Deux contributions en clustering à base de graphe : classification hiérarchique ascendante [1] et apprentissage de matrice d'affinités [2]

Julien Ah-Pine
Université Lyon 2 - Laboratoire ERIC
julien.ah-pine@univ-lyon2.fr

Séminaire LIMOS - Axe SIC
17/03/2022

1. http://www.jmlr.org/papers/v19/18-117.html
2. 10.1016/j.ejor.2021.12.034

Rappel du Sommaire

1. Classification ascendante hiérarchique
2. Apprentissage de matrice d'affinités

Motivations et contributions de ce travail

- Améliorer la **sclabilité** de la Classification Ascendante Hiérarchique (AHC) générique de **Lance et Williams (LW)** dénotée D-AHC pour Dissimilarity based AHC.
- Définition d'un **nouveau cadre formel** de la AHC dénoté SNK-AHC (Sparsified and Normalized Kernels-AHC) :
  - $\Rightarrow$ Utilisation de similarités et de fonctions noyaux.
  - $\Rightarrow$ Nouvelle(s) formule(s) de mise à jour des similarités.
  - $\Rightarrow$ Interprétations en termes de similarités pénalisées.
  - $\Rightarrow$ Proprités étudiées (non discutées dans cette présentation) :
    - ⋆ Conditions suffisantes pour la monotonicité.
    - ⋆ Approche "stored data" (ou feature matrix).
    - ⋆ Propriété d'invariance par rapport aux translations de la diagonale.

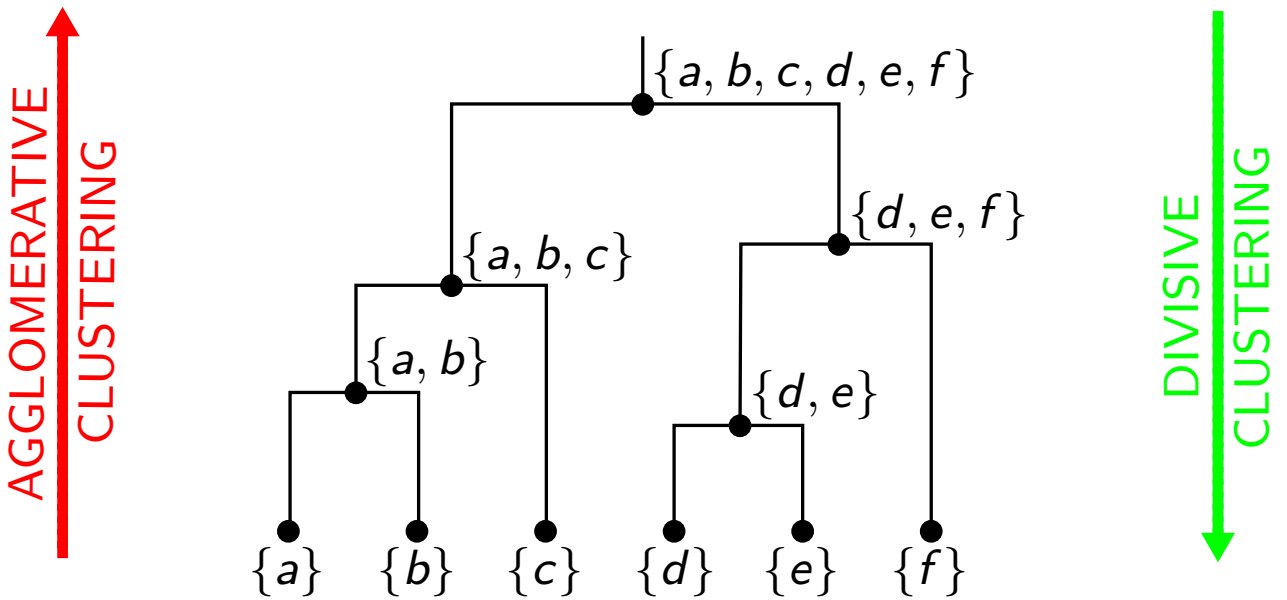
C1 **Hypothèse** : les objets $(a, b, \ldots)$ sont représentés par des vecteurs $(\mathbf{x}^a, \mathbf{x}^b, \ldots)$ dans un RKHS et la matrice de noyaux $\mathbf{S}$ est telle que $\mathbf{S}_{ab} = \langle \mathbf{x}^a, \mathbf{x}^b \rangle$, $\forall a, b$; et la matrice des distances au carré $\mathbf{D}$ est telle que $\mathbf{D}_{ab} = \mathbf{S}_{aa} + \mathbf{S}_{bb} - 2\mathbf{S}_{ab}$, $\forall a, b$.

HC : Hierarchical Clustering

- Famille de méthodes de clustering où on cherche un ensemble de partitions emboîtées.

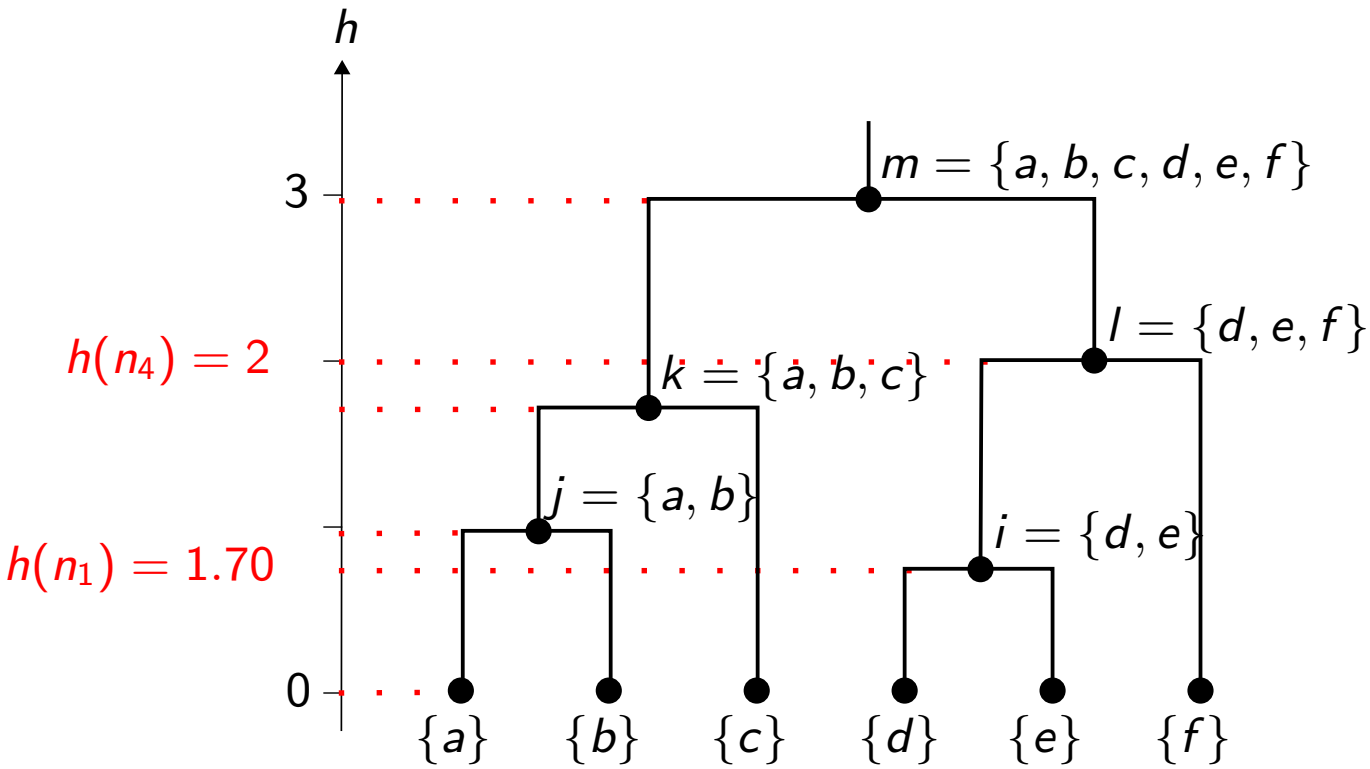
\+ pas de nombre de clusters a priori, cadre flexible (tout type d'objet, plusieurs type de dissimilarités), interprétation (si monotonicité)...

\- $O(n^2)$ en mémoire et $O(n^3)$ en temps de traitement, fonction objectif pas tjs claire...



AHC (Agglomerative HC) et Dendrogramme

- AHC : on part des feuilles et on regroupe itérativement deux clusters.
- Le résultat d'une AHC est un **dendrogramme** : **arbre binaire**, à chaque noeud i on attribue (1) un indice $h(i)$ appelé la **hauteur** et (2) un sous-ensemle d'objets (càd un **cluster**).



D-AHC : Dissimilarity based AHC

- D-AHC : en input on a une matrice $n \times n$ de dissimilarités notée $\mathbf{D}$.

1 Initialisation $\mathbf{D}^1 = \mathbf{D}$.

2 A chaque t, on fusionne la paire (k, l) qui **minimise la dissimilarité** :

$$(k, l) = \underset{(i,j) \in \mathbb{C}^t \times \mathbb{C}^t, i \neq j}{\arg\min} \mathbf{D}^t_{ij}$$

où $\mathbb{C}^t$ est l'ensemble des noeuds à l'itération t.

3 k et l sont fusionnés en un noeud (kl) de hauteur $h((kl)) = \mathbf{D}^t_{kl}$. On **met à jour les dissimilarités** entre (kl) et les clusters existants :

$$\mathbf{D}^{t+1}_{(kl)m}, \forall m \in \mathbb{C}^{t+1}, m \neq (kl)$$

4 On itère pour $t = 2, \ldots, n-1$; la matrice $\mathbf{D}^t$ est d'ordre $n - t + 1$.

La formule générique de Lance et Williams

- Formule paramétrique permettant de regrouper bcp de dissimilarités :

$$\mathbf{D}^{t+1}_{(kl)m} = \alpha(k,l,m)\mathbf{D}^{t}_{km} + \alpha(l,k,m)\mathbf{D}^{t}_{lm} + \beta(k,l,m)\mathbf{D}^{t}_{kl} + \gamma|\mathbf{D}^{t}_{km} - \mathbf{D}^{t}_{lm}|$$

où $\gamma \in \mathbb{R}$ et α, β sont des **fonctions d'ensemble** à valeurs dans $\mathbb{R}$.

