{\lambda_k})_k$
5: $Y = XL$

Exemple sur les arbres de Guyane



Exemple sur les arbres de Guyane



Méthodes non linéaires

- Méthode spectrale (linéaire) $\longrightarrow$ ACP
- Méthodes non linéaires : foisonnement depuis les années 1990 ...
  1. ACP avec noyau (Kernel PCA)
  2. Manifold learning
  3. Isomap
  4. Laplacian eigenmaps
  5. ...

En gros ...

1. travailler de façon linéaire dans un espace de plus grande dimension ...
2. approcher une variété par une collection d'espaces tangents ...
3. naviguer d'un point à un point voisin par une promenade aléatoire ..

Question

Cette enrichissement par des méthodes non linéaires peut-il se transposer en analyse de tableaux de distances ? Si oui, à quel coût de complexité ?

Méthode *Isomap*

Tennebaum J. B. & *al.*, 2000, *Science*, **290** :2319–2323

On se donne un ensemble de distances $d(i,j)$ entre certaines paires de points (ou tous !).

Algorithm 2 pseudocode for Isomap method

1: Have $G = (V, E)$, with weights $w(i,j) = d(i,j)$ on E
2: Select k nearest neighbors of each vertex $i \in V$
3: Build the graph $G' = (V, E')$ induced by edges of k nearest neighbors
4: Compute the shortest path in G' for any pair of vertices (Dijkstra, Floyd)
5: Build D : the pairwise distance matrix with these shortest path lengths
6: Run MDS with D

Bilan :

1. des faiblesses connues : instabilité topologique, sensible à la non convexité de la variété, ...
2. mais plusieurs réussites dans plusieurs domaines ...

De la MDS à la NLM

NLM : *nonlinear mapping*, Sammon, 1969

- MDS $\longrightarrow$ $\|x_i - x_j\| \approx d(i,j), \quad \forall\, d(i,j)$
- NLM :

$$\begin{cases} \phi = \sum_{i<j} \omega_{ij} \left(\|x_i - x_j\|^2 - d_{ij}^2 \right)^2 \\ \omega_{ij} = d_{ij}^{-k}, \quad \exp -\beta d_{ij}, \quad \ldots \end{cases}$$

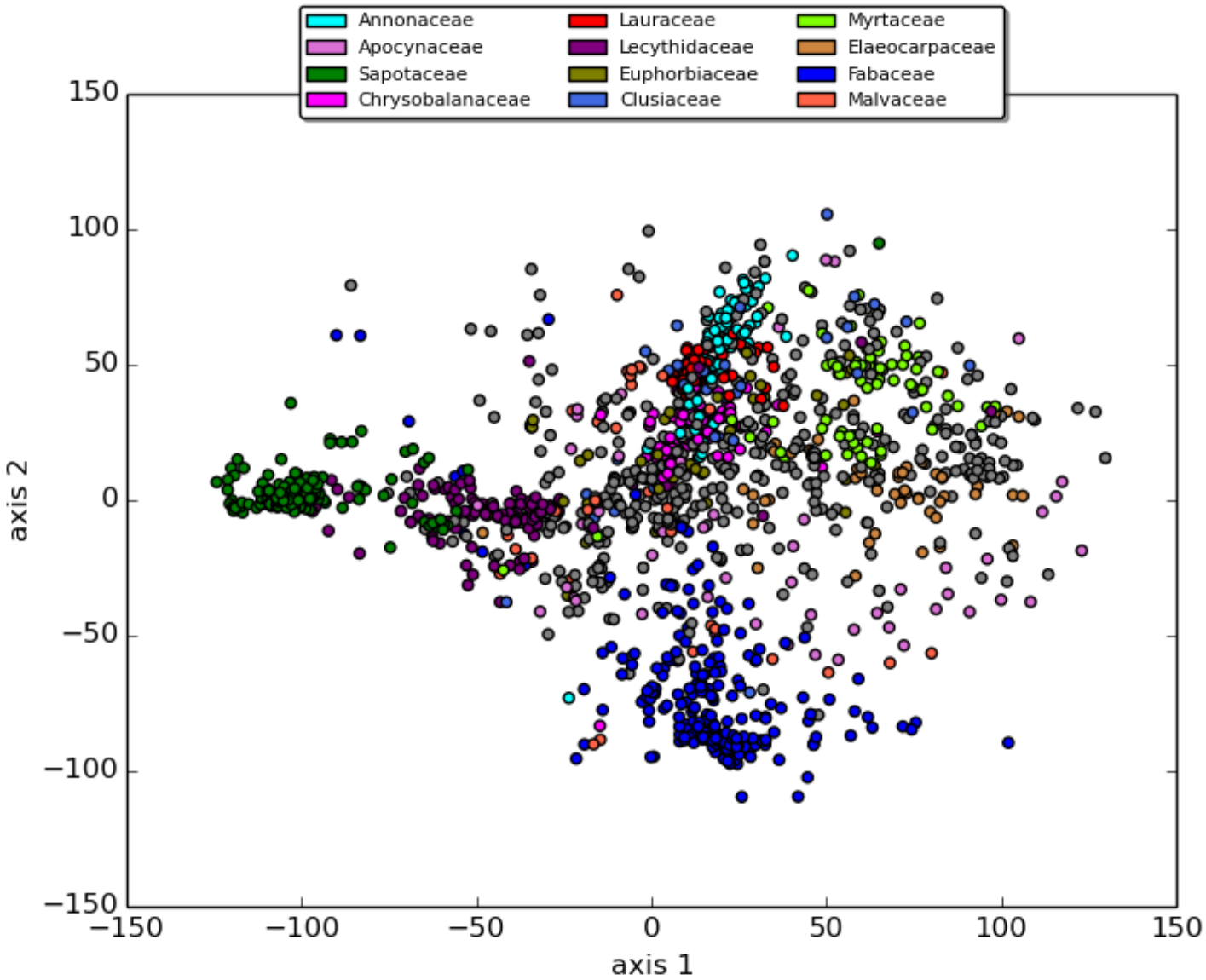
- Résolution (Sammon, 1969) : moindres carrés alternés sur les $[x_{i\alpha}]_i$ (gradient ∇_α sous forme analytique)

