Clustering evidentiel : intérêt, application et variantes

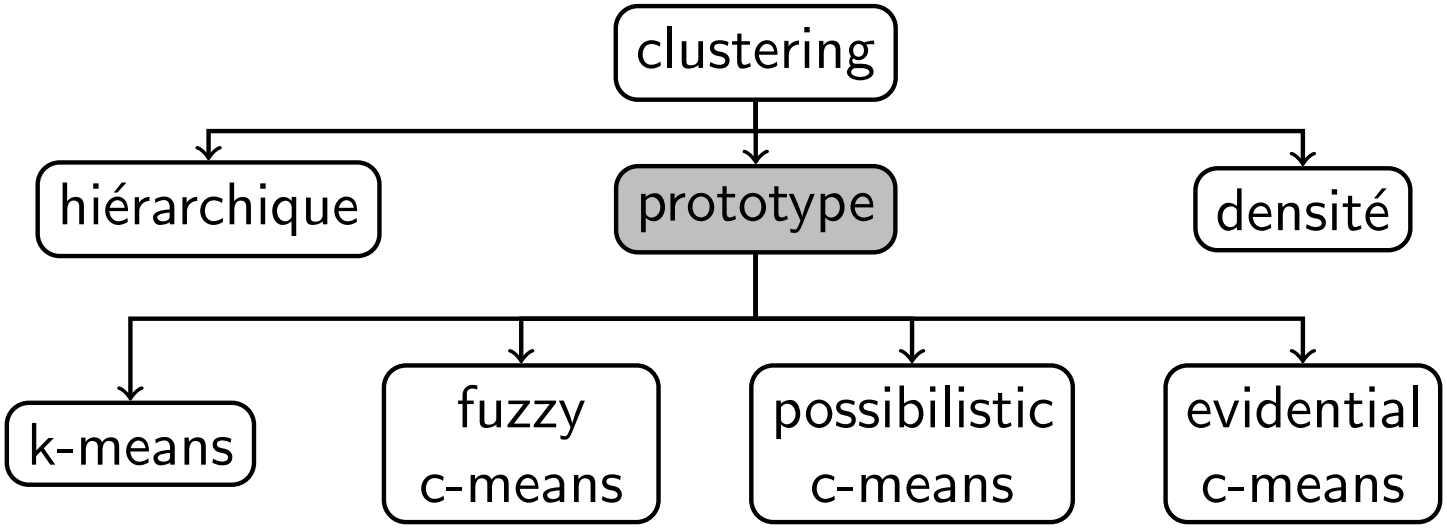
V. Antoine

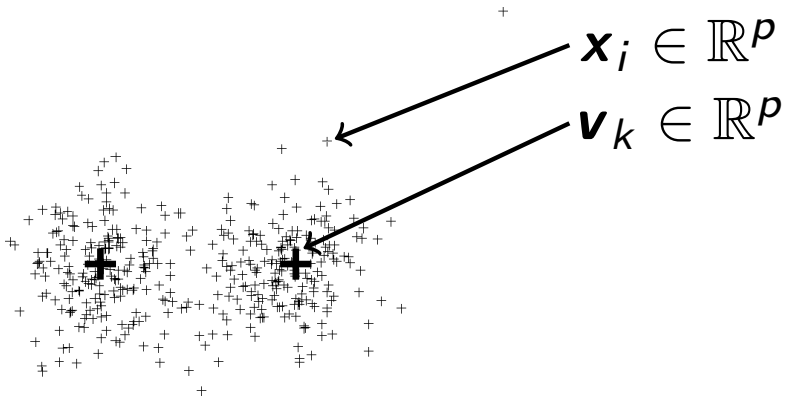
Université Clermont Auvergne, LIMOS, UMR CNRS 6158, France
https://perso.isima.fr/~viantoin

Avril 2024

Clustering

Détermine des groupes d'objets selon une notion de similarité





Clustering basé sur les prototypes

Avantages

- Complexité réduite
- Interprétable

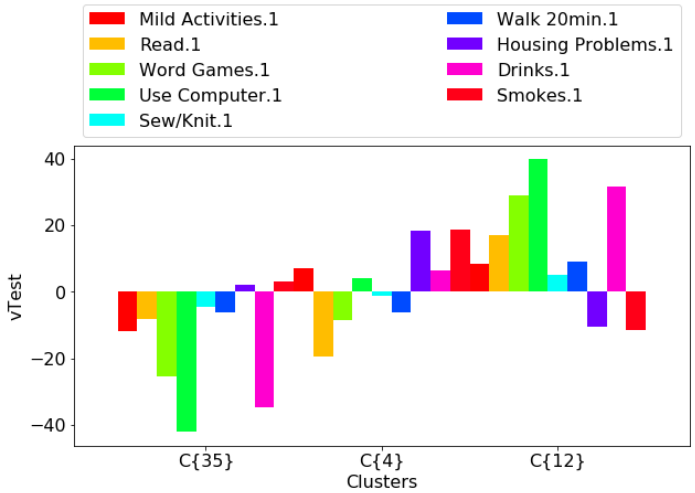
Clustering basé sur les prototypes

Avantages

- Complexité réduite
- Interprétable

Applications

- Vulcanologie
- Biologie
- Santé
- ...



Clustering basé sur les prototypes

Avantages

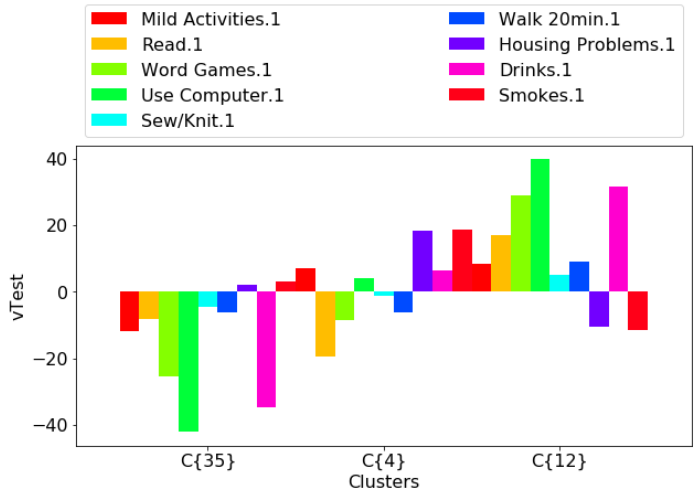
- Complexité réduite
- Interprétable

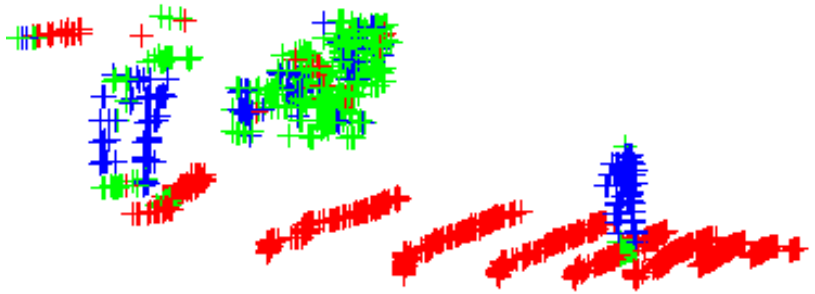
Applications

- Vulcanologie
- Biologie
- Santé
- ...

Bloc dans du ML

- Compression
- Augmentation automatique d'étiquettes





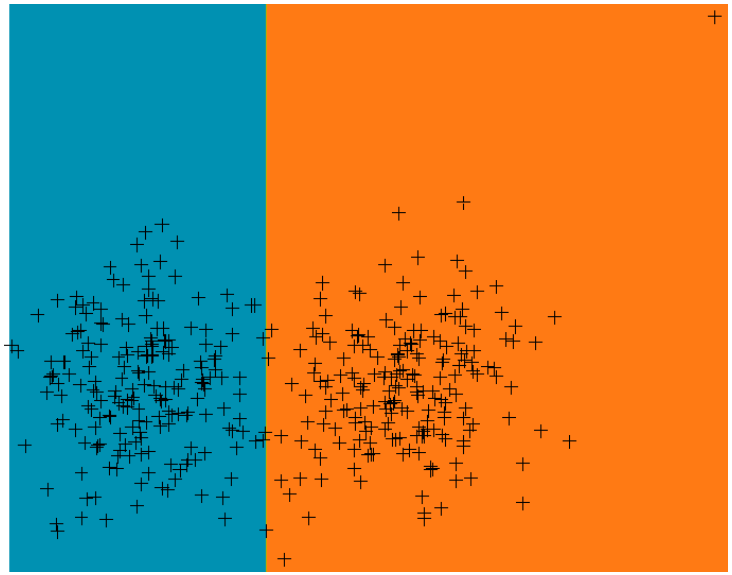
K-means : clustering à partition dure

- Chaque objet est assigné à un et un seul cluster
- $\mathbf{P} = (p_{ik})$ s.t $p_{ik} \in \{0, 1\}, \sum_{k=1}^{c} p_{ik} = 1$

Exemple

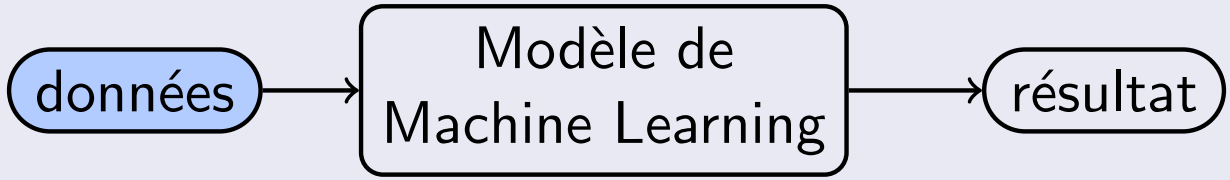
- ω_1 la classe des carrés
- ω_2 la classes des cercles

	p_{i1}	p_{i2}
○	0	1
□	1	0
□	1	0



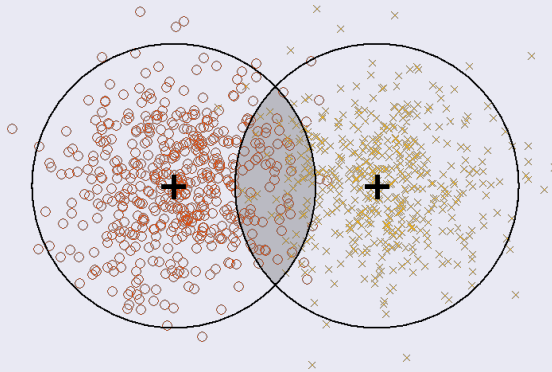
Soft clustering

Modélisation de l'incertitude de la décision



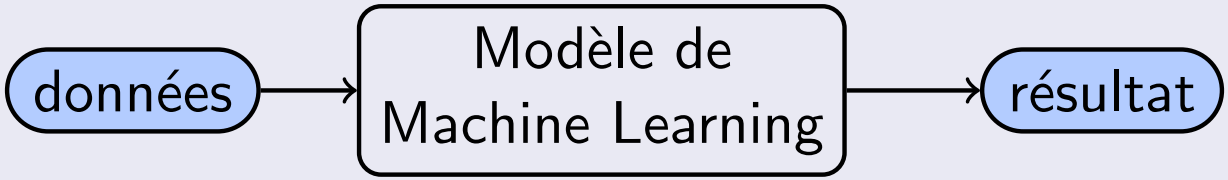
Cause de l'incertitude

- incertitude aléatoire
  - Variabilité naturelle due à des phénomènes aléatoires
  - Irréductible
- incertitude épistémique
  - Incertitude due à un manque de données ou de connaissance



Soft clustering

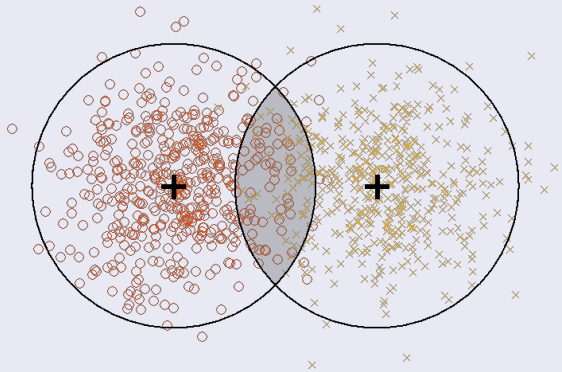
Modélisation de l'incertitude de la décision



Si les données sont incertaines, le résultat sera incertain !

Cause de l'incertitude

- incertitude aléatoire
  - Variabilité naturelle due à des phénomènes aléatoires
  - Irréductible
- incertitude épistémique
  - Incertitude due à un manque de données ou de connaissance



Soft clustering

Variantes de k-means

- Théorie des ensembles flous : FCM
- Théorie des possibilité : PCM
- Théorie des ensembles approximatifs : RKM
- Théorie des fonctions de croyance : ECM
- ...