Method	$\alpha(k,l,m)$	$\beta(k,l,m)$	γ
Single link.	$1/2$	0	$-1/2$
Complete link.	$1/2$	0	$1/2$
Group aver.	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	0	0
Mcquitty	$1/2$	0	0
Centroid	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$-\frac{\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	0
Median	$1/2$	$-1/4$	0
Ward	$\frac{\lvert k\rvert+\lvert m\rvert}{\lvert k\rvert+\lvert l\rvert+\lvert m\rvert}$	$-\frac{\lvert m\rvert}{\lvert k\rvert+\lvert l\rvert+\lvert m\rvert}$	0

- On traite les cas $\gamma = 0$ (single et complete linkage sont exclus). On utilise alors une version équivalente de la formule de LW.

Une version équivalente de la formule de LW

- On fusionne la paire (k, l) qui minimise la **dissimilarité pondérée** :

$$\boxed{(k,l) = \underset{(i,j)\in\mathbb{C}^t\times\mathbb{C}^t, i\neq j}{\arg\min} p(i,j)\mathbf{D}^t_{ij}} \quad (1)$$

- Pour calculer la dissimilarité entre (kl) et $m \in \mathbb{C}^{t+1}, m \neq (kl)$:

$$\boxed{\mathbf{D}^{t+1}_{(kl)m} = \alpha(k,l)\mathbf{D}^t_{km} + \alpha(l,k)\mathbf{D}^t_{lm} + \beta(k,l)\mathbf{D}^t_{kl}} \quad (2)$$

où p , α et β sont des fonctions d'ensemble de **deux arguments**.

Method	$\alpha(k,l)$	$\beta(k,l)$	$p(i,j)$
Group aver.	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	0	1
Mcquitty	$1/2$	0	1
Centroid	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$-\frac{\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	1
Median	$1/2$	$-1/4$	1
Ward	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$-\frac{\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$
W-Median	$1/2$	$-1/4$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$

K-AHC : Kernel based AHC

- K-AHC : en input on a une matrix $n \times n$ de noyaux $\mathbf{S}$.

1 Initialisation : $\mathbf{S}^1 = \mathbf{S}$.

2 A chaque t, on fusionne la paire (k, l) qui **maximise** :

$$\boxed{(k,l) = \underset{(i,j)\in\mathbb{C}^t\times\mathbb{C}^t, i\neq j}{\arg\max} \mathfrak{p}(i,j)\left(\mathbf{S}^t_{ij} - \frac{1}{2}\left(\mathbf{S}^t_{ii} + \mathbf{S}^t_{jj}\right)\right)} \quad (3)$$

3 k et l sont regroupées en (kl) de "*profondeur*" $d((kl)) = \mathfrak{p}(i,j)\left(\mathbf{S}^t_{kl} - \frac{1}{2}\left(\mathbf{S}^t_{kk} + \mathbf{S}^t_{ll}\right)\right)$ et **on met à jour la similarité entre** (kl) **et les autres classes**, $m \in \mathbb{C}^{t+1}$ **et elle même** par :

$$\boxed{\begin{aligned} \mathbf{S}^{t+1}_{(kl)m} &= \mathfrak{a}(k,l)\mathbf{S}^t_{km} + \mathfrak{a}(l,k)\mathbf{S}^t_{lm} \quad (4)\\ \mathbf{S}^{t+1}_{(kl)(kl)} &= \mathfrak{b}(k,l)\mathbf{S}^t_{kl} + \mathfrak{c}(k,l)\mathbf{S}^t_{kk} + \mathfrak{c}(l,k)\mathbf{S}^t_{ll} \quad (5)\end{aligned}}$$

où $\mathfrak{p}, \mathfrak{a}, \mathfrak{b}, \mathfrak{c}$ sont des fonctions d'ensemble.

4 On itère pour $t = 2, \ldots, n-1$; la matrice $\mathbf{S}^t$ est d'ordre $n - t + 1$.

Paramètres des methodes dans le cadre de K-AHC

- **2 formules de mise à jour** de **S** sont nécessaires.
- Les fonctions d'ensemble sont définies comme suit :

$$\mathfrak{a} = \alpha;\ \mathfrak{b} = -2\beta;\ \mathfrak{c} = \alpha + \beta;\ \mathfrak{p} = p;$$

  α, β, p étant les fonctions d'ensemble intervenant dans (1) et (2).
- Les méthodes sont alors définies dans K-AHC par :

Method	$\mathfrak{a}(k,l)$	$\mathfrak{b}(k,l)$	$\mathfrak{c}(k,l)$	$\mathfrak{p}(i,j)$
Group average	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	0	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	1
Mcquitty	1/2	0	1/2	1
Centroid	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$\frac{2\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert k\rvert^2}{(\lvert k\rvert+\lvert l\rvert)^2}$	1
Median	1/2	1/2	1/4	1
Ward	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$\frac{2\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert k\rvert^2}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$
W-Median	1/2	1/2	1/4	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$

Similarités pénalisées

- Rappelons la règle de fusion (3) :

$$(k, l) = \underset{(i,j)\in\mathbb{C}^t\times\mathbb{C}^t, i\neq j}{\arg\max} \overbrace{\mathfrak{p}(i,j)}^{\text{Poids}} \underbrace{\left(\mathbf{S}_{ij}^t - \frac{1}{2}\left(\mathbf{S}_{ii}^t + \mathbf{S}_{jj}^t\right)\right)}_{\text{Sim. pénalisée}}$$

- On a une **matrice de similarités pénalisées** de terme général :

$$\overbrace{\mathbf{S}_{ij}^t}^{\text{Inter-sim.}} - \underbrace{\frac{1}{2}(\overbrace{\mathbf{S}_{ii}^t + \mathbf{S}_{jj}^t}^{\text{Intra-sim.}})}_{\text{Pénalité}}$$

- Pénalité = moyenne arithmétique des intra-similarités.
- Deux clusters fortement homogènes et relativement peu similaires entre eux, se regrouperont tardivement dans la hiérarchie.

Similarités pénalisées : Différentes stratégies d'agrégation

- Rappelons les formules de mise à jour :

$$\mathbf{S}^{t+1}_{(kl)m} = \mathfrak{a}(k,l)\mathbf{S}^{t}_{km} + \mathfrak{a}(l,k)\mathbf{S}^{t}_{lm}$$
$$\mathbf{S}^{t+1}_{(kl)(kl)} = \mathfrak{b}(k,l)\mathbf{S}^{t}_{kl} + \mathfrak{c}(k,l)\mathbf{S}^{t}_{kk} + \mathfrak{c}(l,k)\mathbf{S}^{t}_{ll}$$

- L'ensemble des méthodes à l'étude sont telles que : $\forall k, l \in 2^{\mathbb{O}}$,

$$\begin{cases} \mathfrak{a}(k,l), \mathfrak{b}(k,l), \mathfrak{c}(k,l) \geq 0 \\ \mathfrak{a}(k,l) + \mathfrak{a}(l,k) = 1 \\ \mathfrak{b}(k,l) + \mathfrak{c}(k,l) + \mathfrak{c}(l,k) = 1 \end{cases}$$

  Elles définissent donc des **moyennes pondérées**.

- La différence entre les méthodes peut s'interpréter comme **différentes stratégies d'agrégation** d'inter-similarités et d'intra-similarités :

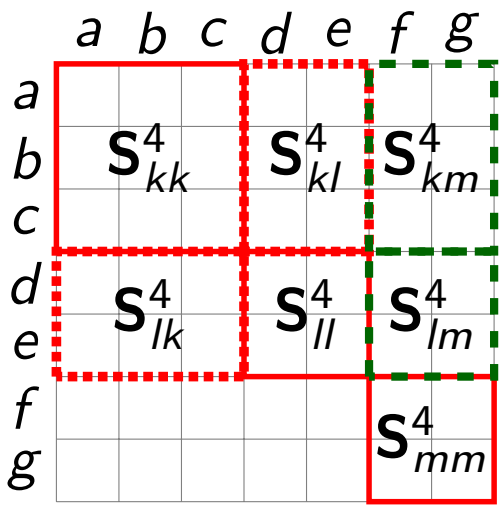
sim. pénalisée=moyenne sim inter - moyenne(moyennes sim intra)

Similarités pénalisées : Illustration

- $k = \{a, b, c\}; l = \{d, e\}; m = \{f, g\}$
- k et l fusionnent.

$$\mathbf{S}^{t+1}_{(kl)m} = \mathfrak{a}(k,l)\mathbf{S}^{t}_{km} + \mathfrak{a}(l,k)\mathbf{S}^{t}_{lm}$$

$$\mathbf{S}^{t+1}_{(kl)(kl)} = \mathfrak{b}(k,l)\mathbf{S}^{t}_{kl} + \mathfrak{c}(k,l)\mathbf{S}^{t}_{kk} + \mathfrak{c}(l,k)\mathbf{S}^{t}_{ll}$$



Method	$\mathfrak{a}(k,l)$	$\mathfrak{b}(k,l)$	$\mathfrak{c}(k,l)$	$\mathfrak{p}(i,j)$
Group average	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	0	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	1
Mcquitty	1/2	0	1/2	1
Centroid	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$\frac{2\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert k\rvert^2}{(\lvert k\rvert+\lvert l\rvert)^2}$	1
Median	1/2	1/2	1/4	1
Ward	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}$	$\frac{2\lvert k\rvert\lvert l\rvert}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert k\rvert^2}{(\lvert k\rvert+\lvert l\rvert)^2}$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$
W-Median	1/2	1/2	1/4	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$

SNK-AHC : Sparsified and Normalized Kernel based AHC

C2 Hypothèse : $\mathbf{S}_{aa} = \mathbf{S}_{bb}, \forall a, b$ (vecteurs sur une hypershpère).

C3 Hypothèse : $\mathbf{S}_{ab} \geq 0, \forall a, b$ (vecteurs dans l'orthant positif).

- Etant donné **S** on peut tjs la **normaliser** (cosinus) puis la **translater**.
- On **sparsifie S** (graphe des plus proches voisins) :
  - D-AHC et K-AHC ont un coût en $O(n^3)$: il faut $n-1$ itérations pour construire le dendrogramme et à chaque itération, il faut trouver la paire optimale qui coûte au pire des cas $O(n^2)$.