Un algorithme séquenciel

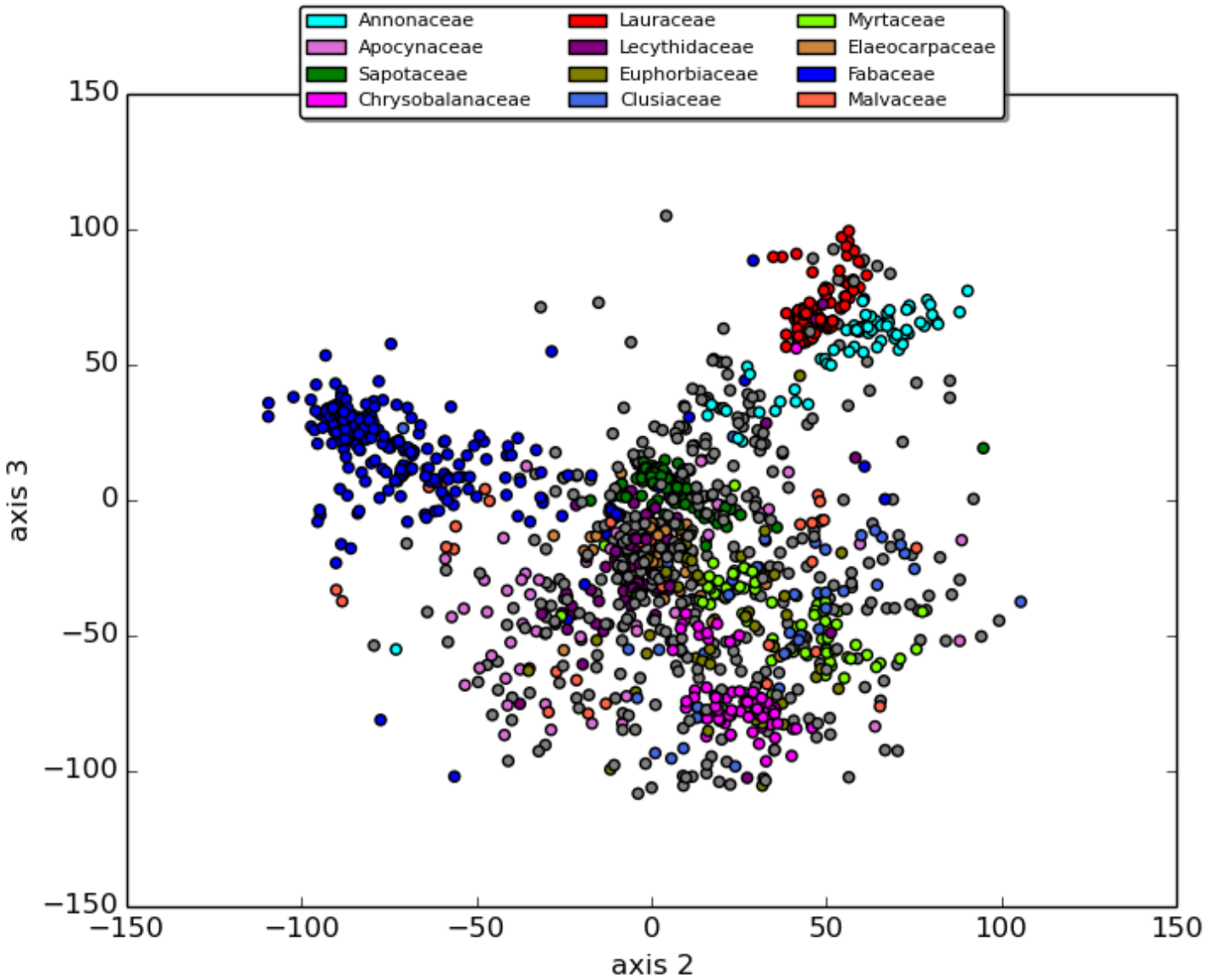
Algorithm 3 pseudocode for sequential optimisation