Notations

- $\boldsymbol{X} = (\boldsymbol{x}_i) \in \mathbb{R}^{n \times p}$ l'ensemble des objets
- $\boldsymbol{V} = (\boldsymbol{v}_k) \in \mathbb{R}^{c \times p}$ l'ensemble des centres associés aux classes
- $\Omega = \{\omega_1, \ldots, \omega_c\}$ l'ensemble des classes

Plan

Plan

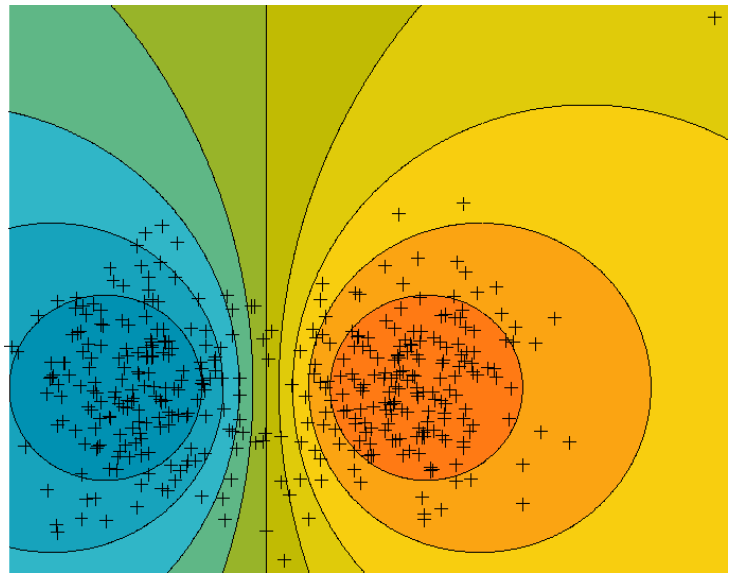
Partition floue

- Chaque objet a un degré d'appartenance à chaque cluster
- $\mathbf{U} = (u_{ik})$ tel que $u_{ik} \in [0, 1]$, $\sum_{k=1}^{c} u_{ik} = 1$

Exemple

- ω_1 la classe des carrés
- ω_2 la classe des cercles

	p_{i1}	p_{i2}
○	0	1
□	1	0
□	0.9	0.1
□	0.5	0.5



Fuzzy c-means (FCM)

Modèle géométrique

- Chaque cluster ω_k est représenté par un centre $\mathbf{v}_k$
- Distance Euclidienne $d_{ik}^2 = (\mathbf{x}_i - \mathbf{v}_k)^T(\mathbf{x}_i - \mathbf{v}_k)$

Fonction objectif

$$J_{FCM}(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{n}\sum_{k=1}^{c} u_{ik}^{\beta} d_{ik}^2$$

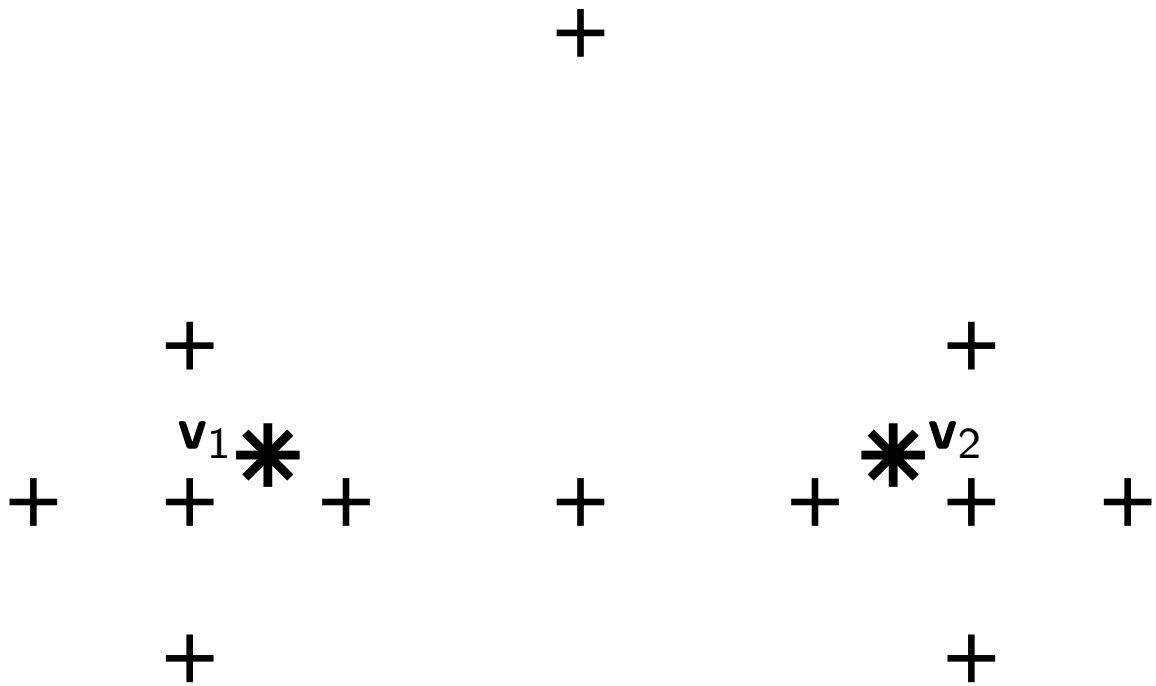
tel que

$$\sum_{k=1}^{c} u_{ik} = 1 \text{ et } u_{ik} \geq 0 \quad \forall i, k$$

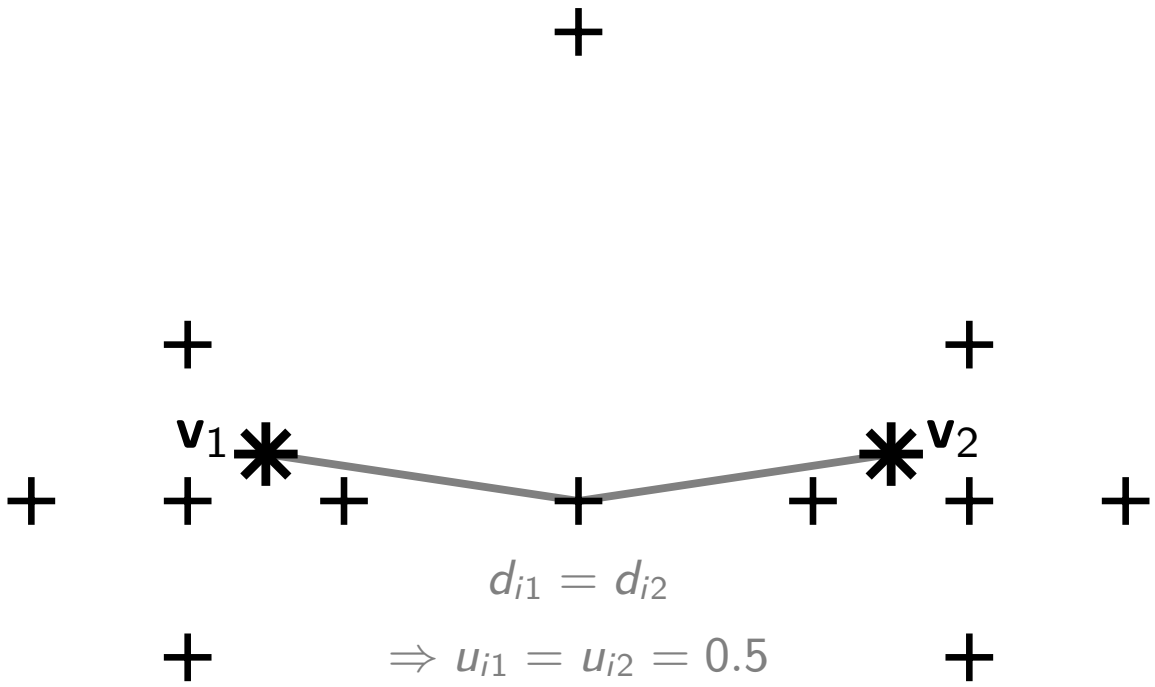
Méthode d'optimisation de type Gauss-Seidel

$$\min_{\mathbf{U}} J_{FCM} \quad \rightarrow \quad \min_{\mathbf{V}} J_{FCM} \quad \rightarrow \quad \ldots$$

Problématique : affectations imprécises et objets atypiques



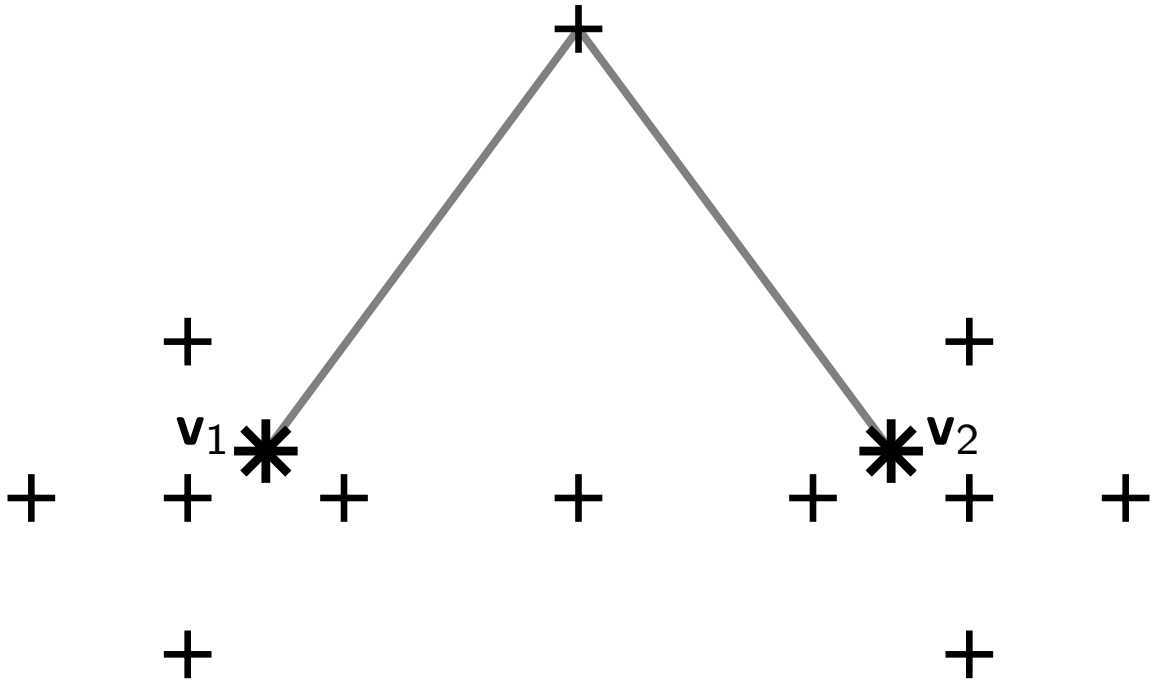
Problématique : affectations imprécises et objets atypiques



Problématique : affectations imprécises et objets atypiques

$$d_{i1} = d_{i2}$$

$$\Rightarrow u_{i1} = u_{i2} = 0.5$$



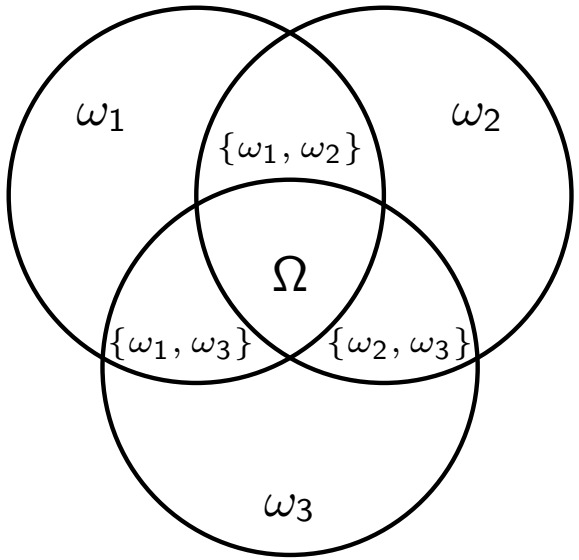
Théorie des fonctions de croyance

Soit Y une variable prenant des valeurs dans l'ensemble fini Ω

Fonction de masse $m : 2^{\Omega} \rightarrow [0, 1]$

$$\sum_{A \subseteq \Omega} m(A) = 1$$

- $m(A)$: degré de croyances spécifique à $Y \in A$
- Si $m(A) > 0$ alors A est un ensemble focal



$\emptyset$

Notions dérivées

Fonction de croyance

Total soutien donné à A :

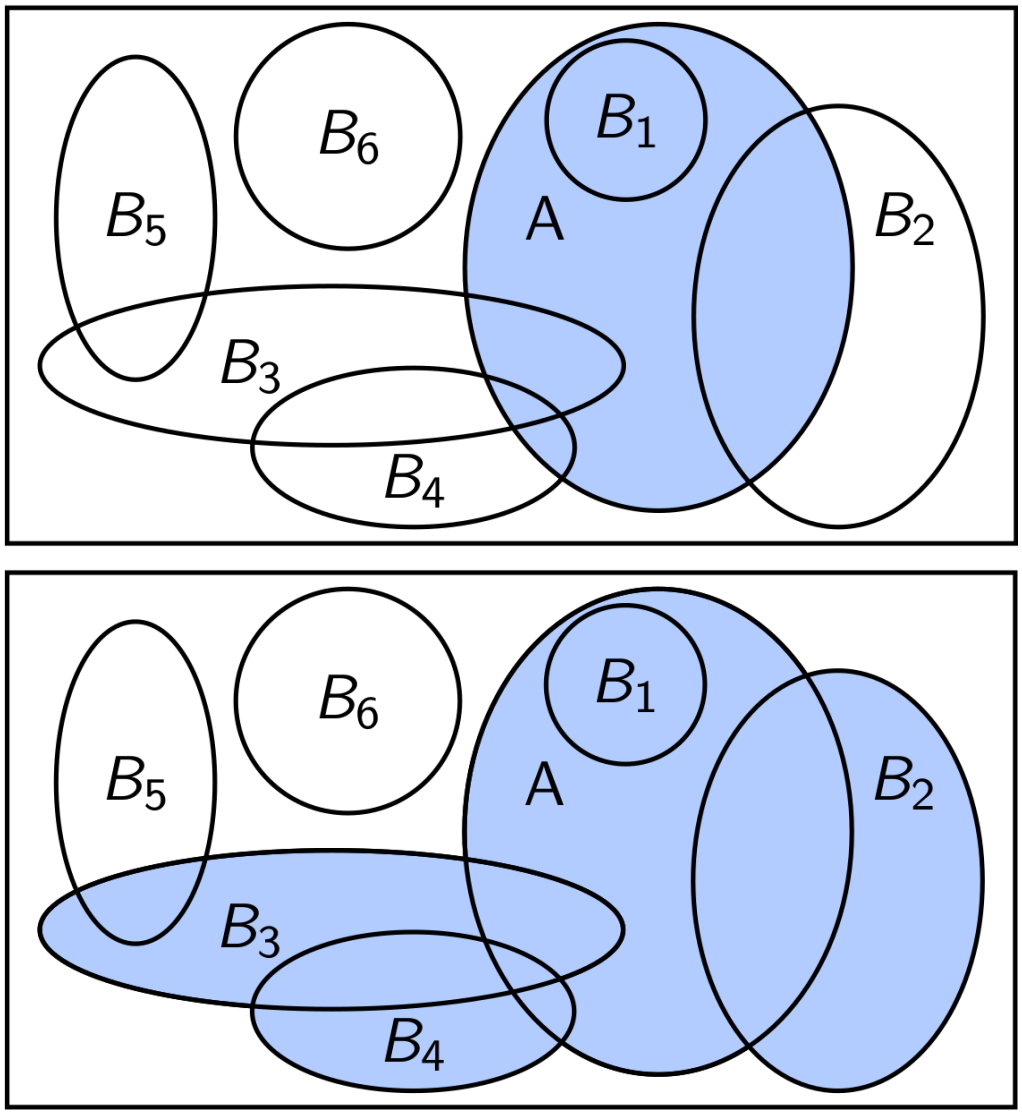
$$bel(A) = \sum_{B \subseteq A} m(B),$$

Fonction de plausibilité

Degré de croyance *potentiel* qui *peut être* donné à A :