  $\Rightarrow$ Cette recherche est le **goulot d'étranglement**. Dans SNK-AHC, on cherche cette paire dans le sous-ensemble des paires de similarités strictement positives. On introduit les sous-ensembles suivants :

$$\mathbb{S}^t = \{(i,j) \in \mathbb{C}^t \times \mathbb{C}^t, \mathbf{S}^t_{ij} > 0\}, \forall t \tag{6}$$

- SNK-AHC est ainsi plus **scalable** que D-AHC ou K-AHC.
- SNK-AHC détermine les **composantes connexes** du graphe.

SNK-AHC : Sparsified and Normalized Kernel based AHC (suite)

- SNK-AHC : en input on a une matrix $n \times n$ de noyaux $\mathbf{S}$.

1 Normaliser, translater et sparsifier $\mathbf{S}$.

2 Initialisation : $\mathbf{S}^1 = \mathbf{S}$ et calcul de $\mathbb{S}^1$.

3 A chaque t, on fusionne la paire (k, l) qui **maximise** :

$$\boxed{(k,l) = \underset{(i,j)\in\mathbb{S}^t\times\mathbb{S}^t, i\neq j}{\arg\max} \mathfrak{p}(i,j)\left(\mathbf{S}^t_{ij} - \frac{1}{2}\left(\mathbf{S}^t_{ii} + \mathbf{S}^t_{jj}\right)\right)} \quad (7)$$

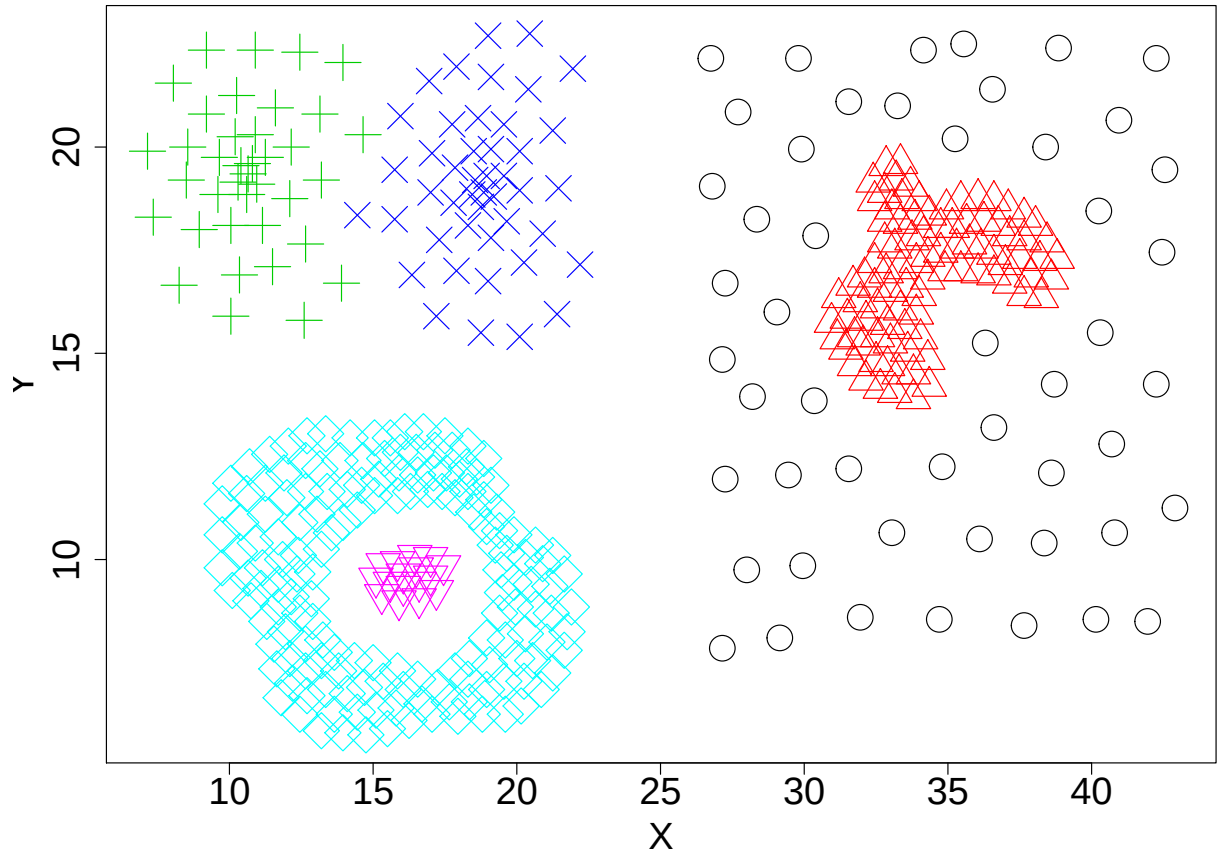
3 k et l sont regroupées en (kl) et on applique les mêmes formules de mise à jour de K-AHC (4) et (5).

4 On itère **jusqu'à** $\mathbb{S}^t = \emptyset$, le **nb de clusters** [3] vaut alors $n - t$.

3. On fusionne deux clusters k et l que s'ils sont connectés ($\mathbf{S}^t_{kl} > 0$) et la procédure AHC est similaire à un algorithme de détection des composantes connexes d'un graphe (càd des clusters).

Données “compound”

- 399 observations en 2D.
- 6 clusters.



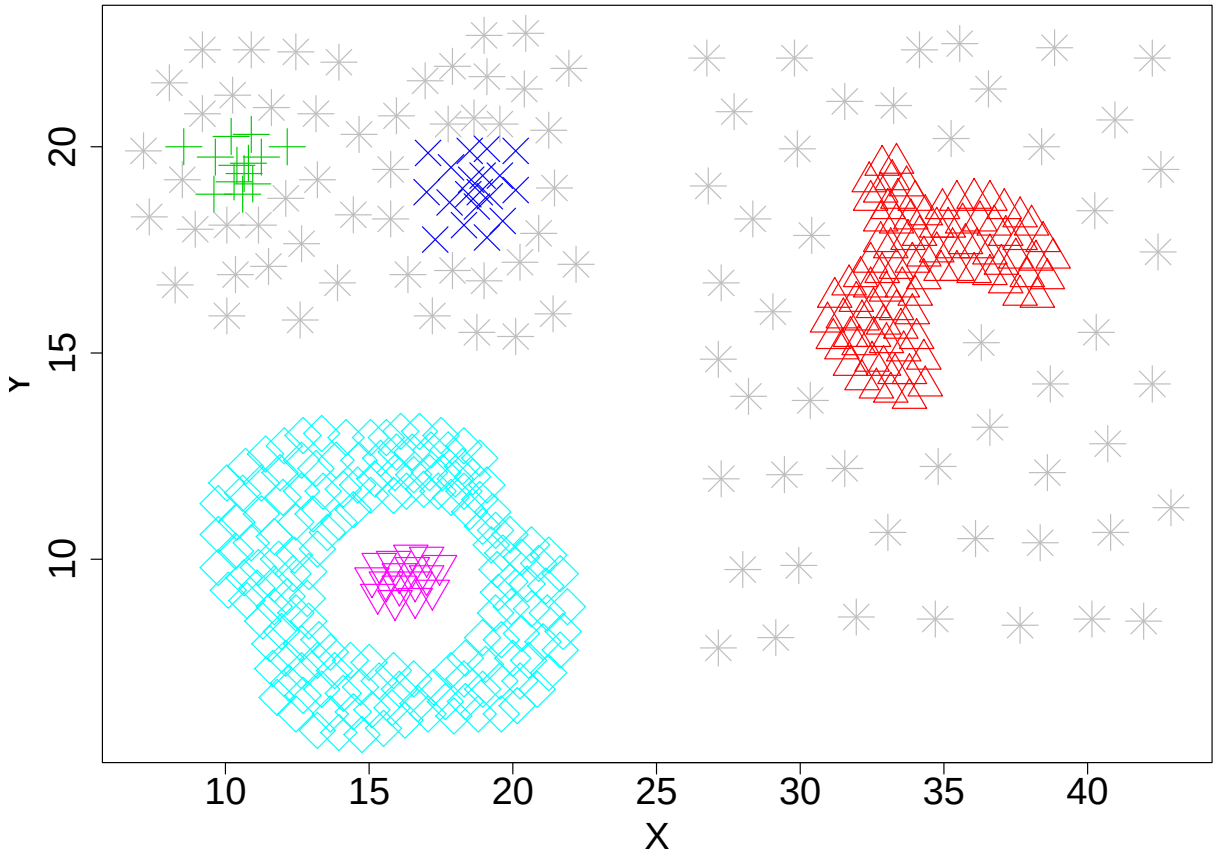
Résultats sur les données "compound"

- Résultats avec noyau Gaussien et la méthode group average.
- θ est le seuil de sparsification : $\mathbf{S}_{ab} \leftarrow \mathbf{S}_{ab}\mathbf{1}_{\{\mathbf{S}_{ab} \geq \theta\}}, \forall a, b$.
- CC est le coefficient de corrélation cophénétique.
- ARI est l'indice de Rand corrigé.
- κ est le nombre de composantes connexes détecté.

Method	θ	CC	ARI	κ
Group average	0.010	1.000	0.811	1
	0.143	1.000	0.811	1
	0.245	1.000	0.811	1
	0.463	0.999	0.811	1
	0.819	0.947	0.802	1
	0.948	-0.766	0.818	3
	0.996	-0.741	**0.906**	99

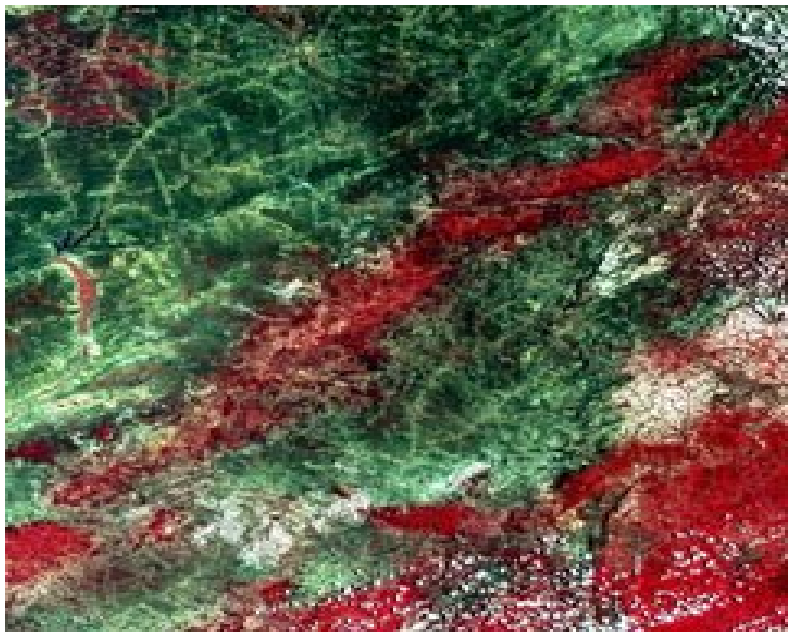
Résultats sur les données "compound" (suite)

- Illustration avec $\theta = 0.996$, on obtient 99 composantes connexes.
- Si les clusters sont de taille ≤ 3, on représente les individus en gris.