1: Sequential optimization
2: **for** $t = 1$ to s **do**
3: **for** $i = 1$ to n **do**
4: select x_i
5: compute $\nabla_i = \nabla_z \phi_i(x_i)$
6: find α such that $\phi_i(x_i + \alpha \nabla_i)$ minimum
7: update $x_i \leftarrow x_i + \alpha \nabla_i$
8: **end for**
9: **end for**

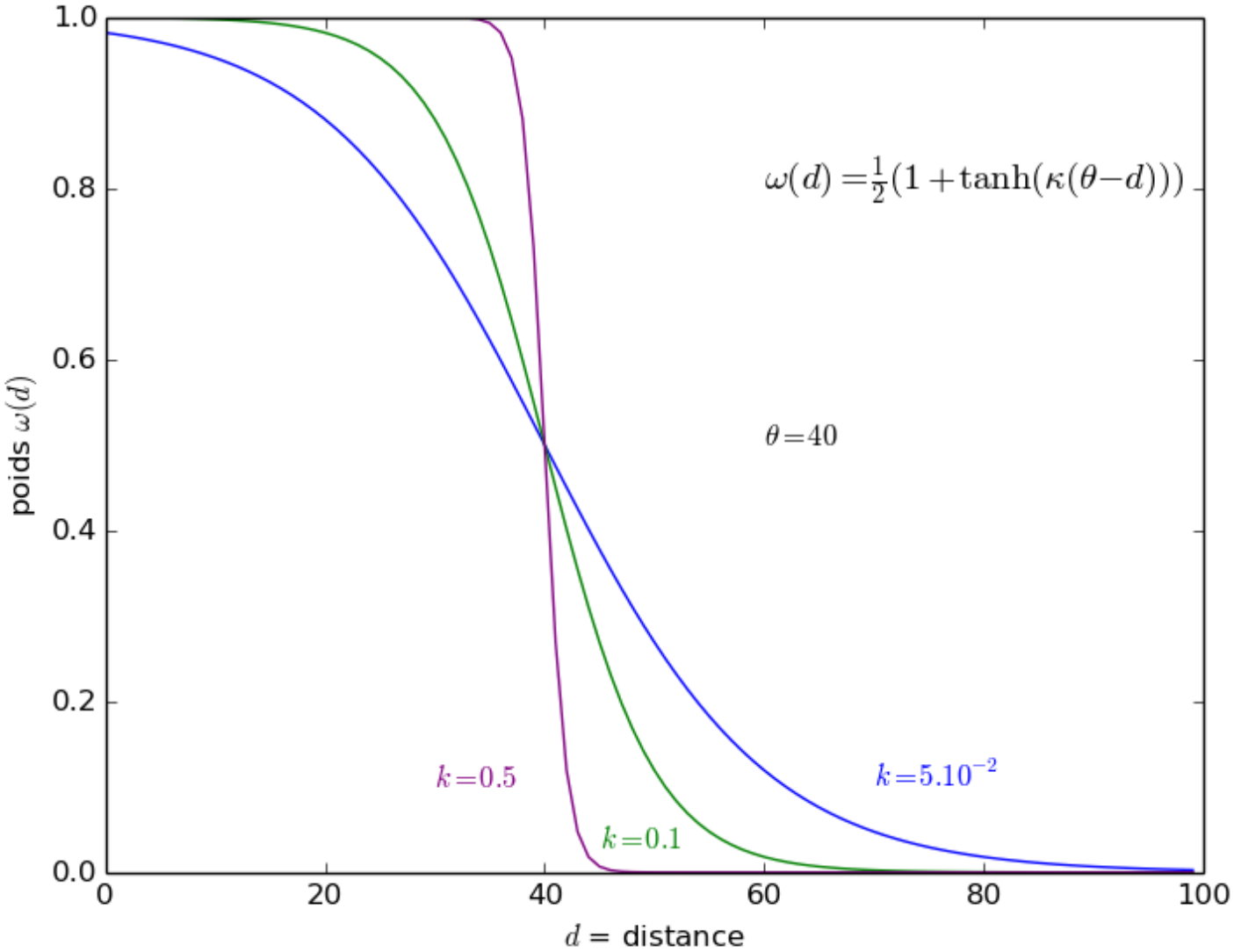
Exemple sur les arbres de Guyane



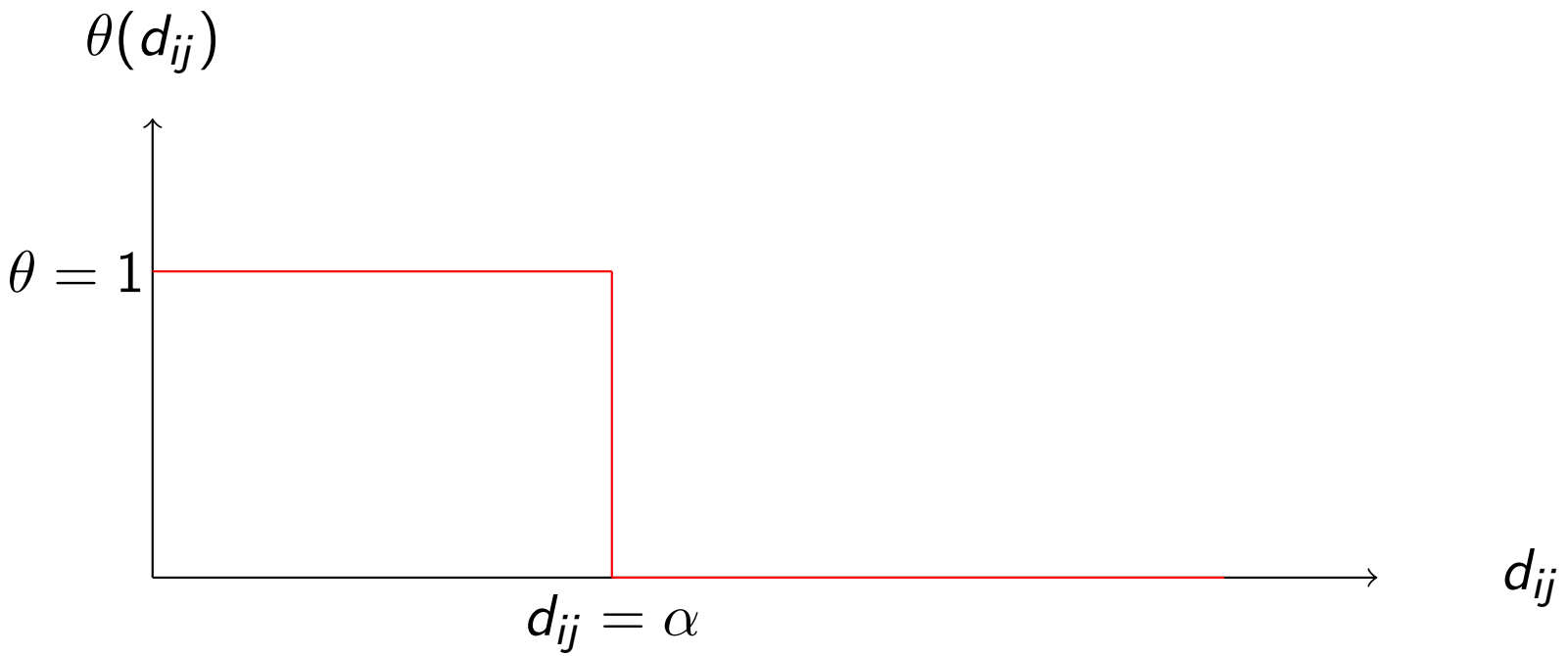
Exemple sur les arbres de Guyane



Convergence vers une distribution



De la *NLM* à la *distance geometry*



$$\begin{cases} d < \alpha & \Rightarrow \quad \theta = 1 \\ d \geq \alpha & \Rightarrow \quad \theta = 0 \end{cases}$$

Convergence en distributions

Soient

$$\begin{cases} \omega(x,\theta,\kappa) = \dfrac{1}{2}\left(1+\tanh(\kappa(\theta-x))\right) \\ H_\theta(x) = H(\theta-x) = \begin{cases} 1 & \text{si} \quad x \leq \theta \\ 0 & \text{si} \quad x > \theta \end{cases} \end{cases}$$

Alors

$$\lim_{\kappa \to +\infty} \omega(x,\theta,\kappa) = H(\theta-x), \qquad \text{au sens des distributions}$$

Question

A chaque poids $\omega(\kappa,\theta)$ on associe un nuage optimal $X(\kappa,\theta)$. A chaque seuil θ, on associe un nuage optimal $\overline{X}(\theta)$ pour les poids H_θ. Comme $\omega(\kappa,\theta) \to H_\theta$, est-ce que $X(\kappa,\theta) \to \overline{X}(\theta)$?