$$pl(A) = \sum_{B \cap A \neq \emptyset} m(B),$$

$$\forall A \subseteq \Omega, A \neq \emptyset$$

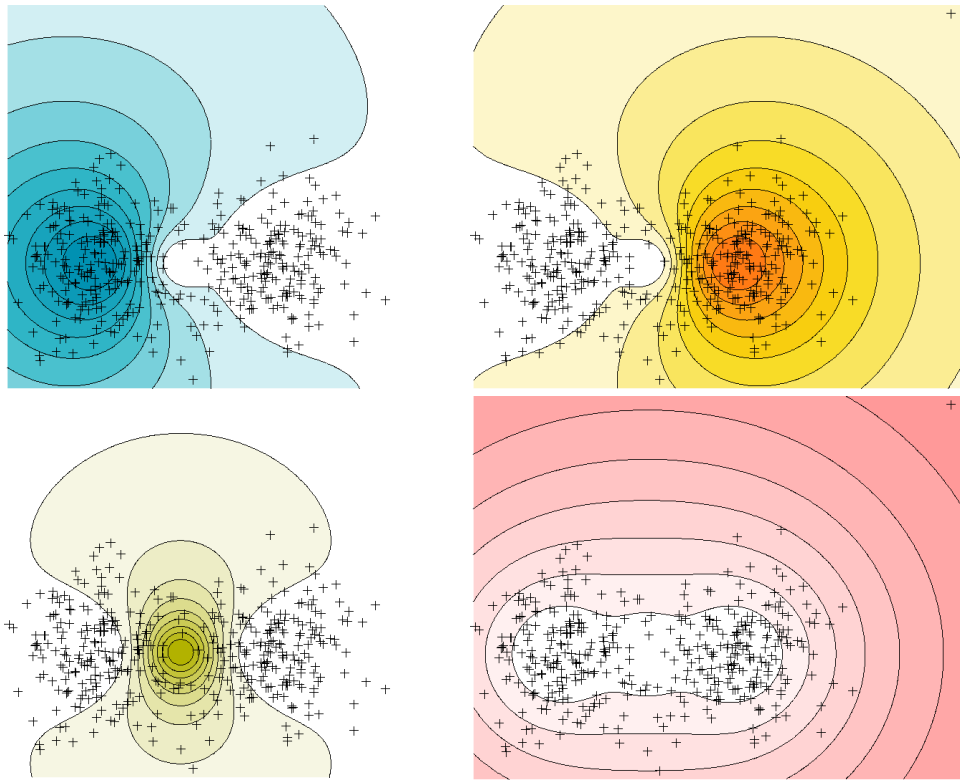


Partition crédale

- Chaque objet a un degré de croyance pour chaque sous-ensemble $A_j \subseteq \Omega$
- $\mathbf{M} = (m_{ij})$ tel que $m_{ij} \in [0, 1]$, $\sum_{A_j \subseteq \Omega} m_{ij} = 1$

Exemple

	$m_{i\emptyset}$	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\Omega}$
○	0	0	1	0
□	0	1	0	0
□	0	0.9	0.1	0
□	0	0	0	1
☆	1	0	0	0



Transformation crédale

Transformation pignistique pour une prise de décision

$$BetP(\omega) = \frac{1}{1 - m(\emptyset)} \sum_{\{A \subseteq \Omega | \omega \in A\}} \frac{m(A)}{|A|}$$

Partition crédale

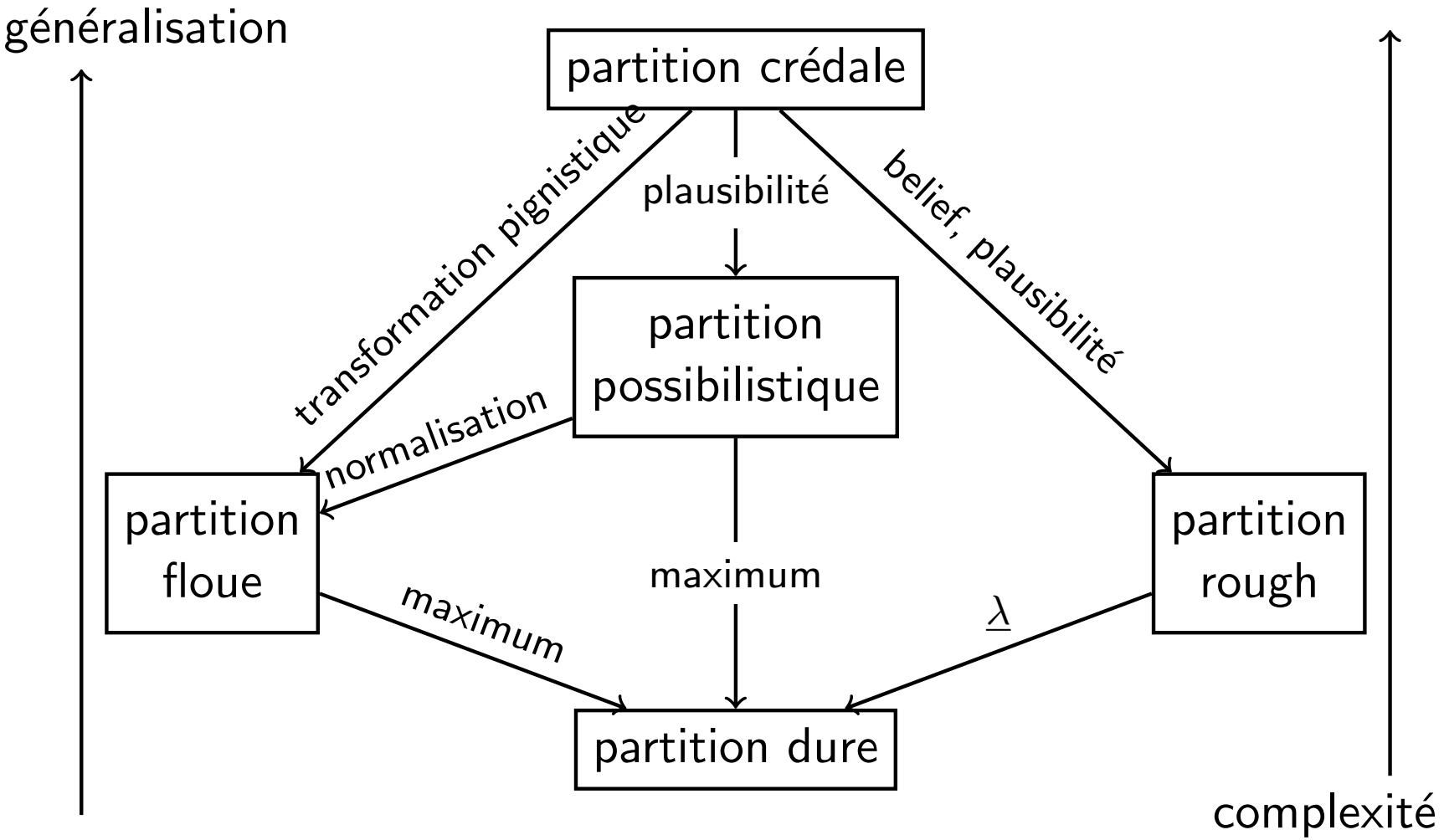
	$m_{i\emptyset}$	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\Omega}$
	0	0	1	0
	0	1	0	0
	0	0.9	0.1	0
	0	0	0	1
	1	0	0	0

transformation pignistique →

Partition floue

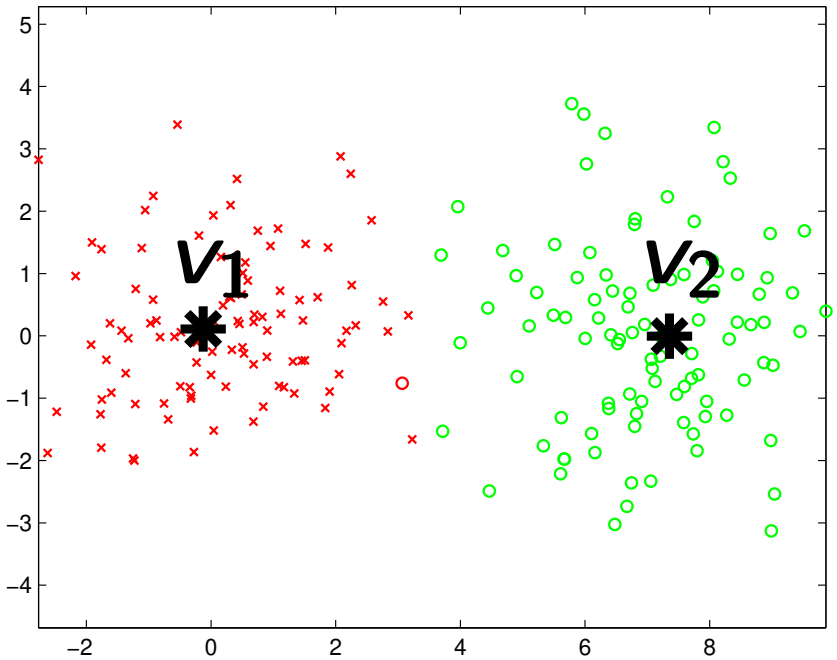
	$u_{i\omega_1}$	$u_{i\omega_2}$
	0	1
	1	0
	0.9	0.1
	0.5	0.5
	0.5	0.5

Transformation crédale



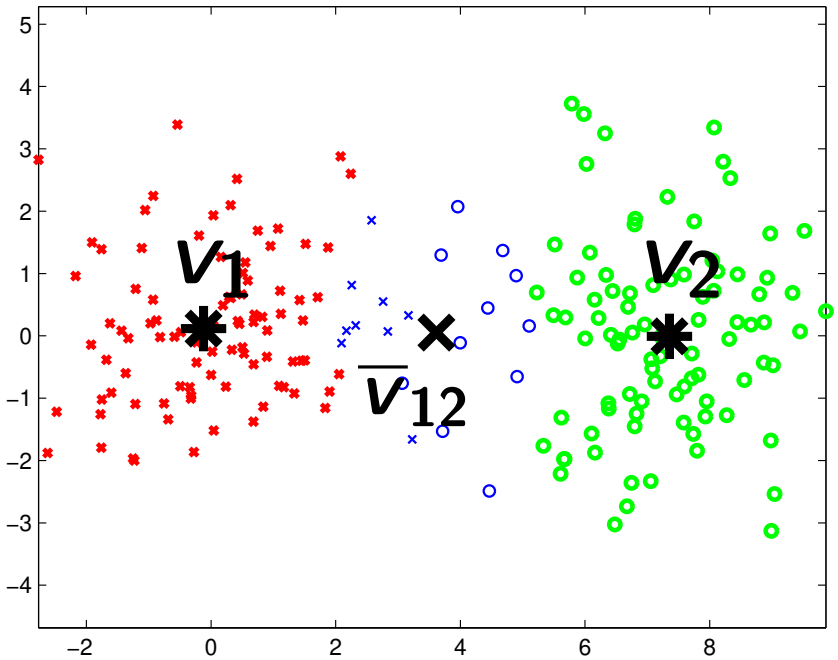
Evidential c-means (ECM)

- Chaque cluster ω_k est representé par un centre $\mathbf{v}_k$
- Centre $\bar{\mathbf{v}}_j$: barycentre des centres associés aux classes composant $A_j \subseteq \Omega$
- Distance d_{ij}^2 entre $\mathbf{x}_i$ et $\bar{\mathbf{v}}_j$



Evidential c-means (ECM)

- Chaque cluster ω_k est representé par un centre $\mathbf{v}_k$
- Centre $\overline{\mathbf{v}}_j$: barycentre des centres associés aux classes composant $A_j \subseteq \Omega$
- Distance d_{ij}^2 entre $\mathbf{x}_i$ et $\overline{\mathbf{v}}_j$



Evidential c-means (ECM)

Fonction objectif

$$J_{ECM}(\mathbf{M},\mathbf{V}) = \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega,\ A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} d_{ij}^2 + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}$$

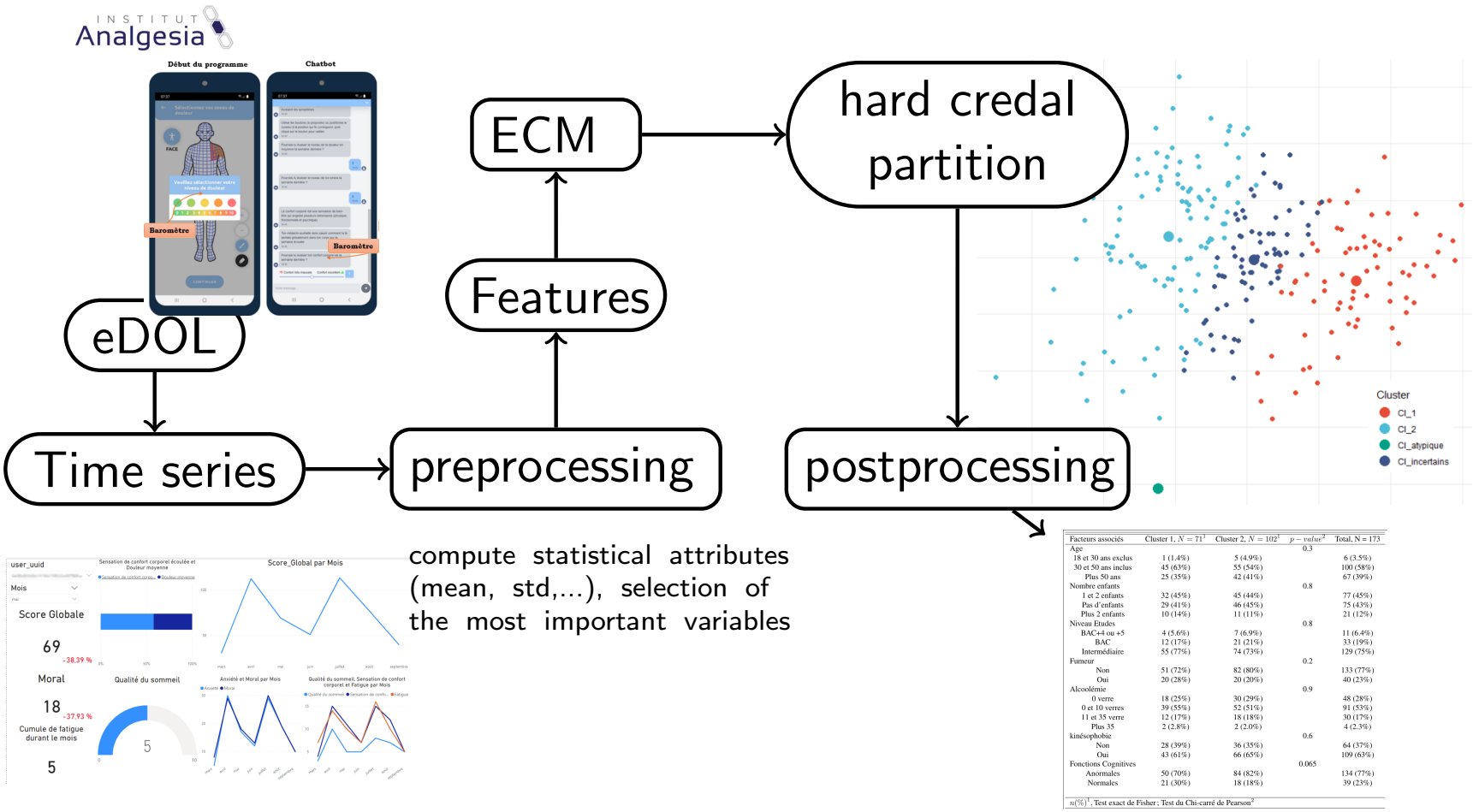
Tel que

$$\sum_{A_j \subseteq \Omega,\ A_j \neq \emptyset} m_{ij} + m_i(\emptyset) = 1,\ m_i(A_j) \geq 0 \quad \forall i,j$$