Données "landsat"

- Images satellites multispectrales. Jeu de données disponible sur UCI [4].
- 6,435 observations (1 obs = patch de 3x3 pixels)
- 36 variables (chaque pixel a 4 valeurs de spectres)
- 6 clusters : red soil, cotton crop, grey soil, damp grey soil, soil with vegetation stubble, very damp grey soil.



4. https://archive.ics.uci.edu/ml/datasets/Statlog+(Landsat+Satellite)

Résultats sur les données "landsat"

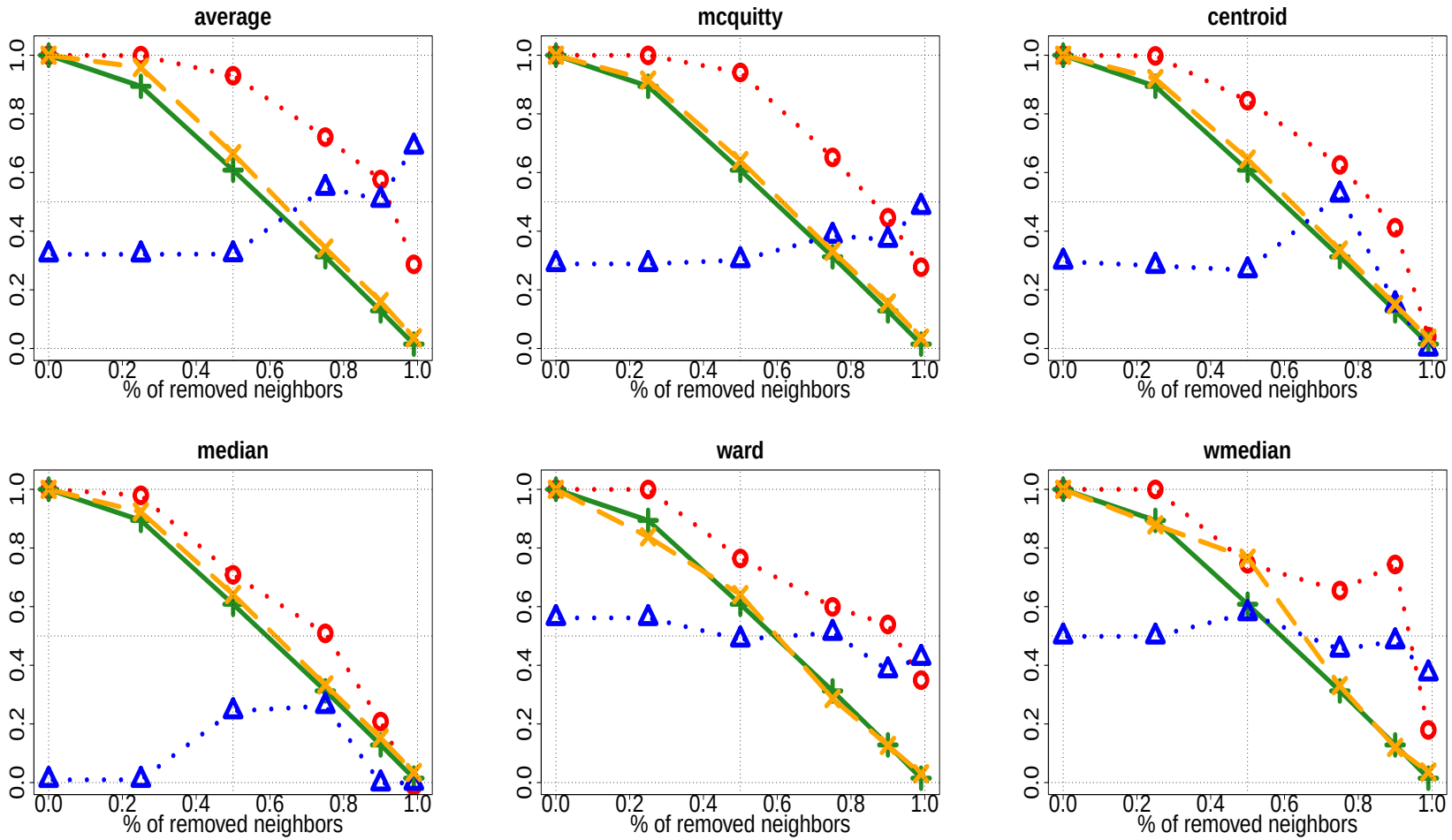


FIGURE – The x-axis corresponds to the % of removed neighbors. The y-axis corresponds to the observed values in $[0, 1]$. Solid lines with plus signs represent the relative memory use, dotted lines with circles indicates the CC values, dotted lines with triangles give the ARI values.

Travaux futurs

- Quelques pistes pour des travaux futurs :
  - Comment sparsifier **S** ? Selon les applications, le voisinage peut être donné par des informations externes (données images, spatiales, temporelles, . . .)
  - Peut-on utiliser d'autres opérateurs d'agrégation que les moyennes pondérées ?
  - Les méthodes invariantes par rapport aux translations de la diagonale semblent mieux marcher que les autres notamment quand **S** est très sparse, pourquoi ?
  - . . .

Rappel du Sommaire

1 Classification ascendante hiérarchique

2 Apprentissage de matrice d'affinités

Motivations

- Graph based clustering :
  - Input is a pairwise similarity or affinity matrix (such as a kernel matrix).
  - Clustering is a graph partitioning problem.
- Bad affinity matrix lead to bad clustering results.
- Task : Transform an initial affinity matrix for clustering purposes.
- This contribution's viewpoint :
  - The **pairwise clustering matrix** is the ideal affinity matrix.
  - The graph partitioning is an NP-hard problem.
  - Learning an affinity matrix is approximating a clustering matrix.
  - Key point : a clustering matrix is a **doubly stochastic and idempotent matrix**.
- This contribution is about a **new relaxed optimization problem** that outputs a **doubly stochastic and nearly-idempotent (DSNI) matrix** in the goal of having a good approximation of the correct clustering matrix. From the DSNI matrix a valid clustering matrix is recovered by using spectral clustering.

Sinkhorn theorem, 1968

Theorem ([Sinkhorn, 1968])

$\mathbf{X} \in \mathbb{R}^{n \times n}$ *is* **doubly stochastic** *and* **idempotent** *iff there exists positive numbers* $n_1, \ldots, n_k$ *which sum to* n *and a permutation mat.* $\mathbf{P}$ *such that :*

$$\mathbf{X} = \mathbf{P} \begin{pmatrix} \mathbf{J}_{n_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{n_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{J}_{n_k} \end{pmatrix} \mathbf{P}^\top \tag{8}$$

where $\mathbf{J}_{n_l} \in \mathbb{R}^{n_l \times n_l}$ *with* $1/n_l$ *in all entries.*

- Example : $\{\{a, b, d\}, \{c\}\} \rightarrow \mathbf{W} =$

$$\begin{array}{c} \\ a \\ b \\ c \\ d \end{array} \begin{array}{c} \begin{array}{cccc} a & b & c & d \end{array} \\ \begin{pmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \end{pmatrix} \end{array}.$$

More detailed example

- Block diagonal representation :

$$\mathbf{X} = \underbrace{\begin{array}{c} \\ a \\ b \\ c \\ d \end{array}\begin{array}{c}\begin{array}{cccc} a & b & c & d \end{array}\\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}\end{array}}_{\mathbf{P}} \overbrace{\begin{array}{c}\begin{array}{cccc} a & b & c & d \end{array}\\ \begin{pmatrix} 1/3 & 1/3 & 0 & 1/3 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 1 & 0 \\ 1/3 & 1/3 & 0 & 1/3 \end{pmatrix}\end{array}}^{\mathbf{W}} \underbrace{\begin{array}{c}\begin{array}{cccc} a & b & c & d \end{array}\\ \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}\end{array}}_{\mathbf{P}^\top} = \begin{array}{c}\begin{array}{cccc} a & b & d & c \end{array}\\ \begin{pmatrix} 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 1/3 & 1/3 & 1/3 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\end{array}$$

- Obervations : Rows and columns sums of $\mathbf{X}$ equal 1 (double stochasticity) and $\mathbf{X}^2 = \mathbf{X}$ (idempotence). Same for $\mathbf{W}$.
- Equivalent Assignment matrix :

$$\mathbf{Y} = \begin{array}{c} \\ a \\ b \\ d \\ c \end{array}\begin{array}{c}\begin{array}{cc} C^1 & C^2 \end{array}\\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\end{array}$$

Prior work

- Notations :
  - $\{\mathbf{x}_i\}_{i=1}^n$, vectors in $\mathbb{R}^p$ we want to cluster.
  - $\phi : \mathbb{R}^p \to \mathbb{F}$, where $\mathbb{F}$ is an RKHS.
  - $\mathbf{Y} \in \{0,1\}^{n \times k}$ the assign. matrix : $y_{ij} = 1$ if $\mathbf{x}_i$ in cluster j ; $y_{ij} = 0$ else.
  - $\mathbf{K} \in \mathbb{R}^{n \times n}$ the kernel matrix : $\mathbf{K}_{ii'} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_{i'}) \rangle_{\mathbb{F}}$.
  - $\mathbf{e}_n$ vector of size n with 1 in all entries.
- The **kernel k-means problem** :

$$\min_{\mathbf{Y} \in \{0,1\}^{n \times k}} \mathrm{SSE}(\mathbf{Y}) = \sum_{j=1}^{k} \sum_{i=1}^{n} \|\phi(\mathbf{x}_i) - \frac{1}{\sum_{i'=1}^{n} y_{i'j}} \sum_{i'=1}^{n} \phi(\mathbf{x}_{i'}) y_{i'j}\|^2 \quad (9)$$

$$\text{s.t. } \mathbf{Y}\mathbf{e}_k = \mathbf{e}_n, \mathbf{Y}^\top \mathbf{e}_n \geq \mathbf{e}_k.$$