Distance Geometry Problem (DGP)

Lavaur & al., *SIAM Review*, **56(1)** :3-69, 2014

Problem

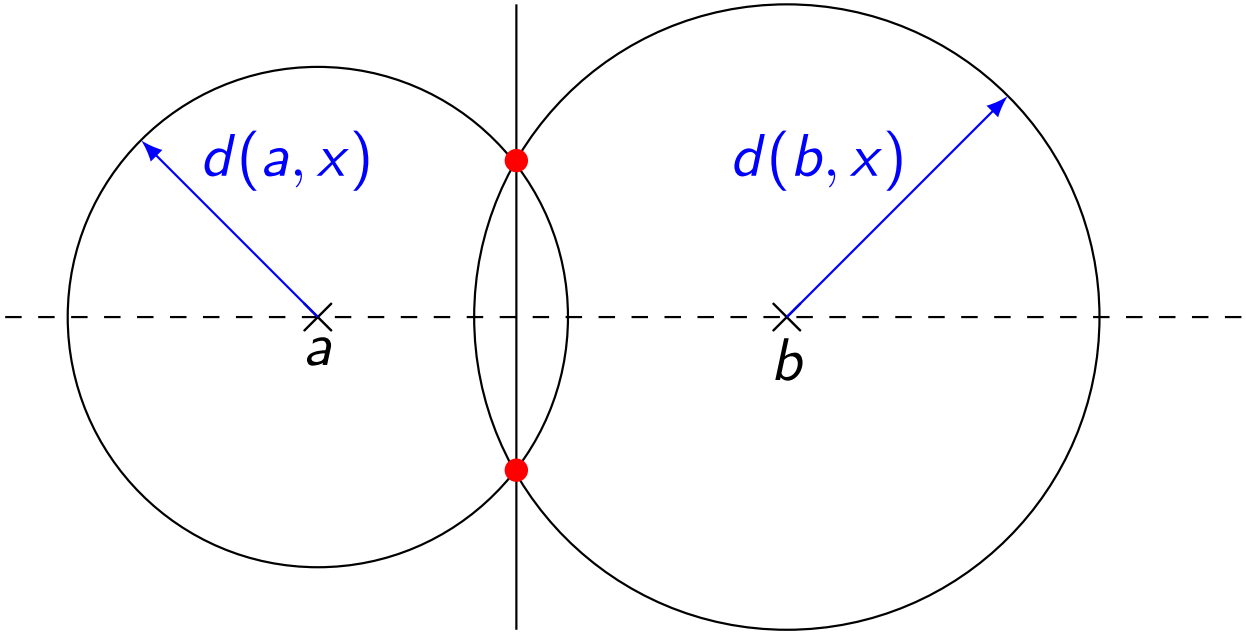
Etant donné un graphe $G = (V, E)$ *avec une fonction de poids* $w(e) \geq 0$ *sur les arêtes, et un entier* $r \geq 1$, *existe-il une fonction*

$$\begin{cases} x : V & \longrightarrow & \mathbb{R}^r \\ i & \longrightarrow & x_i \end{cases}$$

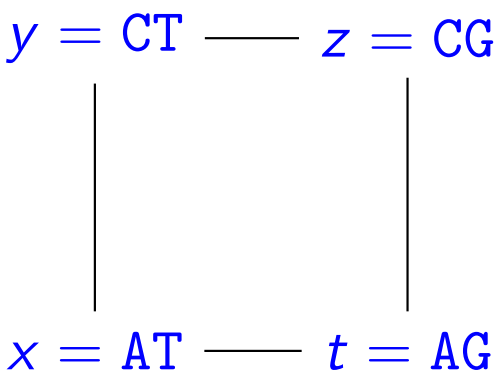
telle que $\|x_i - x_j\| = w(i,j)$ *?*

Méthode	r	représentation
DGP	$n-1$	exacte
MDS	r	approchée
NLM	r	approchée

Triangulation



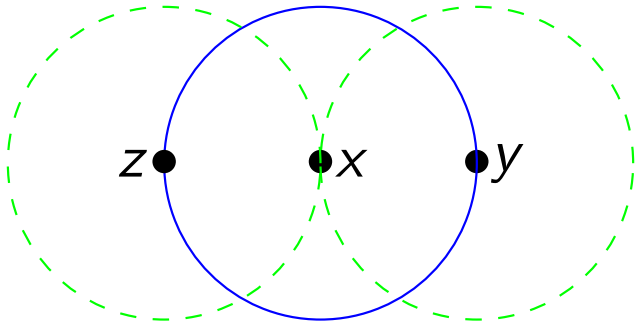
Un exemple d'isométrie impossible ...



$$\forall\, d \in \mathbb{N}, \quad \nexists\, x, y, z, t \in \mathbb{R}^d \quad \text{tq} \quad D(x, y, z, t) = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