Méthode d'optimisation de type Gauss-Seidel

$$\text{opt}(\mathbf{M}) \rightarrow \text{opt}(\mathbf{V}) \rightarrow \ldots$$

Application de ECM pour la santé



compute statistical attributes (mean, std,...), selection of the most important variables

Facteurs associés	Cluster 1, $N = 71^1$	Cluster 2, $N = 102^1$	$p-value^2$	Total, N = 173
Age			0.3	
18 et 30 ans exclus	1 (1.4%)	5 (4.9%)		6 (3.5%)
30 et 50 ans inclus	45 (63%)	55 (54%)		100 (58%)
Plus 50 ans	25 (35%)	42 (41%)		67 (39%)
Nombre enfants			0.8	
1 et 2 enfants	32 (45%)	45 (44%)		77 (45%)
Pas d'enfants	29 (41%)	46 (45%)		75 (43%)
Plus 2 enfants	10 (14%)	11 (11%)		21 (12%)
Niveau Etudes			0.8	
BAC+4 ou +5	4 (5.6%)	7 (6.9%)		11 (6.4%)
BAC	12 (17%)	21 (21%)		33 (19%)
Intermédiaire	55 (77%)	74 (73%)		129 (75%)
Fumeur			0.2	
Non	51 (72%)	82 (80%)		133 (77%)
Oui	20 (28%)	20 (20%)		40 (23%)
Alcoolémie			0.9	
0 verre	18 (25%)	30 (29%)		48 (28%)
0 et 10 verres	39 (55%)	52 (51%)		91 (53%)
11 et 35 verre	12 (17%)	18 (18%)		30 (17%)
Plus 35	2 (2.8%)	2 (2.0%)		4 (2.3%)
kinésophobie			0.6	
Non	28 (39%)	36 (35%)		64 (37%)
Oui	43 (61%)	66 (65%)		109 (63%)
Fonctions Cognitives			0.065	
Anormales	50 (70%)	84 (82%)		134 (77%)
Normales	21 (30%)	18 (18%)		39 (23%)

$n(\%)^1$; Test exact de Fisher ; Test du Chi-carré de Pearson^2

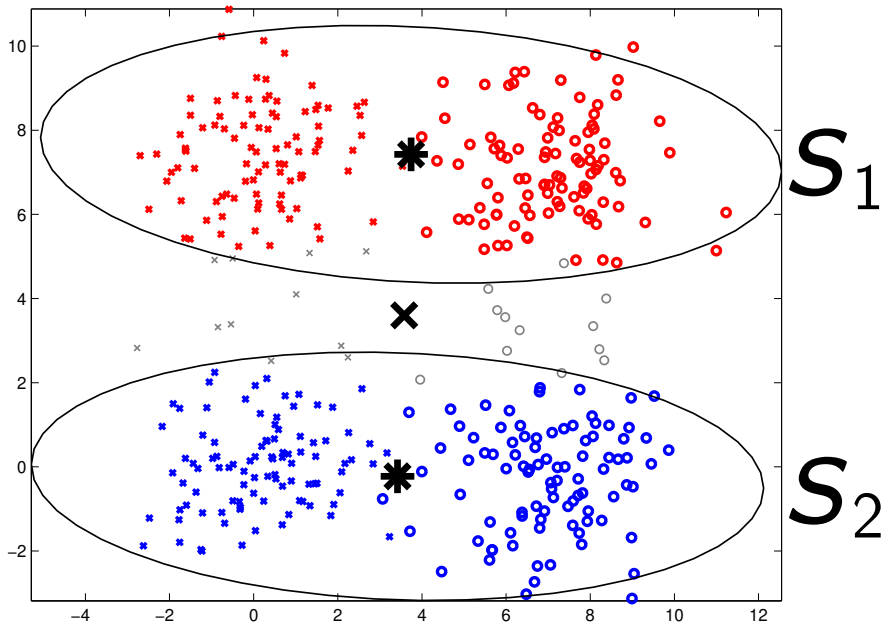
[1] A. Soubeiga, V. Antoine, A. Corteval, N. Kerckhove, S. Moreno, I. Falih, J. Phalip. *Clustering and Interpretation of time-series trajectories of chronic pain using evidential c-means*, Expert Systems With Application (en révision)

Plan

Distance adaptative

Distance de Mahalanobis pour chaque classe ω_k

- Chaque cluster ω_k est représenté par un centre $\boldsymbol{v}_k$
- Chaque cluster ω_k a une matrice de covariance $\boldsymbol{S}_k \succ 0$



Définition $\forall A_j \subseteq \Omega, A_j \neq \emptyset$

$$d_{ij}^2 = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^T \overline{\mathbf{S}}_j (\mathbf{x}_i - \overline{\boldsymbol{v}}_j)$$

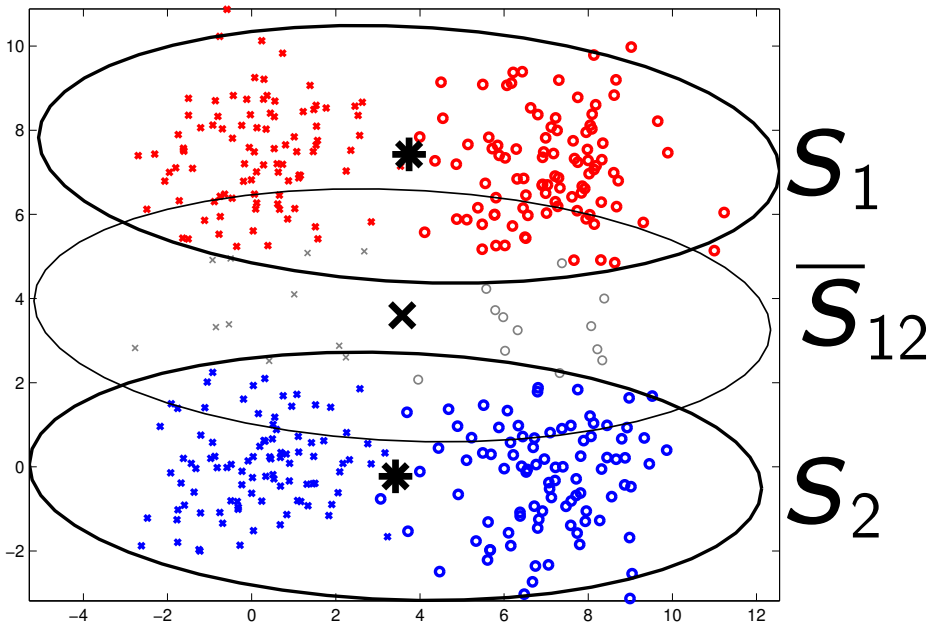
tel que

$$\overline{\mathbf{S}}_j = \frac{1}{|A_j|} \sum_{\omega_k \in A_j} \mathbf{S}_k,$$

Distance adaptative

Distance de Mahalanobis pour chaque classe ω_k

- Chaque cluster ω_k est représenté par un centre $\boldsymbol{v}_k$
- Chaque cluster ω_k a une matrice de covariance $\boldsymbol{S}_k \succ 0$



Définition $\forall A_j \subseteq \Omega, A_j \neq \emptyset$

$$d_{ij}^2 = (\mathbf{x}_i - \overline{\mathbf{v}}_j)^T \overline{\mathbf{S}}_j (\mathbf{x}_i - \overline{\boldsymbol{v}}_j)$$

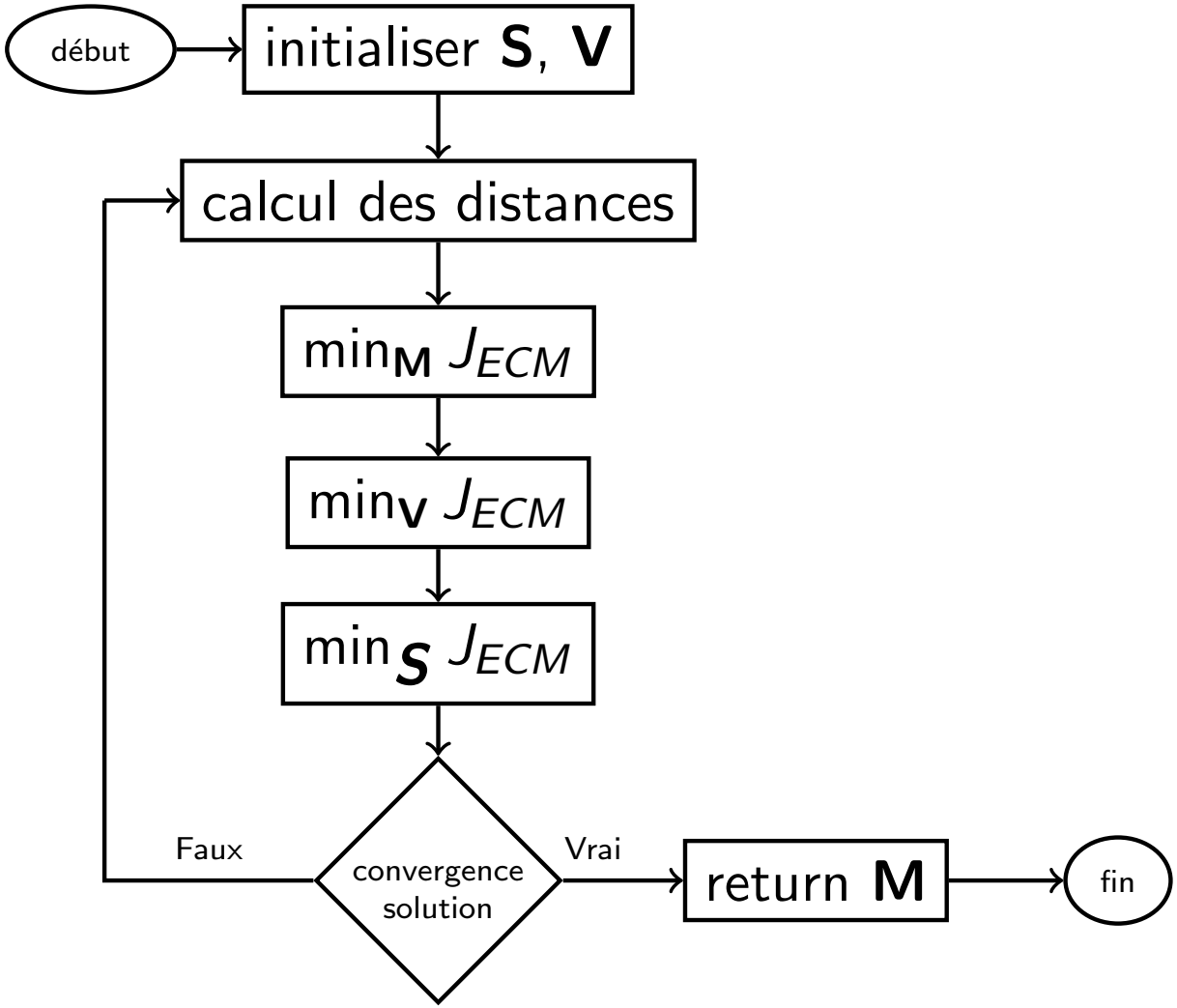
tel que

$$\overline{\mathbf{S}}_j = \frac{1}{|A_j|} \sum_{\omega_k \in A_j} \mathbf{S}_k,$$

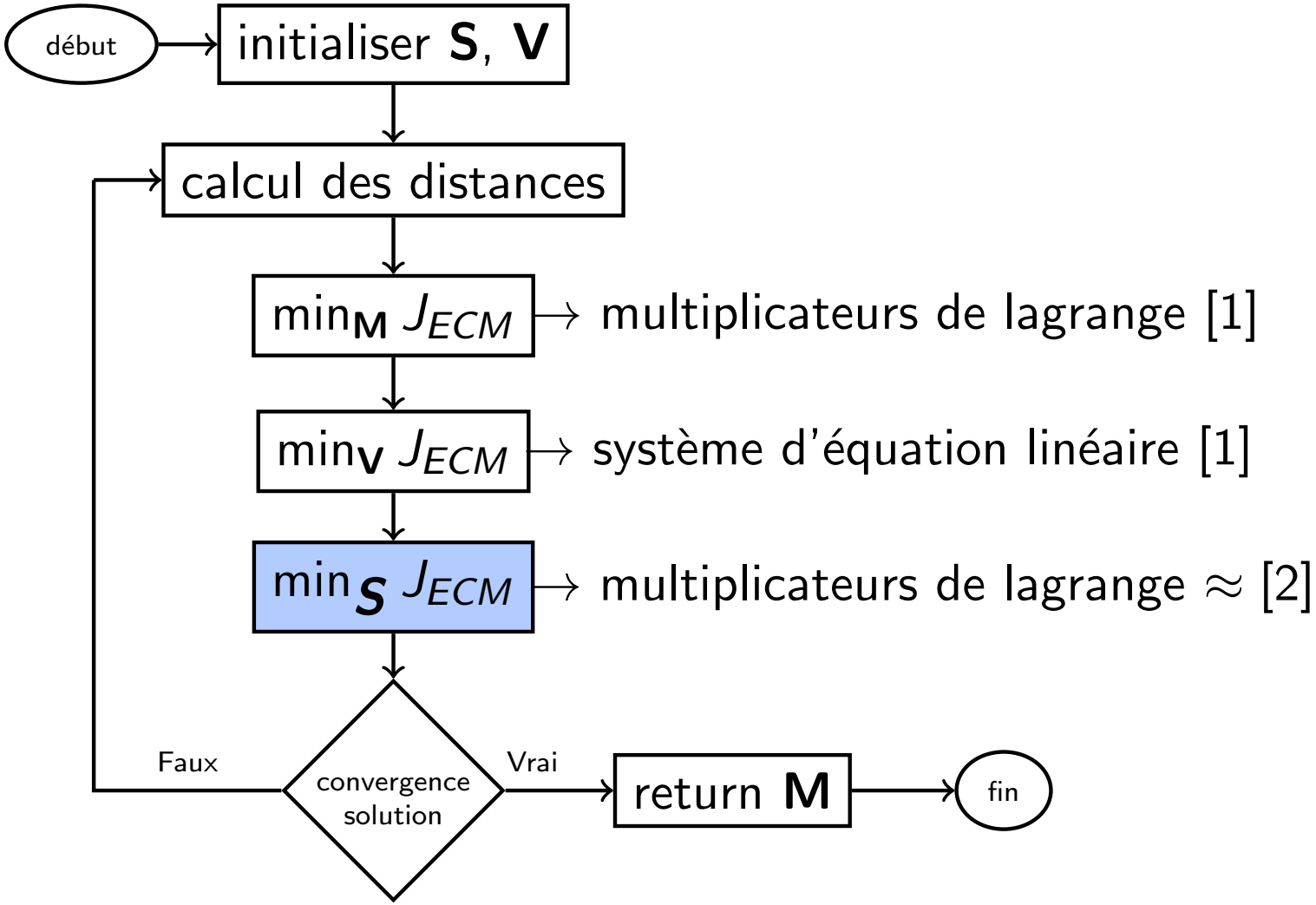
Nouvelle fonction objectif

Minimiser $J_{ECM}(\boldsymbol{M}, \boldsymbol{V}, \boldsymbol{S})$ tq $\boldsymbol{S}_k \succ 0$, $\det(\mathbf{S}_k) = 1 \quad \forall k = 1, c$