- An equivalent **graph based formulation** by [Peng and Xia, 2005] :

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \mathrm{SSE}(\mathbf{X}) = \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \quad (10)$$

$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X} = \mathbf{X}^2, \mathrm{Tr}(\mathbf{X}) = k.$$

Example of relaxation

- Pbs (9) and (10) are NP-hard.
- Key observation : $\mathbf{X} = \mathbf{Y}(\mathbf{Y}^\top\mathbf{Y})^{-1}\mathbf{Y}^\top$.
- $\mathbf{X}$ is an **orthogonal projection matrix (OPM)** and thus :

$$\mathbf{X} \text{ is symmetric : } \mathbf{X}^\top = \mathbf{X} \tag{11}$$

$$\mathbf{X} \text{ is idempotent : } \mathbf{X}^2 = \mathbf{X} \tag{12}$$

$$\mathrm{Rk}(\mathbf{X}) = \mathrm{Tr}(\mathbf{X}) = k \tag{13}$$

$$\mathbf{X} \text{ is positive semi-definite : } \mathbf{X} \succeq \mathbf{0}_n \tag{14}$$

$$\text{if } \mathbf{X} \text{ is the OPM on } \mathbb{E} \text{ then } \mathbf{I}_n - \mathbf{X} \text{ is the OPM on } \mathbb{E}^\perp \tag{15}$$

$$\mathbf{X}(\mathbf{I}_n - \mathbf{X}) = (\mathbf{I}_n - \mathbf{X})\mathbf{X} = \mathbf{0}_n \tag{16}$$

- *SDP relaxation* of Pb (10) :

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) \tag{17}$$

$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{I}_n \succeq \mathbf{X} \succeq \mathbf{0}_n, \mathrm{Tr}(\mathbf{X}) = k.$$

Doubly stochastic and nearly indempotent matrix - DSNI

- We drop the constraint $\mathrm{Tr}(\mathbf{X}) = k$ (unknown nb of clusters).
- But $\mathrm{SSE}(\mathbf{X}) = \mathrm{Tr}(\mathbf{K}(\mathbf{I}_n - \mathbf{X})) = 0$ for $\mathbf{X} = \mathbf{I}_n$.
- We use the **Frobenius dist.** $\|.\|_F$ instead and tackle the NP-hard pb :

$$\min_{\mathbf{X} \in \mathbb{R}^{n \times n}} \|\mathbf{K} - \mathbf{X}\|_F^2 \tag{18}$$
$$\text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n, \mathbf{X}^2 = \mathbf{X}.$$

- Key observations :
  - If $\mathbf{X}\mathbf{e}_n = \mathbf{e}_n$ then the **Laplacian matrix** is $\mathbf{L}_\mathbf{X} = \mathbf{I}_n - \mathbf{X}$.
  - Since $\mathbf{X}$ is an OPM then (16) implies : $\mathbf{X}\mathbf{L}_\mathbf{X} = \mathbf{0}$.
- Moreover, Pb (18) can be reformulated wrt $\mathbf{L}_\mathbf{X}$:

$$\min_{\mathbf{L}_\mathbf{X} \in \mathbb{R}^{n \times n}} \|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_\mathbf{X}\|_F^2 \tag{19}$$
$$\text{s.t. } \mathbf{L}_\mathbf{X} \leq \mathbf{I}_n, \mathbf{L}_\mathbf{X} = \mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n = \mathbf{n}_n, \mathbf{L}_\mathbf{X}^2 = \mathbf{L}_\mathbf{X}.$$

where $\mathbf{n}_n$ vector of size n with 0 in all entries.

Doubly stochastic and nearly indempotent matrix - DSNI

- Let consider the following Pb where $\mathbf{X}$ and $\mathbf{L}$ are *independent*...

$$\min_{\mathbf{X},\mathbf{L}\in\mathbb{R}^{n\times n}} \|\mathbf{K}-\mathbf{X}\|_F^2 + \|\mathbf{I}_n-\mathbf{K}-\mathbf{L}\|_F^2 \tag{20}$$

$$\text{s.t.} \begin{cases} \mathbf{X}\geq\mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n, \mathbf{X}^2=\mathbf{X}, \\ \mathbf{L}\leq\mathbf{I}_n, \mathbf{L}=\mathbf{L}^\top, \mathbf{L}\mathbf{e}_n=\mathbf{n}_n, \mathbf{L}^2=\mathbf{L}. \end{cases}$$

- ...it is more interesting to learn $\mathbf{X}$ and $\mathbf{L}_\mathbf{X}$ jointly ...

$$\min_{\mathbf{X},\mathbf{L}_\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{K}-\mathbf{X}\|_F^2 + \|\mathbf{I}_n-\mathbf{K}-\mathbf{L}_\mathbf{X}\|_F^2 \tag{21}$$

$$\text{s.t.} \begin{cases} \mathbf{X}\geq\mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n, \\ \mathbf{L}_\mathbf{X}\leq\mathbf{I}_n, \mathbf{L}_\mathbf{X}=\mathbf{L}_\mathbf{X}^\top, \mathbf{L}_\mathbf{X}\mathbf{e}_n=\mathbf{n}_n, \\ \mathbf{X}+\mathbf{L}_\mathbf{X}=\mathbf{I}_n, \mathbf{X}\mathbf{L}_\mathbf{X}=\mathbf{0}_n. \end{cases}$$

Doubly stochastic and nearly indempotent matrix - DSNI

- ...because we can provide the following **relaxation** !

$$\min_{\mathbf{X},\mathbf{L_X}\in\mathbb{R}^{n\times n}} \frac{1}{2}\|\mathbf{K}-\mathbf{X}\|_F^2+\frac{1}{2}\|\mathbf{I}_n-\mathbf{K}-\mathbf{L_X}\|_F^2+\frac{\mu}{2}\|\mathbf{XL_X}\|^2{}_F \qquad (22)$$

$$\text{s.t.} \begin{cases} \mathbf{X}\geq\mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n, \\ \mathbf{L_X}\leq\mathbf{I}_n, \mathbf{L_X}=\mathbf{L_X}^\top, \mathbf{L_X}\mathbf{e}_n=\mathbf{n}_n, \\ \mathbf{X}+\mathbf{L_X}=\mathbf{I}_n. \end{cases}$$

- $\mathbf{X}+\mathbf{L_X}=\mathbf{I}_n$ remains a constraint but $\mathbf{XL_X}=\mathbf{0}$ **is penalized**.
- The solution is **Doubly Stochastic and Nearly Idempotent**.
- This pb is biconvex in $\mathbf{X}$ and $\mathbf{L_X}$ and can be handled by the **Alternating Direction Method of Multipliers (ADMM)** (see for eg [Boyd et al., 2011]).

Optimization procedure based on (scaled) ADMM

1. Update $\mathbf{L}_{\mathbf{X}}^{t+1}$ with $\mathbf{X}^t$ fixed :

$$\mathbf{L}_{\mathbf{X}}^{t+1} \leftarrow \underset{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}^t \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\rho}{2}\|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \quad \text{s.t. } \mathbf{L}_{\mathbf{X}} \leq \mathbf{I}_n, \mathbf{L}_{\mathbf{X}} = \mathbf{L}_{\mathbf{X}}^\top, \mathbf{L}_{\mathbf{X}}\mathbf{e}_n = \mathbf{n}_n. \tag{23}$$

2. Update $\mathbf{X}^{t+1}$ with $\mathbf{L}_{\mathbf{X}}^{t+1}$ fixed :

$$\mathbf{X}^{t+1} \leftarrow \underset{\mathbf{X} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 + \frac{\rho}{2}\|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \quad \text{s.t. } \mathbf{X} \geq \mathbf{0}_n, \mathbf{X} = \mathbf{X}^\top, \mathbf{X}\mathbf{e}_n = \mathbf{e}_n. \tag{24}$$

3. Update $\mathbf{U}^{t+1}$:

$$\mathbf{U}^{t+1} \leftarrow \mathbf{U}^t + \mathbf{X}^{t+1} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n \tag{25}$$

4. Repeat 1., 2., 3. until a stopping criterion is satisfied.

Projection on convex sets

- Sub-problems (23) and (24) are convex and can be solved by a **Projection On Convex Sets (POCS)** procedure.
- Notations :
  - $\mathcal{U}_{\mathbf{I}} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} \leq \mathbf{I}_n\}$,
  - $\mathcal{L}_{\mathbf{0}} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} \geq \mathbf{0}_n\}$,
  - $\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X} = \mathbf{X}^\top\}$,
  - $\mathcal{D}_{\mathbf{n}} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X}\mathbf{e}_n = \mathbf{X}^\top \mathbf{e}_n = \mathbf{n}_n\}$,
  - $\mathcal{D}_{\mathbf{e}} = \{\mathbf{X} \in \mathbb{R}^{n \times n} : \mathbf{X}\mathbf{e}_n = \mathbf{X}^\top \mathbf{e}_n = \mathbf{e}_n\}$.
- We can solve Sub-problem (23) as follows :
  1. Solve the unconstrained problem :

$$\widehat{\mathbf{L}_{\mathbf{X}}} \leftarrow \underset{\mathbf{L}_{\mathbf{X}} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{I}_n - \mathbf{K} - \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}^t \mathbf{L}_{\mathbf{X}}\|_F^2 + \frac{\rho}{2}\|\mathbf{X}^t + \mathbf{L}_{\mathbf{X}} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \tag{26}$$

  2. Sequentially project on the convex subsets until convergence :

$$\widehat{\mathbf{L}_{\mathbf{X}}} \leftarrow \Pi_{\mathcal{U}_{\mathbf{I}}}\left(\Pi_{\mathcal{D}_{\mathbf{n}}}\left(\Pi_{\mathcal{S}}\widehat{\mathbf{L}_{\mathbf{X}}}\right)\right) \tag{27}$$

Projection on convex sets

- We can solve Sub-problem (24) as follows :
  1. Solve the unconstrained problem :

$$\widehat{\mathbf{X}} \leftarrow \underset{\mathbf{L_X} \in \mathbb{R}^{n \times n}}{\arg\min} \frac{1}{2}\|\mathbf{K} - \mathbf{X}\|_F^2 + \frac{\mu}{2}\|\mathbf{X}\mathbf{L}_{\mathbf{X}}^{t+1}\|_F^2 + \frac{\rho}{2}\|\mathbf{X} + \mathbf{L}_{\mathbf{X}}^{t+1} - \mathbf{I}_n + \mathbf{U}^t\|_F^2 \tag{28}$$

  2. Sequentially project on the convex subsets until convergence :

$$\widehat{\mathbf{X}} \leftarrow \Pi_{\mathcal{L}_0}\left(\Pi_{\mathcal{D}_\mathbf{n}}\left(\Pi_{\mathcal{S}}\widehat{\mathbf{X}}\right)\right) \tag{29}$$

- Key properties :
  - $\Pi_{\mathcal{D}_\mathbf{n}}, \Pi_{\mathcal{U}_\mathbf{l}}, \Pi_{\mathcal{L}_\mathbf{0}}$ preserve symmetry so after one iteration, $\Pi_{\mathcal{S}}$ needs not to be applied.
  - All projection problems $\Pi_{\mathcal{D}_\mathbf{n}}, \Pi_{\mathcal{U}_\mathbf{l}}, \Pi_{\mathcal{L}_\mathbf{0}}, \Pi_{\mathcal{S}}$ have **closed-form solutions** !