Démonstration courte dans $\mathbb{R}^2$

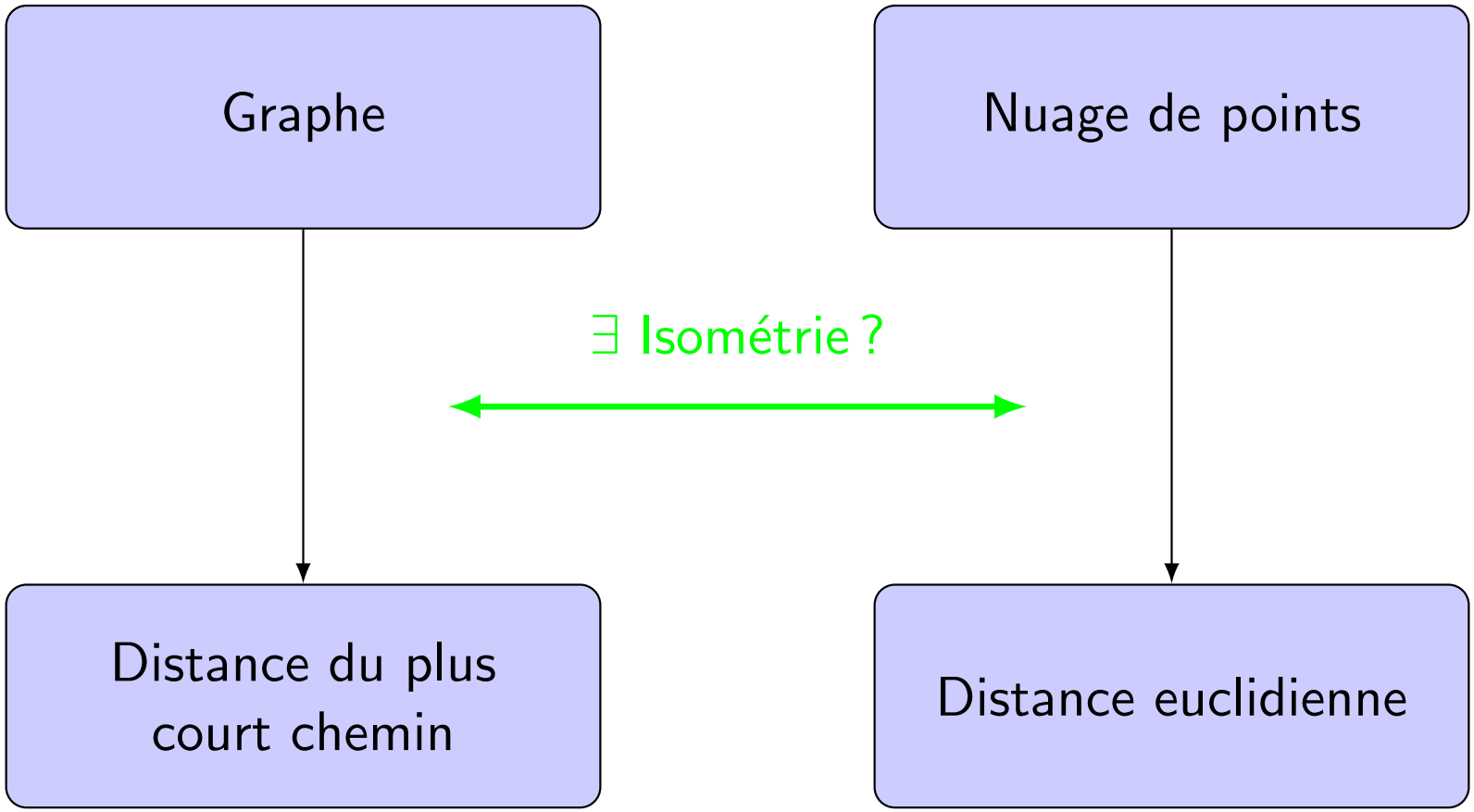
$$\left\{\begin{array}{lllllll} d(x,y) & = & 1 & d(x,z) & = & 1 & d(y,z) = 2 \\ d(y,t) & = & 1 & d(z,t) & = & 1 & \\ d(x,t) & = & 2 & & & & \end{array}\right.$$



Remarque

Cette construction se généralise à l'hypercube $\{0,1\}^n$, avec la distance de Hamming ...

Plus généralement ...



Un théorème du à Fréchet

Notation : $x \in \mathbb{R}^d, \qquad \|x\|_\infty = \max_i |x_i|$

Théorème (Fréchet, 1906 ; Kuratowski, 1935)

Tout espace métrique (X, d) de n points peut être plongé par une isométrie dans $\ell_\infty^n = (\mathbb{R}^n, \|.\|_\infty)$

□ Preuve : On définit

$$\begin{array}{rccc} f : & X & \longrightarrow & \mathbb{R}^n \\ & x & \longrightarrow & (f_1(x), \ldots, f_n(x)) \end{array}$$