Méthode d'optimisation de type Gauss-Seidel



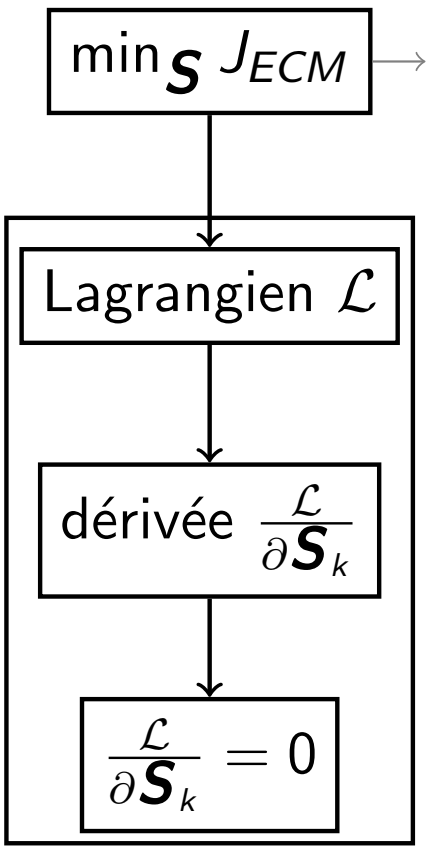
Méthode d'optimisation de type Gauss-Seidel



[1] M.-H. Masson & al, *ECM: An evidential version of the fuzzy c-means algorithm*, 2008

[2] D. Gustafson & al, *Fuzzy clustering with a fuzzy covariance matrix*, 1978

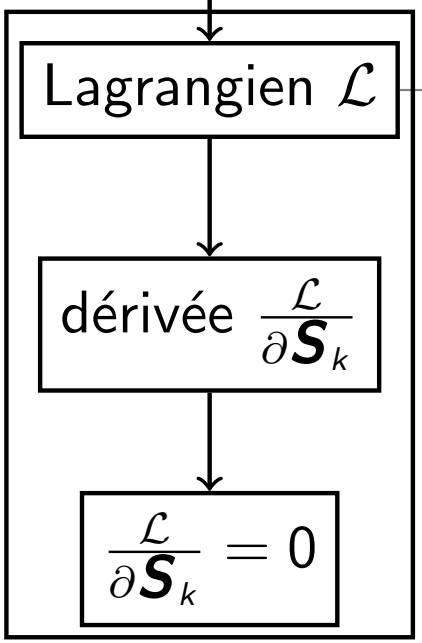
Méthode d'optimisation de type Gauss-Seidel



$$J_{ECM}(\boldsymbol{S}) = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T \overline{\boldsymbol{S}}_j (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j) + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}$$

tq $\det(\boldsymbol{S}_k) = 1$, $\boldsymbol{S}_k \succ 0$, $\forall k \in 1, c$

Méthode d'optimisation de type Gauss-Seidel



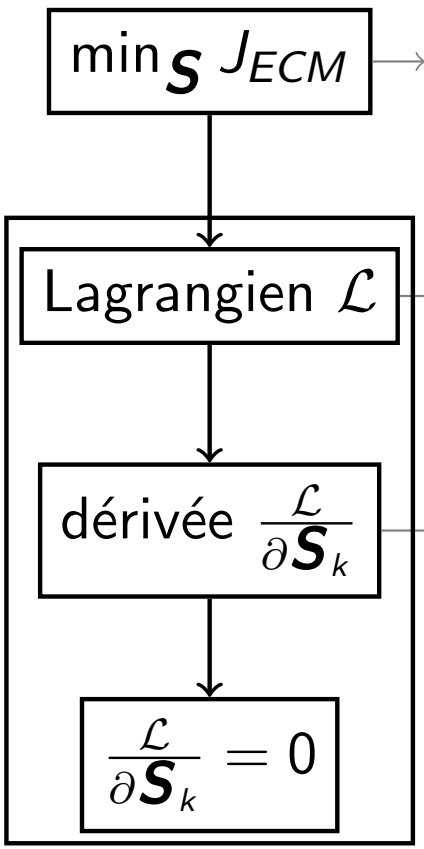
$$J_{ECM}(\boldsymbol{S}) = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T \overline{\boldsymbol{S}}_j (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j) + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}$$

tq $\det(\boldsymbol{S}_k) = 1$, $\boldsymbol{S}_k \succ 0$, $\forall k \in 1, c$

Omission des contraintes $\boldsymbol{S}_k \succ 0 \quad \forall k \in 1, c$

$$\mathcal{L}(\boldsymbol{S}, \lambda_1, \ldots, \lambda_c) = J_{ECM} - \sum_{k=1}^{c} \lambda_k (\det(\boldsymbol{S}_k) - 1)$$

Méthode d'optimisation de type Gauss-Seidel



$$J_{ECM}(\boldsymbol{S}) = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T \overline{\boldsymbol{S}}_j (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j) + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}$$

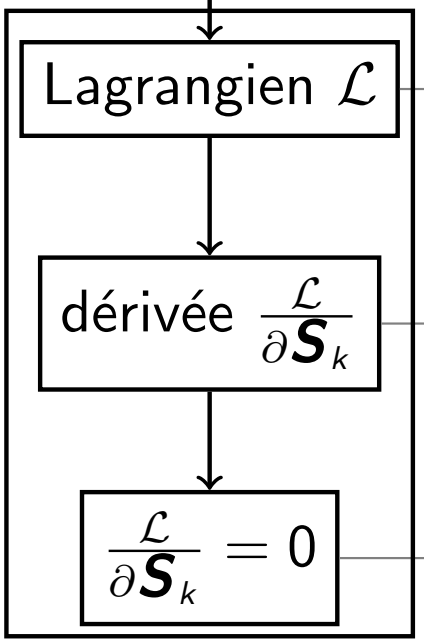
tq $\det(\boldsymbol{S}_k) = 1$, $\boldsymbol{S}_k \succ 0$, $\forall k \in 1, c$

Omission des contraintes $\boldsymbol{S}_k \succ 0 \quad \forall k \in 1, c$

$$\mathcal{L}(\boldsymbol{S}, \lambda_1, \ldots, \lambda_c) = J_{ECM} - \sum_{k=1}^{c} \lambda_k (\det(\boldsymbol{S}_k) - 1)$$

$$= \sum_{i=1}^{n} \sum_{\omega_k \ni A_j} |A_j|^{\alpha - 1} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)(\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T - \lambda_k \det(\boldsymbol{S}_k) \boldsymbol{S}_k^{-1}$$

Méthode d'optimisation de type Gauss-Seidel



$$J_{ECM}(\boldsymbol{S}) = \sum_{i=1}^{n} \sum_{A_j \neq \emptyset} |A_j|^{\alpha} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T \overline{\boldsymbol{S}}_j (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j) + \sum_{i=1}^{n} \delta^2 m_{i\emptyset}^{\beta}$$

tq $\det(\boldsymbol{S}_k) = 1$, $\boldsymbol{S}_k \succ 0$, $\forall k \in 1, c$

Omission des contraintes $\boldsymbol{S}_k \succ 0 \quad \forall k \in 1, c$

$$\mathcal{L}(\boldsymbol{S}, \lambda_1, \ldots, \lambda_c) = J_{ECM} - \sum_{k=1}^{c} \lambda_k (\det(\boldsymbol{S}_k) - 1)$$

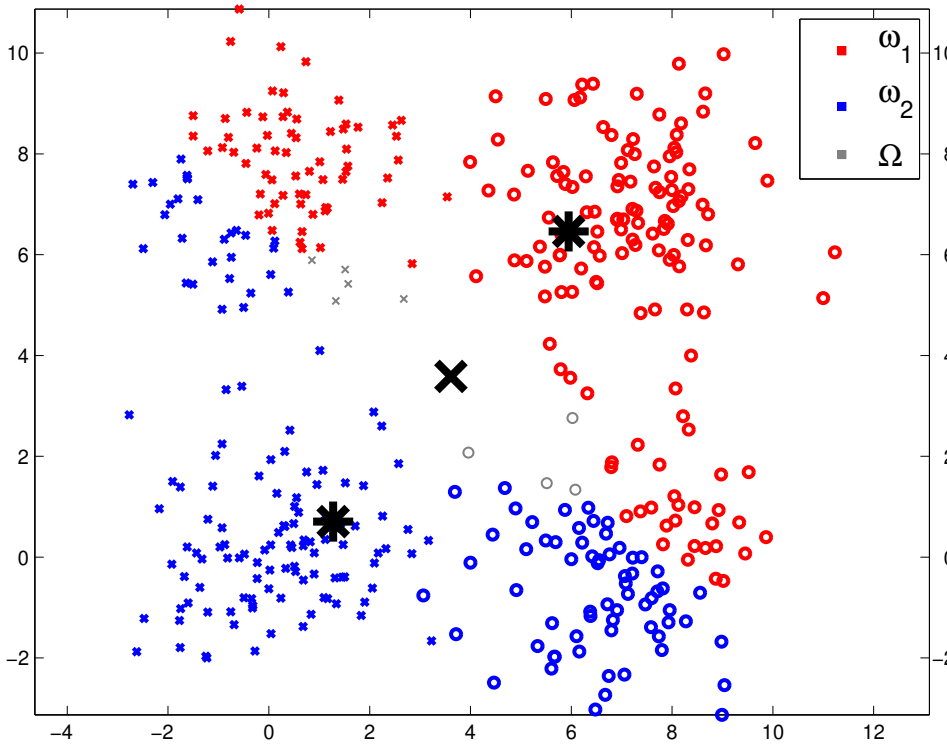
$$= \sum_{i=1}^{n} \sum_{\omega_k \ni A_j} |A_j|^{\alpha - 1} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)(\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T - \lambda_k \det(\boldsymbol{S}_k) \boldsymbol{S}_k^{-1}$$

$$\boldsymbol{S}_k = \det(\Sigma_k)^{\frac{1}{p}} \Sigma_k^{-1}$$

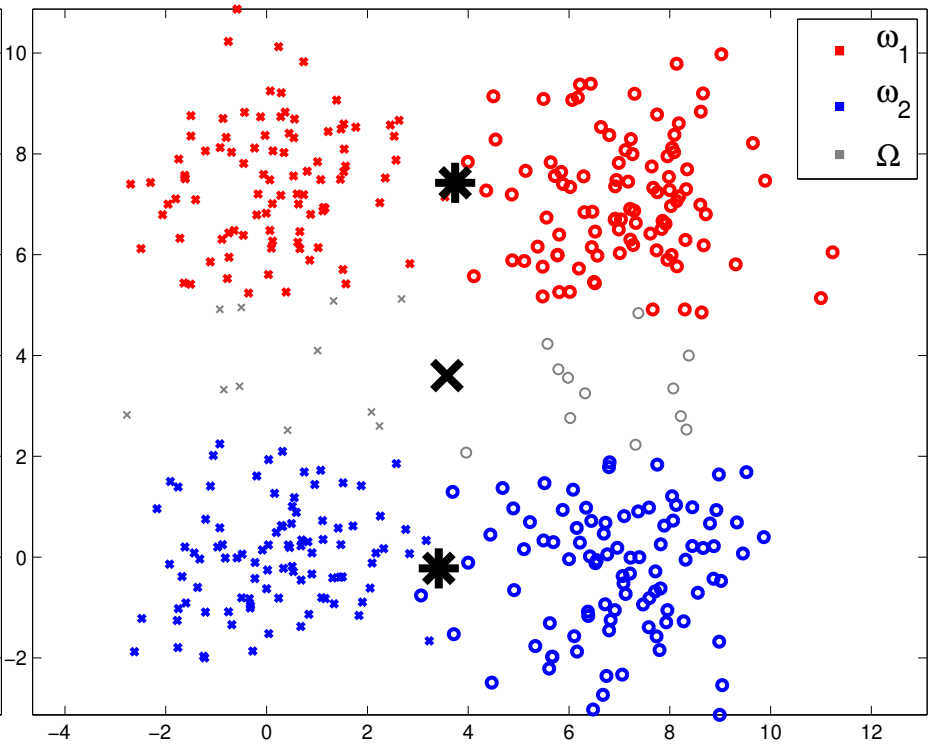
$$\Sigma_k = \sum_{i=1}^{n} \sum_{A_j \ni \omega_k} |A_j|^{\alpha - 1} m_{ij}^{\beta} (\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)(\boldsymbol{x}_i - \overline{\boldsymbol{v}}_j)^T$$

Et vérification $\boldsymbol{S}_k \succ 0$

Expérience



ECM+Euclidean distance



ECM+Mahalanobis distance

[1] V. Antoine, B. Quost, M.-H. Masson and T. Denoeux. *CECM: Constrained Evidential C-Means algorithm*. Computational Statistics and Data Analysis, Vol. 56, Issue 4, pages 894-914, 2012.