Competing methods and experimental protocol

- Doubly Stochastic Normalization (**DSN**) [Zass and Shashua, 2007] :

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \|\mathbf{K}-\mathbf{X}\|_F^2 \tag{30}$$
$$\text{s.t. } \mathbf{X}\geq \mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n.$$

- Symmetrized Sinkhorn and Knopp algorithm (**SSN**)
  [Zass and Shashua, 2005, Sinkhorn and Knopp, 1967]

$$\min_{\mathbf{X}\in\mathbb{R}^{n\times n}} \mathrm{KL}(\mathbf{X}|\mathbf{K}) = \sum_{i=1}^{n}\sum_{i'=1}^{n} x_{ii'}\log\frac{x_{ii'}}{k_{ii'}} + k_{ii'} - x_{ii'} \tag{31}$$
$$\text{s.t. } \mathbf{X}\geq \mathbf{0}_n, \mathbf{X}=\mathbf{X}^\top, \mathbf{X}\mathbf{e}_n=\mathbf{e}_n.$$

- Test the affinity matrices in the **spectral clustering** framework :
  - Compute the affinity matrix using DSN or SSK or DSNI.
  - Compute the spectral decomposition of the Laplacian matrix.
  - Apply the regular k-means algorithm on the first k eigenvectors.
  - Compare the output with the ground-truth using NMI measure (external validation).

Experimental results

- Experiments setting :
  - No parameter for DSN or SSK.
  - For DSNI by default, $\mu = \sqrt{n}$ and $\rho = 1$.
  - **Gaussian kernel** is used as initial affinity, by default $\sigma^2 = p$.
  - Baseline : Spectral Clustering (SC) with full initial Gaussian kernel.
- Experiments with 5 real-world datasets :

Dataset	n	p	k	SC	SSK	DSN	DSNI
Glass	214	9	6	0.253	0.276	0.243	**0.297**
Ionosphere	351	34	2	0.038	0.066	0.076	**0.131**
Breast cancer	569	30	2	0.010	0.010	0.010	**0.670**
Yeast	1484	8	10	0.070	0.258	0.256	**0.263**
Digits	1797	64	10	0.015	0.044	0.743	**0.767**

TABLE – Datasets statistics and NMI measures.

- DSNI which promotes **idempotency** performs better.

Future work

- Some ideas for future work :
  - Sensitivity analysis w.r.t. the hyper-parameter ρ.
  - The properties of the DSNI matrix : it is sparse, it seems to capture the intrinsic geoemtry of the data.
  - Benchmarks of a couping of DSNI and SNK-AHC.

Thank you for your attention ! Any questions ?

References I

Boyd, S., Parikh, N., and Chu, E. (2011).
Distributed optimization and statistical learning via the alternating direction method of multipliers.
Now Publishers Inc.

Peng, J. and Xia, Y. (2005).
A new theoretical framework for k-means-type clustering.
In Foundations and advances in data mining, pages 79–96. Springer.

Sinkhorn, R. (1968).
Two results concerning doubly stochastic matrices.
The American Mathematical Monthly, 75(6) :632–634.

Sinkhorn, R. and Knopp, P. (1967).
Concerning nonnegative matrices and doubly stochastic matrices.
Pacific Journal of Mathematics, 21(2) :343–348.

Zass, R. and Shashua, A. (2005).
A unifying approach to hard and probabilistic clustering.
In Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, volume 1, pages 294–301. IEEE.

Zass, R. and Shashua, A. (2007).
Doubly stochastic normalization for spectral clustering.
In Advances in neural information processing systems, pages 1569–1576.

Conditions suffisantes pour la monotonicité

- Pour D-AHC un dendrogramme est monotone si :

$$\mathbf{D}^{t+1}_{(kl)m} \geq \mathbf{D}_{kl}, \forall t, \forall k, l, m \in \mathbb{C}^t$$

- Un dendrogramme monotone est préférable en termes d'interprétations. Group average, Mcquitty et Ward sont guaranties de produire des dendogrammes monotones.
- Pour K-AHC, on dira qu'un dendrogramme est monotone si :

$$\mathfrak{p}((kl), m)\left(\mathbf{S}^{t+1}_{(kl)m} - \frac{1}{2}\left(\mathbf{S}^{t+1}_{(kl)(kl)} + \mathbf{S}^{t+1}_{mm}\right)\right)$$
$$\leq \mathfrak{p}(k, l)\left(\mathbf{S}^{t}_{kl} - \frac{1}{2}\left(\mathbf{S}^{t}_{kk} + \mathbf{S}^{t}_{ll}\right)\right), \forall t, \forall k, l, m \in \mathbb{C}^t$$

Conditions suffisantes pour la monotonicité (suite)

Proposition

Si les fonctions d'ensemble $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}$ *et* $\mathfrak{p}$ *satisfont aux conditions suivantes :*

$$\begin{cases} \mathfrak{a}(k,l), \mathfrak{b}(k,l), \mathfrak{c}(k,l), \mathfrak{p}(k,l) \geq 0 \\ \mathfrak{a}(k,l) + \mathfrak{a}(l,k) = 1 \\ \mathfrak{b}(k,l) - \mathfrak{b}(l,k) = 0 \\ \mathfrak{c}(k,l) - \mathfrak{a}(k,l) + \frac{1}{2}\mathfrak{b}(k,l) = 0 \\ \frac{\mathfrak{a}(k,l)}{\mathfrak{p}(k,m)} + \frac{\mathfrak{a}(l,k)}{\mathfrak{p}(l,m)} - \frac{\mathfrak{b}(k,l)}{2\mathfrak{p}(k,l)} \geq 0 \\ \mathfrak{p}((kl),m)\left(\frac{\mathfrak{a}(k,l)}{\mathfrak{p}(k,m)} + \frac{\mathfrak{a}(l,k)}{\mathfrak{p}(l,m)} - \frac{\mathfrak{b}(k,l)}{2\mathfrak{p}(k,l)}\right) \geq 1 \end{cases} \qquad \forall k, l, m \in 2^{\mathbb{O}}$$

alors sous l'hypothèse C1, K-AHC produit des dendrogrammes monotones.

- On peut vérifier que les 3 méthodes citées précédemment satisfont à ces conditions.
- Ce résulat est valable pour SNK-AHC également.

Approche "stored data"

- D-AHC et la formule de LW suppose une matrice de dissimilarités mais certaines approches peuvent utiliser la représentation vectorielle ("fetaure matrix"). En effet, du fait de leur interprétation géométrique, centroid, Ward et median utilisent implicitement des vecteurs représentants de cluster (barycentre ou isobarycentre -mid-point-) et les dissimilarités entre clusters sont en fait des distances entre représentants de clusters.
- Dans le cas de K-AHC, nous pouvons exprimer toutes les méthodes à l'étude selon une approche vectorielle.

Approche "stored data" (suite)

Method	$\mathbf{S}_{ij}^t, i \neq j$	$\mathbf{S}_{ii}^t$	$\mathfrak{p}(i,j)$	$\mathbf{g}^{(kl)}$	$\mathbf{s}_{(kl)}^{t+1}$
Group average	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\mathbf{s}_i^t$	1	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^k + \frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^l$	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{s}_k^t + \frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{s}_l^t$
Mcquitty	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\mathbf{s}_i^t$	1	$\frac{1}{2}\mathbf{g}^k + \frac{1}{2}\mathbf{g}^l$	$\frac{1}{2}\mathbf{s}_k^t + \frac{1}{2}\mathbf{s}_l^t$
Centroid	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\langle \mathbf{g}^i, \mathbf{g}^i \rangle$	1	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^k + \frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^l$	NA
Median	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\langle \mathbf{g}^i, \mathbf{g}^i \rangle$	1	$\frac{1}{2}\mathbf{g}^k + \frac{1}{2}\mathbf{g}^l$	NA
Ward	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\langle \mathbf{g}^i, \mathbf{g}^i \rangle$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$	$\frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^k + \frac{\lvert k\rvert}{\lvert k\rvert+\lvert l\rvert}\mathbf{g}^l$	NA
W-Median	$\langle \mathbf{g}^i, \mathbf{g}^j \rangle$	$\langle \mathbf{g}^i, \mathbf{g}^i \rangle$	$\frac{\lvert i\rvert\lvert j\rvert}{\lvert i\rvert+\lvert j\rvert}$	$\frac{1}{2}\mathbf{g}^k + \frac{1}{2}\mathbf{g}^l$	NA

TABLE – Particular settings in the stored data matrix based on K-AHC and defined by (3) and representative vectors updates.

SNK-AHC

- On introduit **S**parse **N**ormalized **K**ernels based **A**gglo. **H**ier. **C**lust.