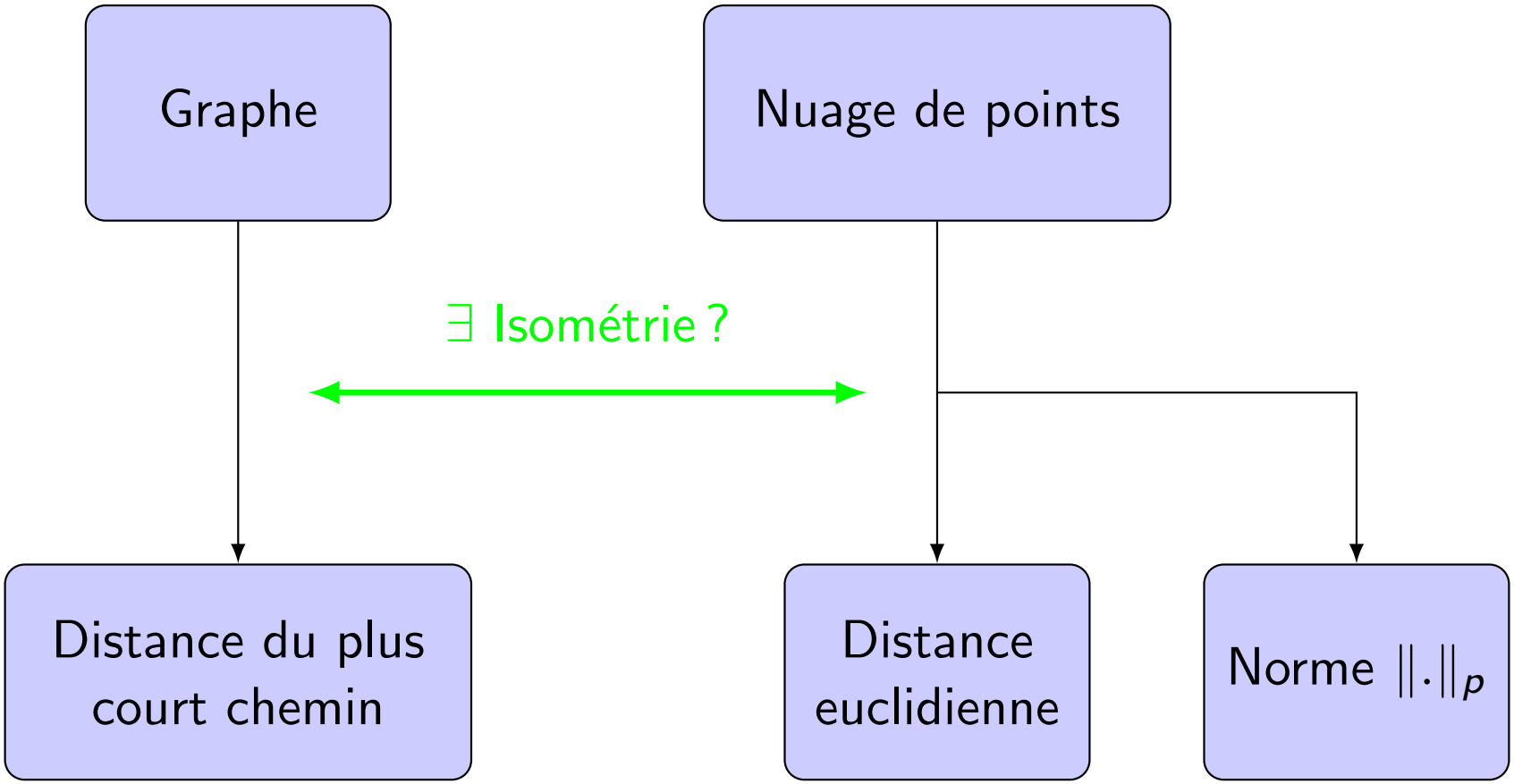
avec

$$f_i(x_j) = d(x_i, x_j)$$

$$\begin{array}{ccc} y = \mathtt{CT} & \text{---} & z = \mathtt{CG} \\ | & & | \\ x = \mathtt{AT} & \text{---} & t = \mathtt{AG} \end{array}$$

$$f(x, y, z, t) = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

Plus généralement ...



Isométries et $D-$plongements

$$f \ : \ x \in X \longrightarrow y = \in Y$$

tel que

$$\forall\, (x, x') \in X, \quad d_Y(y, y') = d_X(x, x')$$

$D-$plongement

$D \geq 1 \in \mathbb{R}$

$$\exists\, r > 0 \quad \text{tq} \quad \forall\, (x, x'), \quad r.d_X(x, x') < d_Y(y, y') < D.r.d_X(x, x')$$

Distorsion de f : infimum sur D tel que f soit un $D-$plongement

$$c_Y(X) = \inf_D \{D \ : \ \exists\, D - \text{plongement } X \longrightarrow Y\}$$

Déterminer $c_Y(X)$ connaissant (X, d_X) et (Y, d_Y) est un problème difficile.

Complexe simplicial abstrait

Mise en bouche ...

- on se donne un nuage de points $\mathscr{X} = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
- on choisit un rayon θ
- on construit l'ensemble des boules $\mathcal{B}_i(\theta) = \{x \in \mathbb{R}^d \ : \ d(x_i, x) \leq \theta\}$

Complexe simplicial abstrait

Mise en bouche ...