Plan

Clustering semi-supervisé

Problématique du clustering

Aucune connaissance a priori

- comment définir la notion de similarité ?
- comment choisir une solution parmi plusieurs partition possible ?

Clustering semi-supervisé

Problématique du clustering

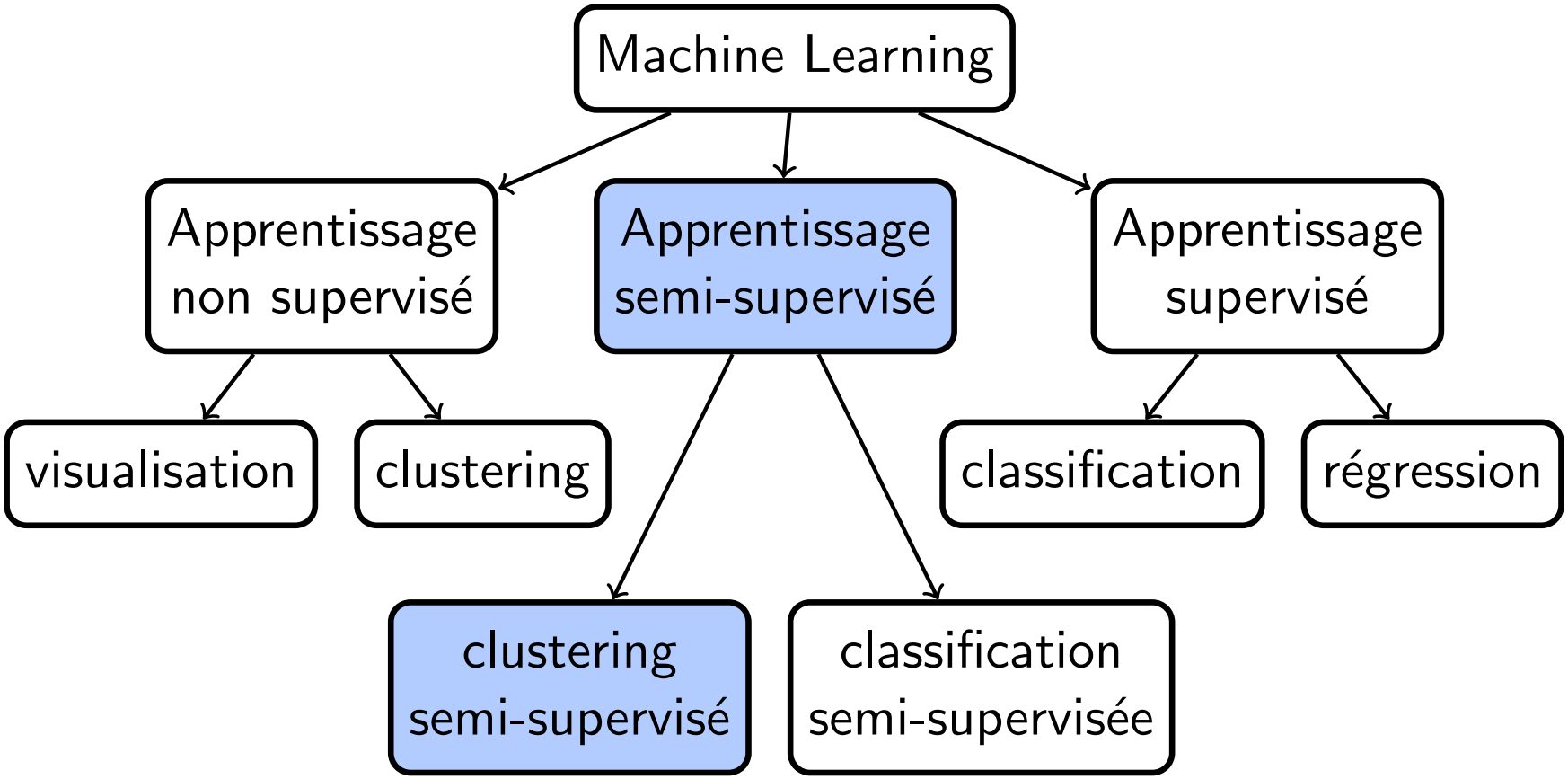
Aucune connaissance a priori

- comment définir la notion de similarité ?
- comment choisir une solution parmi plusieurs partition possible ?

Information provenant de l'expert

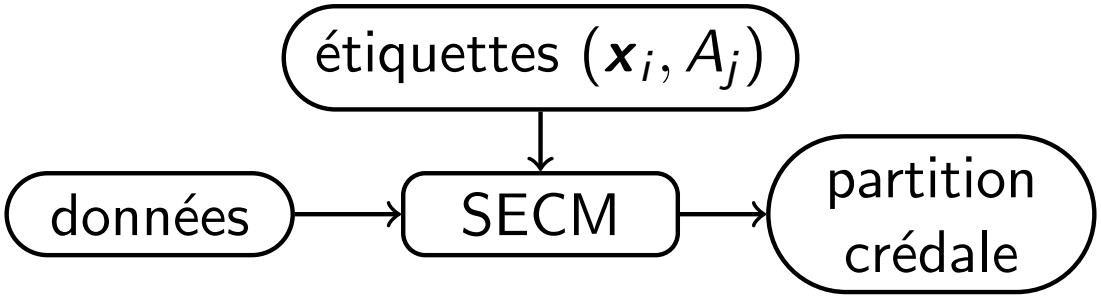
- étiquettes,
- contraintes par pair,
- classes équilibrées,...

Clustering semi-supervisé



Motivation

L'expert fournit des étiquettes imprécises A_j

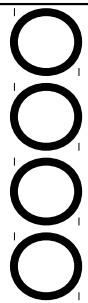


Exemple d'annotation d'expert

ω_1 pour les carrés, ω_2 pour les cercles, ω_3 pour les pentagones

	ω_1	ω_2	ω_3	A_j
	✓	✗	✗	ω_1
	✗	✓	✗	ω_2
	?	?	✗	$\omega_{12} = \{\omega_1, \omega_2\}$

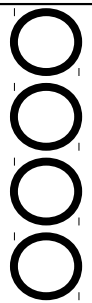
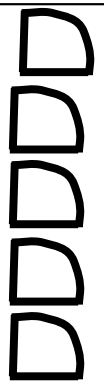
Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette			
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	Ω	A_j			
	1	0	0	0	0	0	0	ω_1	++		
	0	0	1	0	0	0	0	ω_1	+		
	0	0	0	0	0	0	1	ω_1	=		
	0	1	0	0	0	0	0	ω_1	-		

Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette			
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	Ω	A_j			
O	1	0	0	0	0	0	0	ω_1	++		
O	0	0	1	0	0	0	0	ω_1	+		
O	0	0	0	0	0	0	1	ω_1	=		
O	0	1	0	0	0	0	0	ω_1	-		
D	0	1	0	0	0	0	0	ω_{12}	++		
D	0	0	1	0	0	0	0	ω_{12}	+		
D	0	0	0	0	1	0	0	ω_{12}	=		
D	0	0	0	0	0	0	1	ω_{12}	=		
D	0	0	0	1	0	0	0	ω_{12}	-		

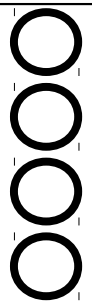
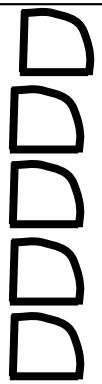
Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette		r=1
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	Ω	A_j		T_{ij}
	1	0	0	0	0	0	0	ω_1	++	1
	0	0	1	0	0	0	0	ω_1	+	1/2
	0	0	0	0	0	0	1	ω_1	=	1/3
	0	1	0	0	0	0	0	ω_1	-	0
	0	1	0	0	0	0	0	ω_{12}	++	1
	0	0	1	0	0	0	0	ω_{12}	+	$\sqrt{2}/2$
	0	0	0	0	1	0	0	ω_{12}	=	1/2
	0	0	0	0	0	0	1	ω_{12}	=	$\sqrt{2}/3$
	0	0	0	1	0	0	0	ω_{12}	-	0

Mesure de cohérence

$$T_{ij} = T_i(A_j) = \sum_{A_\ell \cap A_j \neq \emptyset} \frac{|A_j \cap A_\ell|^{r/2}}{|A_\ell|^r} m_{i\ell},\ r \geq 0 \text{ un hyperparamètre}$$

Cohérence entre étiquettes et partition crédale dure

	partition crédale							étiquette		r=1	r=0
	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	Ω	A_j		T_{ij}	T_{ij}
	1	0	0	0	0	0	0	ω_1	++	1	1
	0	0	1	0	0	0	0	ω_1	+	1/2	1
	0	0	0	0	0	0	1	ω_1	=	1/3	1
	0	1	0	0	0	0	0	ω_1	-	0	0
	0	1	0	0	0	0	0	ω_{12}	++	1	1
	0	0	1	0	0	0	0	ω_{12}	+	$\sqrt{2}/2$	1
	0	0	0	0	1	0	0	ω_{12}	=	1/2	1
	0	0	0	0	0	0	1	ω_{12}	=	$\sqrt{2}/3$	1
	0	0	0	1	0	0	0	ω_{12}	-	0	0

Mesure de cohérence

$$T_{ij} = T_i(A_j) = \sum_{A_\ell \cap A_j \neq \emptyset} \frac{|A_j \cap A_\ell|^{r/2}}{|A_\ell|^r} m_{i\ell},\ r \geq 0 \text{ un hyperparamètre}$$

Étude de l'hyperparamètre r

	$m_{i\omega_1}$	$m_{i\omega_2}$	$m_{i\omega_{12}}$	$m_{i\omega_3}$	$m_{i\omega_{13}}$	$m_{i\omega_{23}}$	Ω	A_j	r=1, T_{ij}		r=0, T_{ij}	
O	1	0	0	0	0	0	0	ω_1	++	1	+	1
O	0	0	1	0	0	0	0	ω_1	+	1/2	+	1
O	0	0	0	0	0	0	1	ω_1	=	1/3	+	1
O	0	1	0	0	0	0	0	ω_1	-	0	-	0
D	0	1	0	0	0	0	0	ω_{12}	++	1	+	1
D	0	0	1	0	0	0	0	ω_{12}	+	$\sqrt{2}/2$	+	1
D	0	0	0	0	1	0	0	ω_{12}	=	1/2	+	1
D	0	0	0	0	0	0	1	ω_{12}	=	$\sqrt{2}/3$	+	1
D	0	0	0	1	0	0	0	ω_{12}	-	0	-	0

Mesure de cohérence

- $r = 0 \Rightarrow$ ne pénalise pas les sous-ensembles de grandes cardinalités. Utile en cas de bruit dans les étiquettes.
- $r > 0 \Rightarrow$ pénalise les sous-ensembles de grandes cardinalités. Etiquettes certaines.

Semi-supervised evidential clustering: SECM

Idée globale

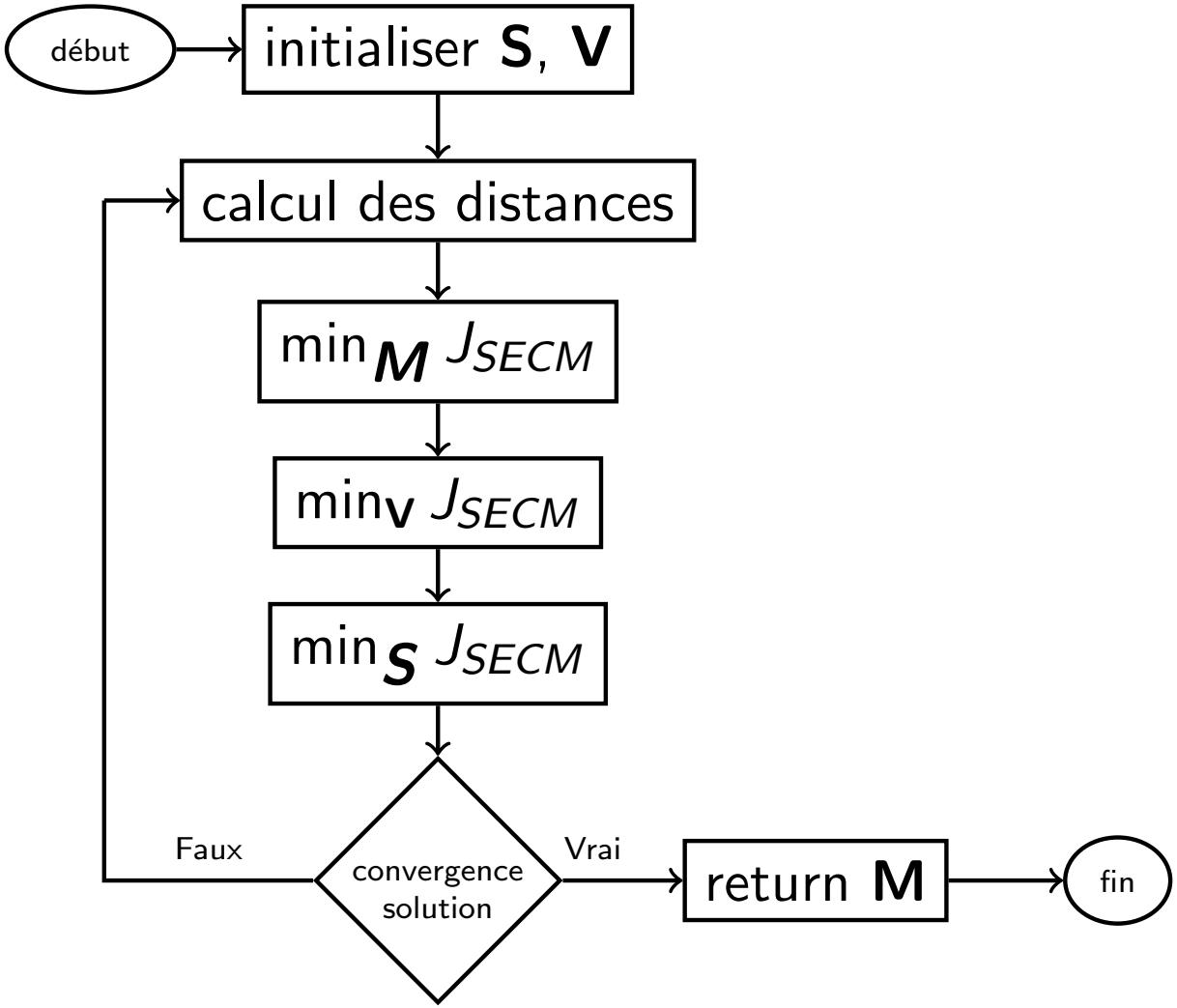
Si $\mathbf{x}_i \in A_j \Rightarrow T_{ij}$ doit être élevé