Input: **S** (kernel matrix), sparsification method, AHC method
Output: D a dendrogram

1 **if** the diagonal of **S** is not constant **then**
2 | Normalize **S** using cosine normalization;
3 **end**
4 Translate **S** in order to have non negative values;
5 Sparsify **S** in order to have a sparse **S**;
6 Initialize D with n leaves;
7 Set $\mathbf{S}^1 = \mathbf{S}$;
8 Determine $\mathbb{S}^1$ according to (6);
9 **while** $\mathbb{S}^t \neq \emptyset$ **do**
10 | Find the pair of clusters (k, l) according to (**??**) ;
11 | Merge (k, l) into (kl) and update D;
12 | Update $\mathbb{S}^{t+1}$ from $\mathbb{S}^t$;
13 | Compute $\mathbf{S}^{t+1}$ by applying (4) and (5).
14 **end**

Complexité

Proposition

Soit **S** *la matrice de similarités sparse obtenue après l'étape 5 de l'algorithme SNK-AHC. Soit* z *le nombre d'entrées non-nulles de* **S**. *La construction du dendrogramme donnée par les étapes 6 à 14 de l'algorithme SNK-AHC, a une complexité en mémoire en* $O(z)$ *et une complexité en temps de traitement en* $O(nz)$.

Invariance vis à vis de translations de la diagonale

- Après sparsification **S** n'est plus sdp.
- On peut **augmenter sa diagonale** pour la rendre sdp à nouveau mais ceci aboutit à une "distorsion" de l'espace initial.
- Certaines méthodes sont **invariantes vis à vis de la translation de la diagonale** !

Proposition

Soit $\mathbf{S}$ *la matrice de similarités sparse obtenue après l'étape 5 de l'algorithme SNK-AHC. Pour les méthodes group average, Mcquitty et Ward, l'algorithme SNK-AHC produit deux dendrogrammes equivalents si l'on prend comme matrices sparses* $\mathbf{S}$ *et* $\mathbf{T} = \mathbf{S} + w\mathbf{I}_n$, $w \in \mathbb{R}$.

- Donc pour ces trois techniques, les hypothèses géométriques C1 et C2 restent valides même si **S** n'est pas sdp.

Composantes connexes et nombre de clusters

- Nous interprétons **S** telle la matrice d'adjacence pondérée d'un graphe non orienté sur $\mathbb{O}$. Si **S** est sparse, le graphe n'est pas complet et peut contenir **plusieurs composantes connexes**.
- En fait, les étapes de fusion de SNK-AHC sont identiques à celles d'un algorithme permettant de détecter les **sous-ensembles disjoints** des sommets d'un graphe non connexe (s'il existe un chemin rejoignant deux sommets alors ils sont mis dans le même sous-ensemble).

Proposition

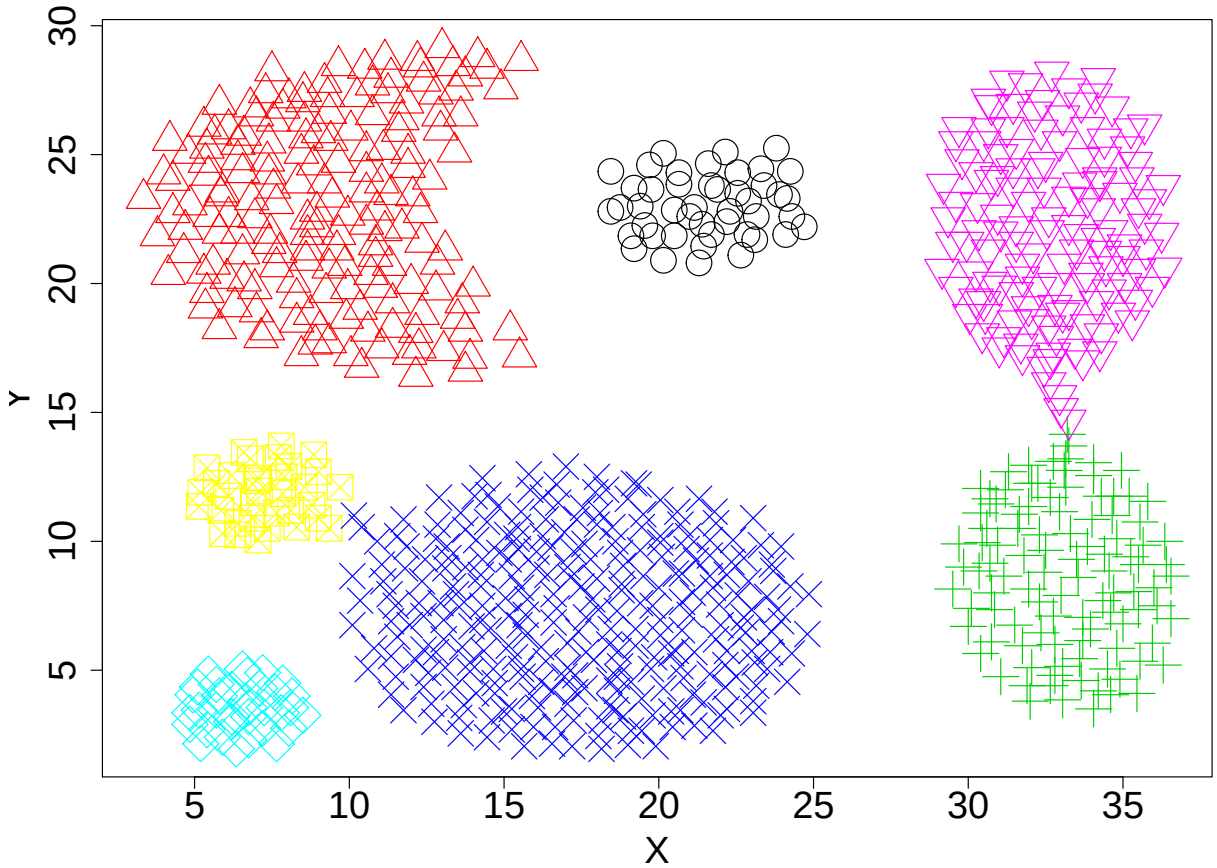
Soit **S** *la matrice de similarités sparse obtenue après l'étape 5 et* $\mathbb{S} = \mathbb{S}^1$ *le sous-ensemble de couples d'objets obtenu après l'étape 8 de l'algo. SNK-AHC. Soit* $G = (\mathbb{O}, \mathbb{S})$ *le graphe non orienté sur* $\mathbb{O}$ *et de matrice d'adjacence* $\mathbb{S}$*. Si* G *possède* κ *composantes connexes alors l'algo. SNK-AHC s'arrête à l'iteration* $n - \kappa - 1$ *et donne comme résultat une fôret d'arbres binaires où chaque arbre est une composante connexe.*

Protocol expérimental

- Illustrations des propriétés sur deux jeux de données artifielles ("aggregation" et "compound") et deux jeux de données réelles ("landsat", "pendigits").
- **Plusieurs niveaux de sparsification** utilisant soit le seuil θ soit les k plus proches voisins.
- **Coefficient cophénétique (CC)** pour mesurer la similarité entre le dendrogramme de la baseline et ceux donnés par SNK-AHC. Le résultat de référence est celui obtenu avec la matrice complète **S** (qui est donc équivalent à l'approche classique D-AHC).
- L'**indice de Rand corrigé (ARI)** pour mesurer la qualité du résultat de clustering vis à vis de la vérité terrain (on coupe le dendrogramme au nombre correct de clusters).
- Les **diminutions** de stockage **mémoire** et de **temps de traitement** lorsque **S** est de plus en plus sparse, sont mesurées relativement aux performances obtenues avec la matrice complète **S**.

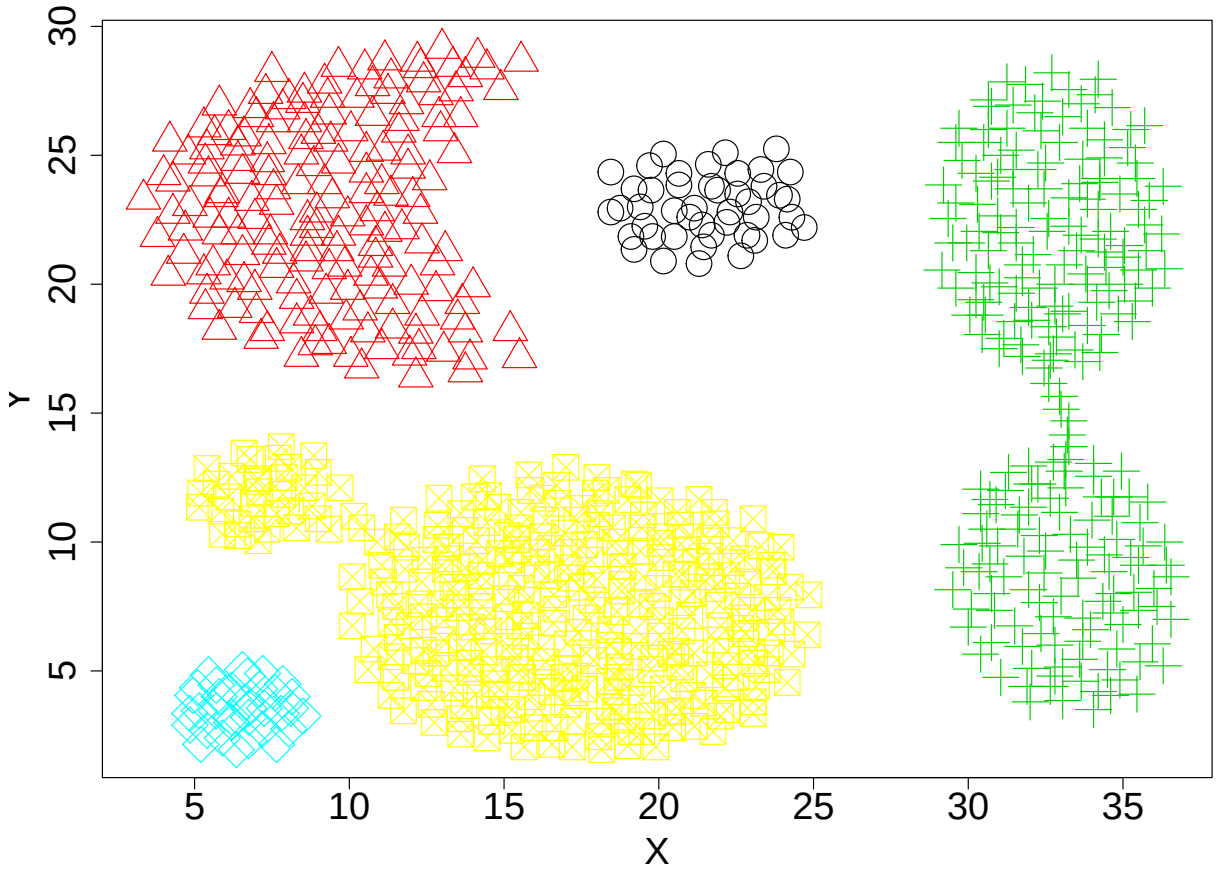
Données "aggregation"

- 788 observations en 2D.
- 7 clusters.