- on se donne un nuage de points $\mathscr{X} = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
- on choisit un rayon θ
- on construit l'ensemble des boules $\mathcal{B}_i(\theta) = \{x \in \mathbb{R}^d \ : \ d(x_i, x) \leq \theta\}$

Entrée ...

- On définit $I = \{1, n\} \subset \mathbb{N}$
- on définit l'ensemble $\mathscr{A}$ des parties $A \subset I$ telles que

$$A \in \mathscr{A} \quad \Longleftrightarrow \quad \forall\, i, j \in A, \quad \mathcal{B}_i \cap \mathcal{B}_j \neq \emptyset$$

Complexe simplicial abstrait

Mise en bouche ...

- on se donne un nuage de points $\mathscr{X} = \{x_1, \ldots, x_n\}, \quad x_i \in \mathbb{R}^d$
- on choisit un rayon θ
- on construit l'ensemble des boules $\mathcal{B}_i(\theta) = \{x \in \mathbb{R}^d \ : \ d(x_i, x) \leq \theta\}$

Entrée ...

- On définit $I = \{1, n\} \subset \mathbb{N}$
- on définit l'ensemble $\mathscr{A}$ des parties $A \subset I$ telles que

$$A \in \mathscr{A} \quad \Longleftrightarrow \quad \forall\, i, j \in A, \quad \mathcal{B}_i \cap \mathcal{B}_j \neq \emptyset$$

Plat de résistance ...Complexe Simplicial Abstrait ... (Alexander, 1932)

On vérifie que

$$A \in \mathscr{A}, \quad A' \subset A \quad \Longrightarrow \quad A' \in \mathscr{A}$$

Alors

$$A, A' \in \mathscr{A} \quad \Longrightarrow \quad A \cap A' \in \mathscr{A}$$

Complexes abstraits et taxonomie

$\rightarrow$ Vocabulaire :

- $A \in \mathscr{A}$ est un simplexe
- sa dimension est $\dim A = |A| - 1$
- $\dim \mathscr{A} = \max_A \dim A$

$\rightarrow$ Remarque : les complexes abstraits sont naturels dans l'étude d'intersections d'ensembles, par exemple

- ensemble de cliques
- ensemble d'ensembles convexes

$\rightarrow$ En taxonomie, un taxon peut être représenté naturellement comme une clique, ou un ensemble convexe (sans trous). La TDA permet de repérer les trous, donc la non convexité ...

Quels complexes sur un nuage de points ?

Edelsbrunner & Harer - 2010 - Computational Topology - AMS

On se donne un ensemble de points x_i avec $i \in I = \{1, n\}$. On se fixe $\theta > 0$ et définit donc $\mathcal{B}_i = \{x \ : \ \|x - x_i\| \leq \theta\}$.
On définit alors

- le complexe de Čech

$$C = \left\{ A \subset I \ : \ \bigcap_{i \in A} \mathcal{B}_i \neq \emptyset \right\}$$

- Le complexe de Vietoris-Ripps

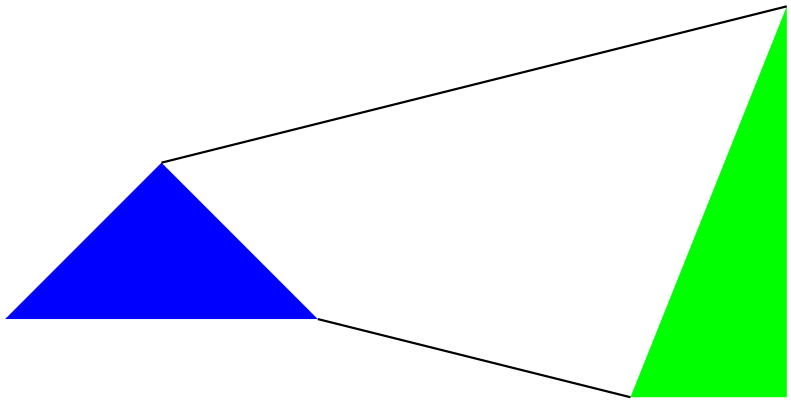
$$R = \{A \subset I \ : \ \forall \, i, j \in A, \ \mathcal{B}_i \cap \mathcal{B}_j \neq \emptyset\}$$

- La complexe α ($\mathcal{V}_i$ est le voisinage de i dans le pavage de Voronoï des centres des sphères)

$$\alpha = \left\{ A \subset I \ : \ \bigcap_{i \in A} (\mathcal{B}_i \cap \mathcal{V}_i) \neq \emptyset \right\}$$

Vocabulaire de graphes

Complexes	graphes
Vietoris-Ripps	Cliques
Čech	Cliques “pleines” (hypergraphes)



Remerciements

- Equipe Pleiade : Philippe Chaumeil, Jean-Marc Frigerio, Franck Salin, David Sherman
- Groupe de Thonon et Uppsala (diatomées) : Frédéric Rimet, Agnès Bouchez, Maria Kahlert
- Groupe de Cayenne : Jean-François Molino, Daniel Sabatier