Fonction objectif

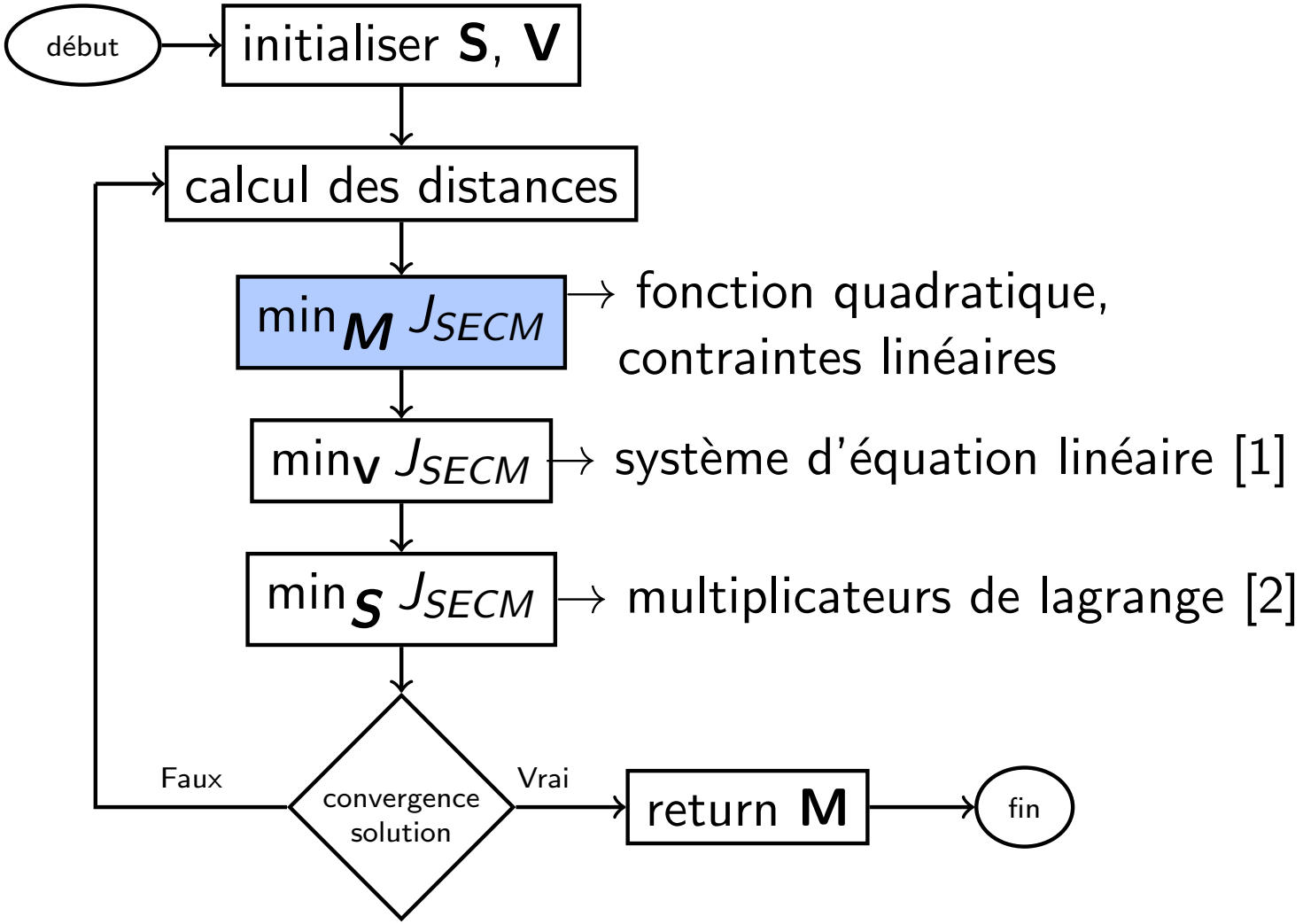
$$J_{SECM} = (1-\gamma)J_{ECM} + \gamma \sum_{i=1}^{n} \sum_{A_j \subseteq \Omega, A_j \neq \emptyset} b_{ij}(1 - T_{ij})$$

tel que $b_{ij} = \begin{cases} 1 & \text{si } \mathbf{x}_i \text{ est contraint avec } A_j, \\ 0 & \text{sinon.} \end{cases}$

Méthode d'optimisation de type Gauss-Seidel



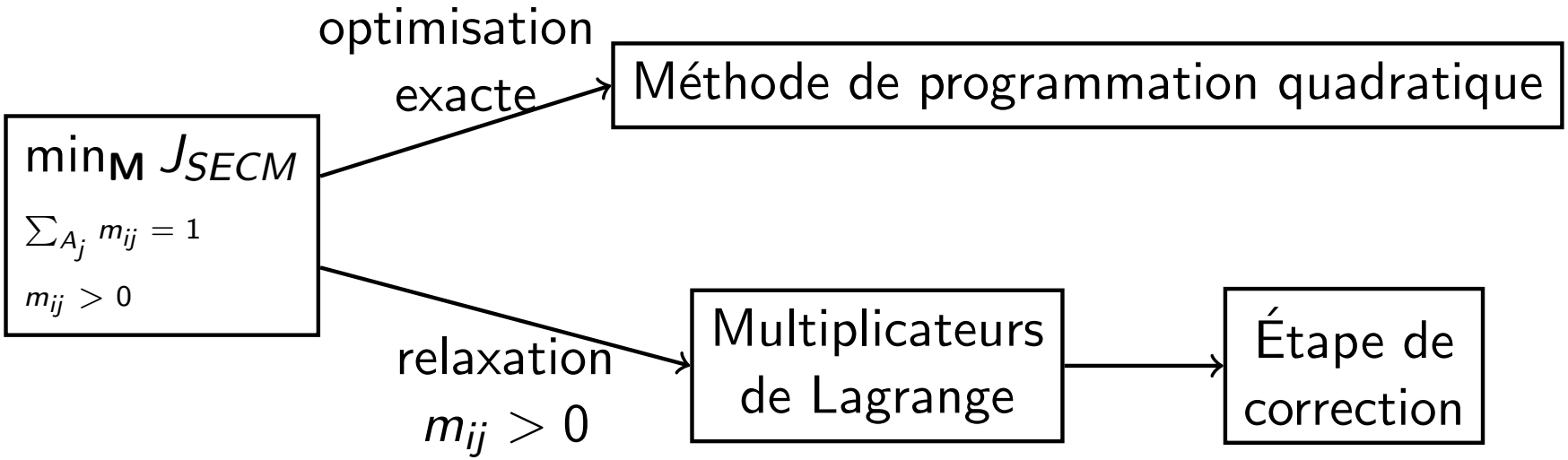
Méthode d'optimisation de type Gauss-Seidel



[1] M.-H. Masson & al, *ECM: An evidential version of the fuzzy c-means algorithm*, 2008

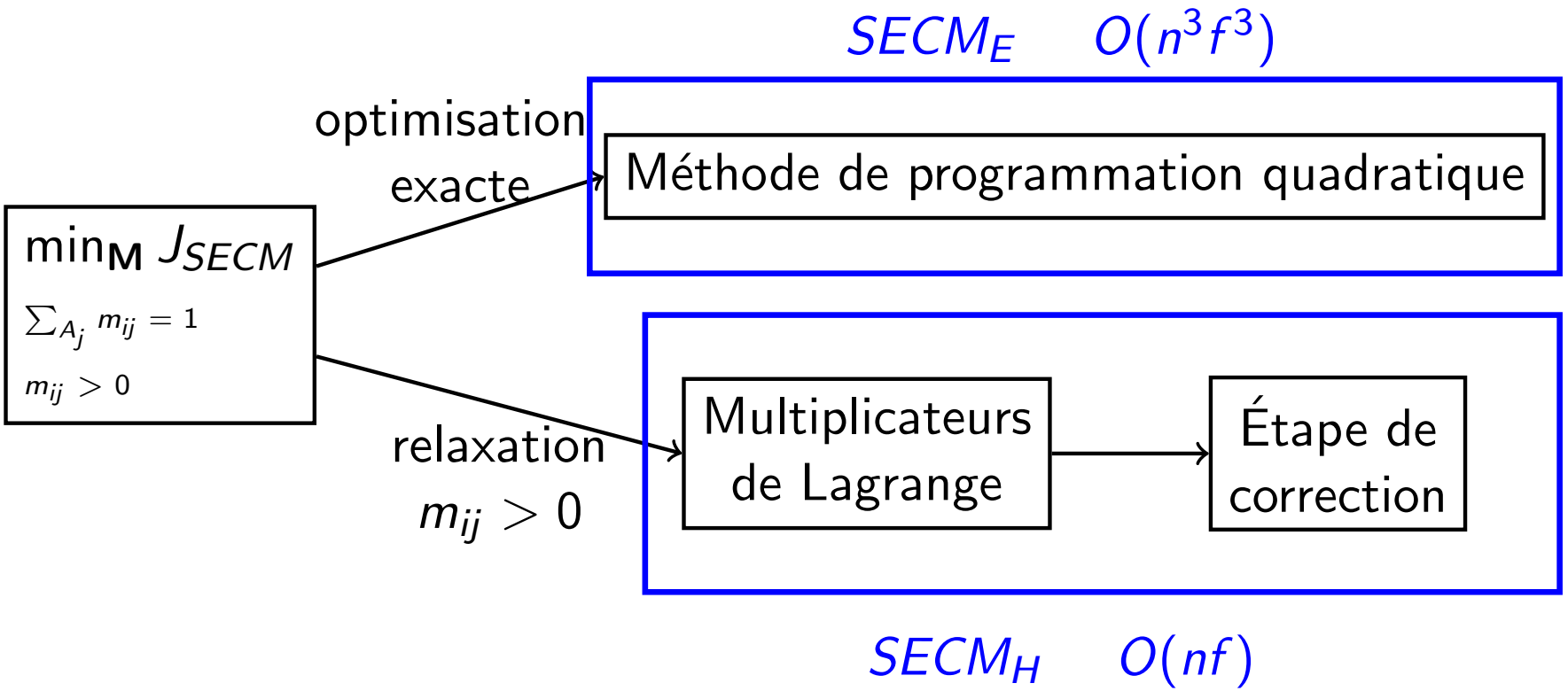
[2] V. Antoine, & al, *CECM: Constrained Evidential C-Means algorithm*, 2012.

Optimisation de la partition crédale



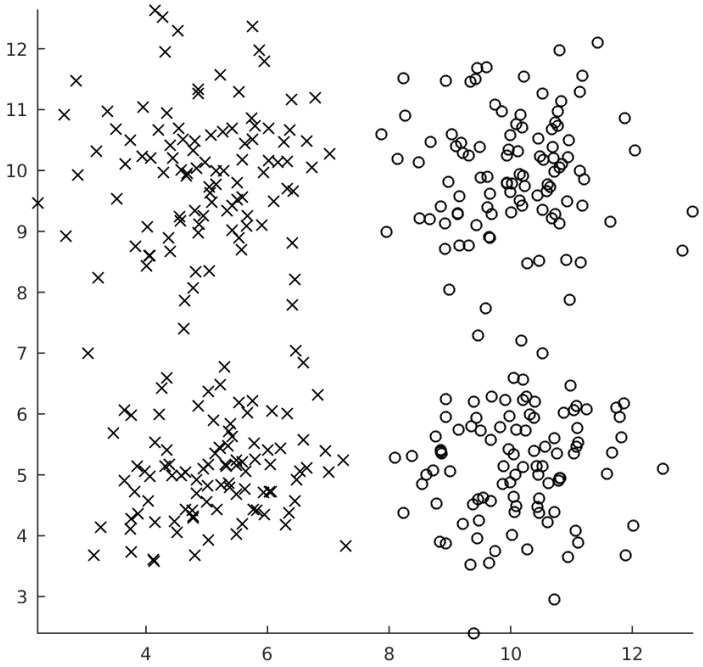
[1] V. Antoine, J. Guerrero, J. Xie. *Fast semi-supervised evidential clustering.* International Journal of Approximate Reasonning, Vol. 133, pp 116-132, 2021.

Optimisation de la partition crédale

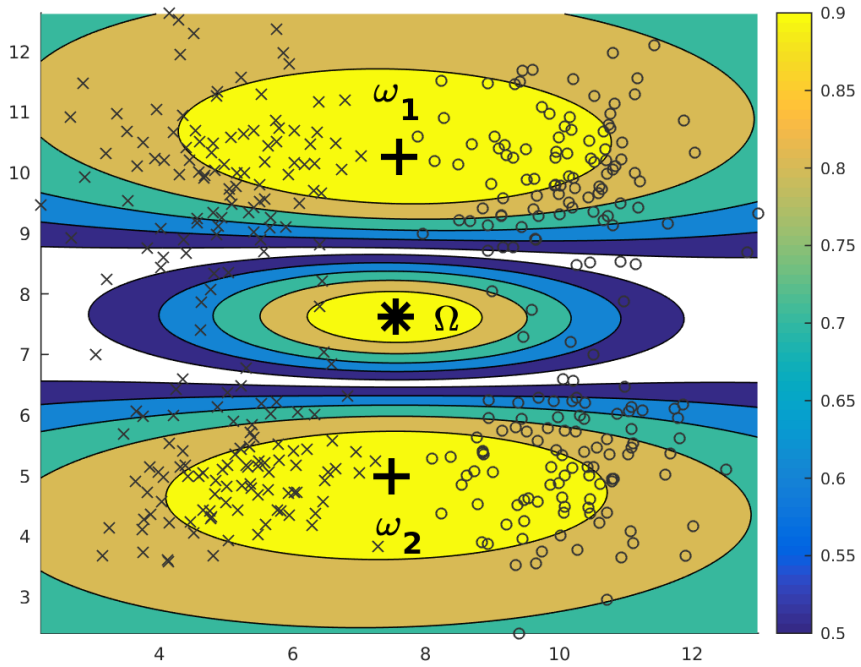


[1] V. Antoine, J. Guerrero, J. Xie. *Fast semi-supervised evidential clustering.* International Journal of Approximate Reasonning, Vol. 133, pp 116-132, 2021.