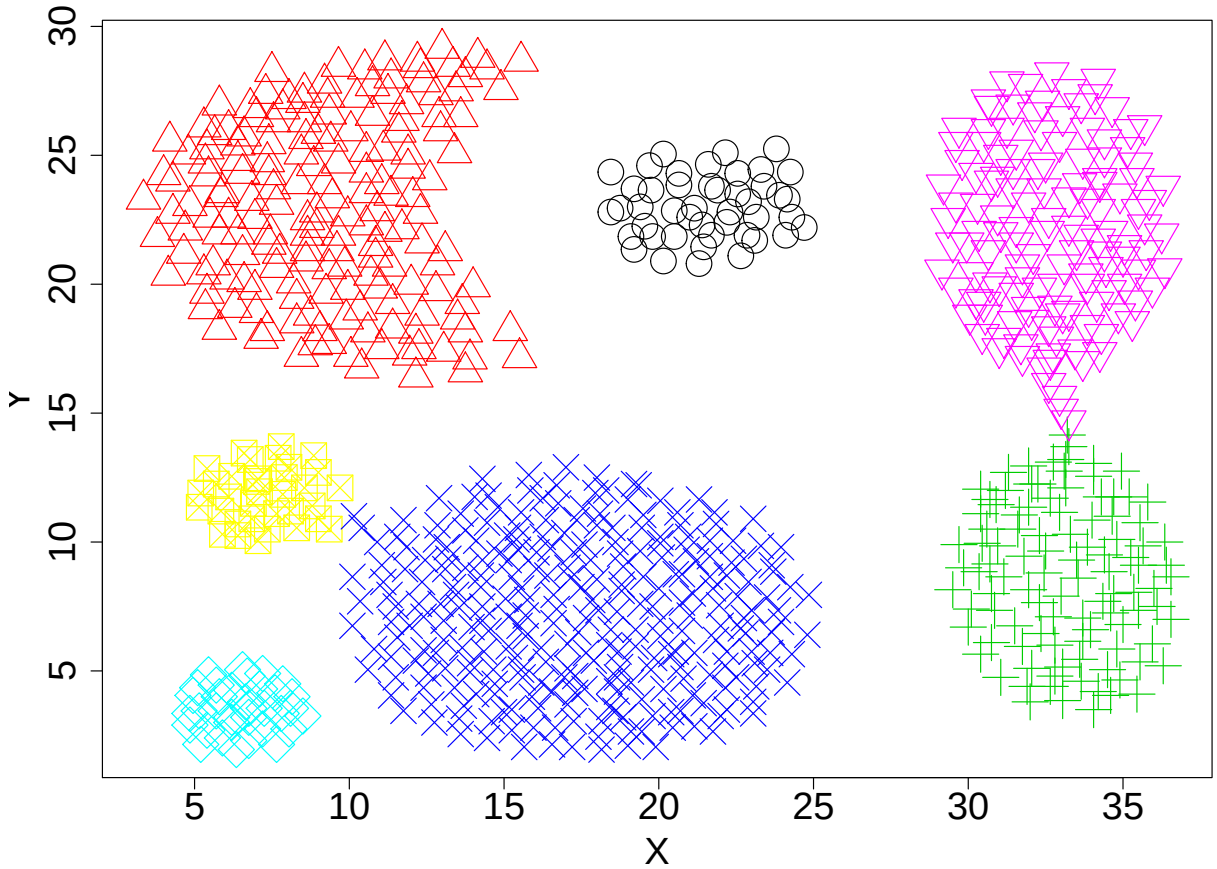
Résultats sur les données "aggregation"

- Résultats avec noyau Gaussien, $k = 8$ ($\sim$10% de proches voisins).
- 5 composantes connexes sont détectées automatiquement.



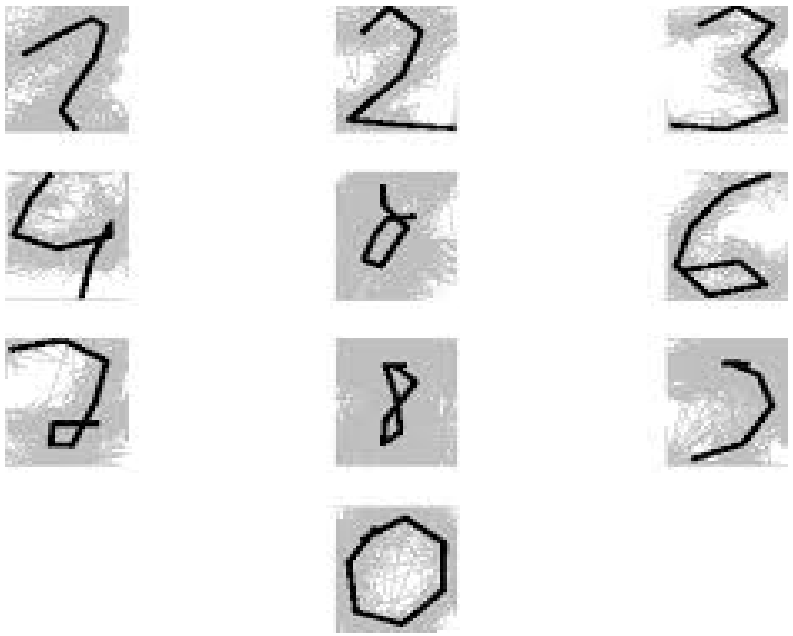
Résultats sur les données "aggregation"

- Résultats avec noyau Gaussien, $k = 8$ ($\sim$10% de proches voisins).
- Si on coupe à 7 clusters on obtient la solution exacte.



Données "pendigits"

- Reconnaissance de chiffres écrits à la main par 44 personnes différentes. Jeu de données disponible sur UCI [5].
- 10,992 observations (1 obs = 1 chiffre = $\{x_t, y_t\}_{t=1,\ldots,8}$)
- 16 variables (chaque pixel a 4 valeurs de spectres)
- 10 clusters : 0,. . . ,9.



5. https://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits

Résultats sur les données "pendigits"

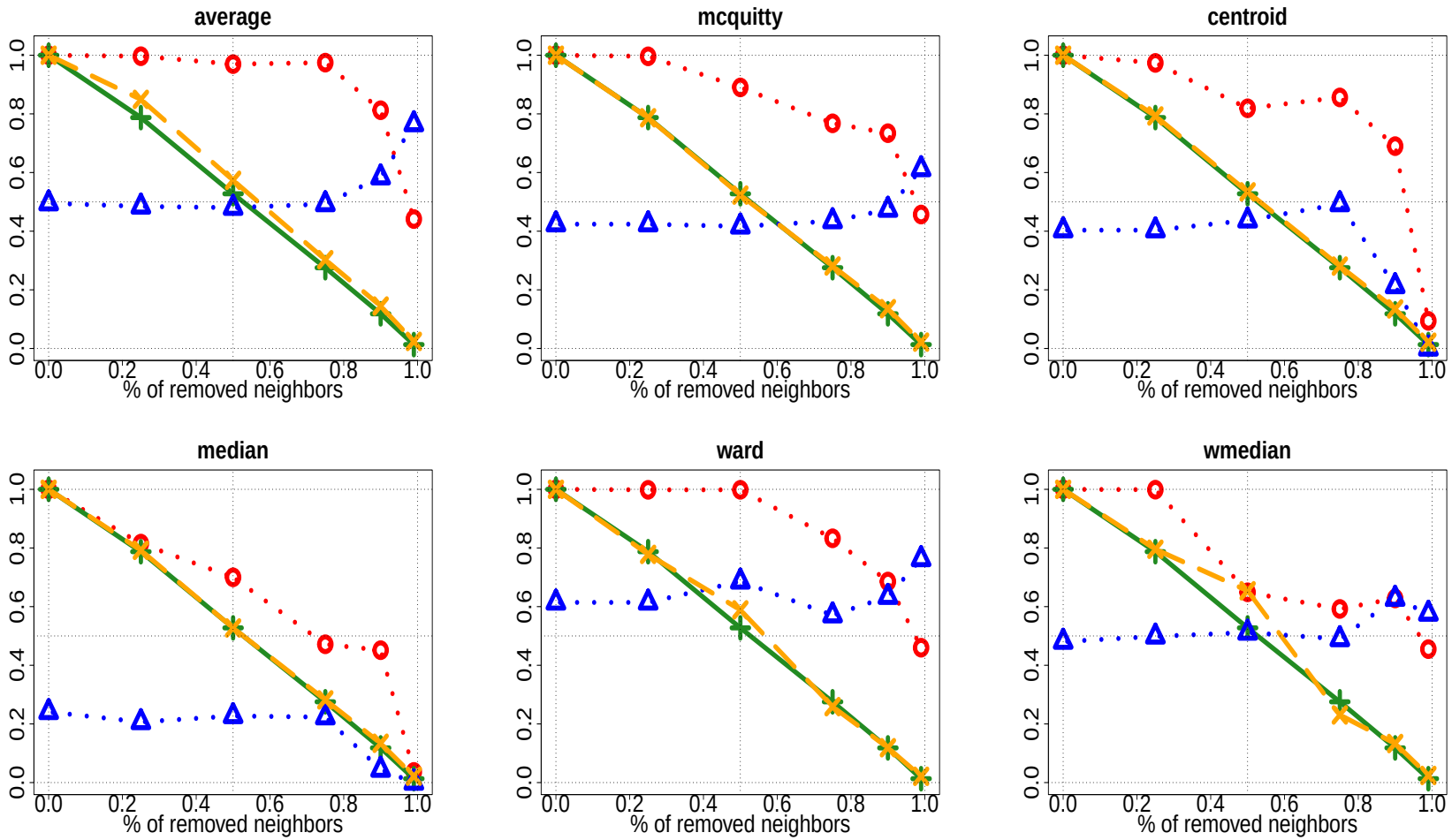


FIGURE – Solid lines with plus signs represent the relative memory use, dashed lines with cross signs show the relative running time, dotted lines with circles indicate the CC values, dotted lines with triangles give the ARI values.