Intérêt des contraintes

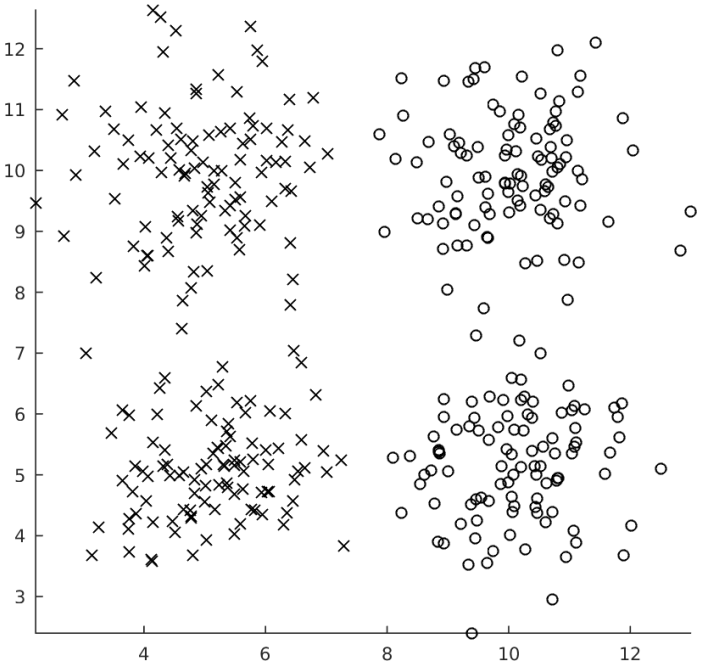


gaussK2

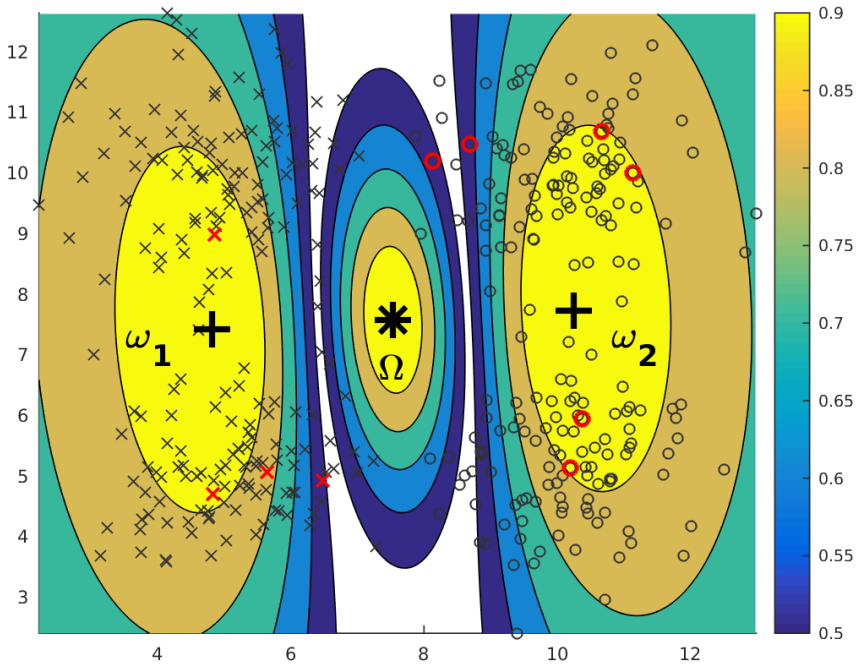


ECM

Intérêt des contraintes



gaussK2



SECM

Protocole experimental

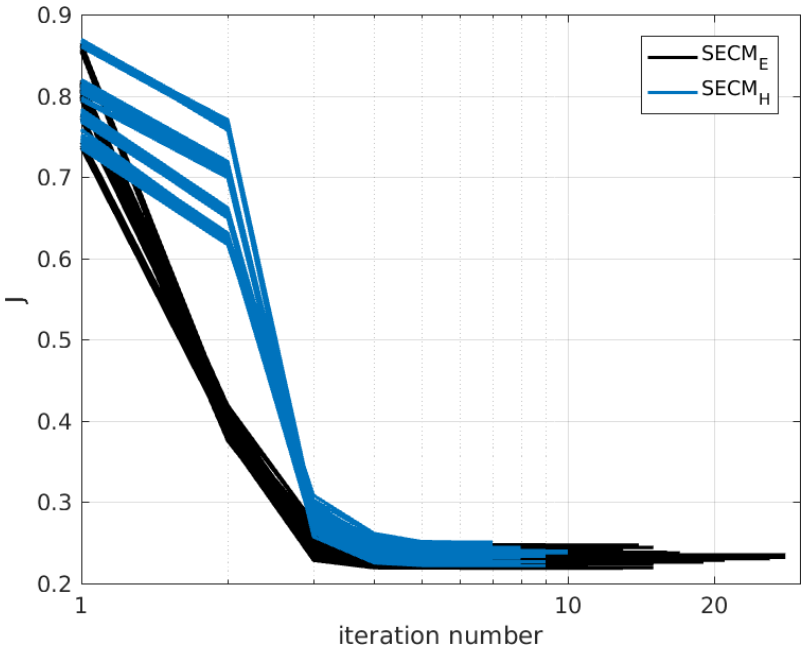
Jeux de données

	# objets	# attributs	# classes
Column	310	6	3
Wine	178	13	3

Méthode d'évaluation basée sur les vraie classes

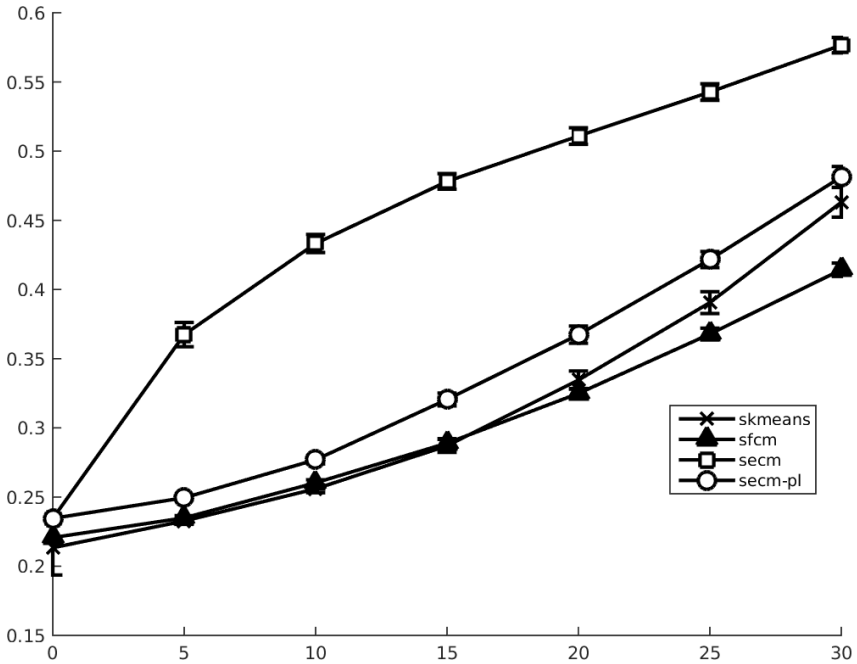
- sélection aléatoire des contraintes
- mesure d'évaluation:
  - transformation pignistique $\Rightarrow$ partition floue
  - maximum de probabilité $\Rightarrow$ partition dure
  - ARI $\in [0, 1]$

Analyse de l'optimization sur le jeu de données Wine

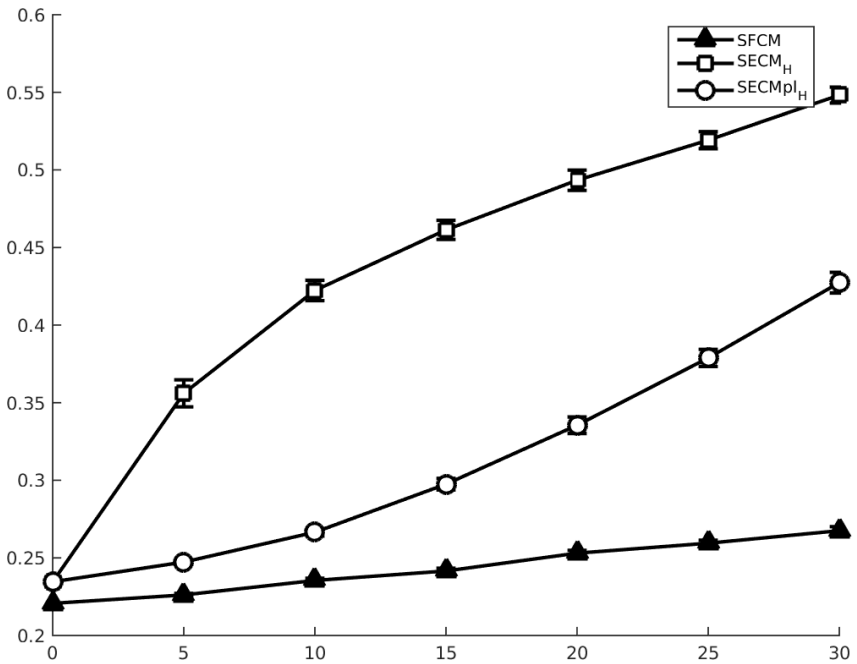


30 cont.	$SECM_H$	$SECM_E$
J_{SECM}	236.3[1.1]	232.7[1.1]
CPU (s)	0.19[0.00]	0.89[0.03]
ARI	0.92[0.02]	0.92[0.03]

Comparaison d'algorithme sur le jeu de données Column



étiquettes simples

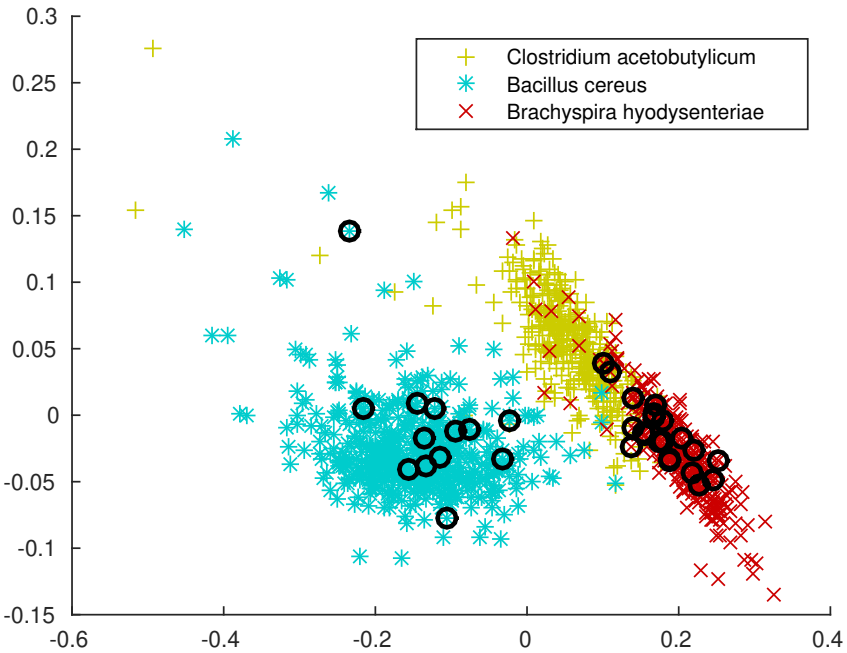


étiquettes doubles

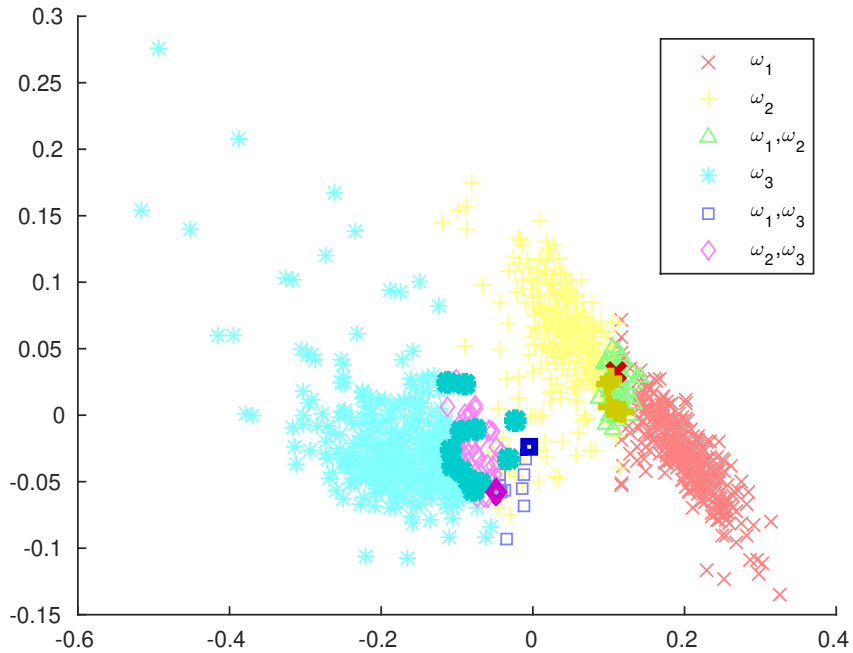
Application en génomique

Jeu de données tetragen

Séquences d'ADN dont les plus grandes ont été divisées en plusieurs objets ⇒ génération d'étiquettes



ECM



SECM

[1] V. Antoine, K. Gravouil, N. Labroche. *On evidential clustering with partial supervision.* BELIEF, 2018.

Plan

Conclusion ...

ECM avec une distance adaptative

- clustering évidentiel
+ généralisation de la distance Euclidienne
+ permet de trouver des clusters de forme ellipsoïdale
- complexité
- plus sensible au minima locaux

SECM

- clustering évidentiel
- ajout d'étiquettes
+ partition crédale comprend de nombreuses informations
+ les étiquettes améliorent les performances
- complexité
- sensibilité à la sélection d'étiquettes

... et perspectives

Perspectives à court terme

- subspace ECM
- définition améliorée des centres de gravité

Perspectives à long terme

- prendre en compte des données floues en entrée de ECM
- clustering évidentiel pour des données multisources de santé
  - notamment données ordinales et séries longitudinales

Merci


instance-level
partial
entropy
cevclus
constrained
labeled
methods
signal
constraint
labels
neural
classification
xie-beni
maximal
subsets
measure
processing
method
machine
anomalies
function
semi
supervised
k-nn
optimization
evidence
c-means
image
clustering
possibilistic
information
uncertainty
dairy
video
detect
adaptive
bayes
segmentation
sequences
theory
fuzzy
learning
cows
categorical
pairwise
evidential
belief
data
mahalanobis
series
relational
ensemble
constraints
algorithms
networks
detection
supervision
proximity
label
time
model
functions
distance
coherent
out-of-domain