Contrôle écrit - Apprentissage non supervisé

Durée : 1h45

Documents non autorisés, Calculettes autorisées, Répondre directement sur les feuilles

Nom :
Prénoms :

Questions de cours (7 points)

1. Quel lien existe-il entre la méthode SOM et l'analyse en composantes principales ?

2. Quelle similitude et quelle différence existe t-il entre l'algorithme SOM et la version séquentielle des k-means ?

3. En classification spectrale, quelle heuristique est généralement utiliée pour sélectionner le nombre de classes ?

4. Les algorithmes de classification spectrale se terminent généralement par une application de l'algorithme des k-means. Donner la raison théorique liée à cette utilisation finale des k-means.

5. Quelles propriétés théoriques nous garantissent-elles que la partition obtenue à la convergence de l'algorithme des k-means correspond bien à un optimum du critère d'inertie intra-classes ?

6. Dans quelle situation l'algorithme PAM peut-il s'avérer plus performant que l'algorithme des k-means ?

7. Proposer une version des nuées dynamiques permettant de partitionner des données en classes de forme elliptiques.

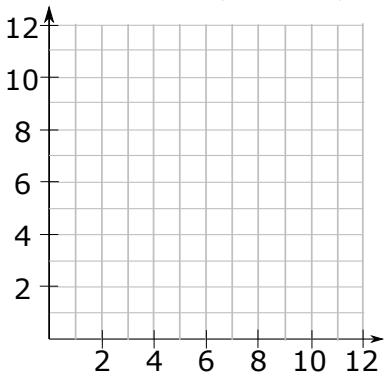
Exercice 1 (6 points)

Considérons le tableau de données suivant constitué de 6 individus $(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$ décrits par 2 variables quantitatives :

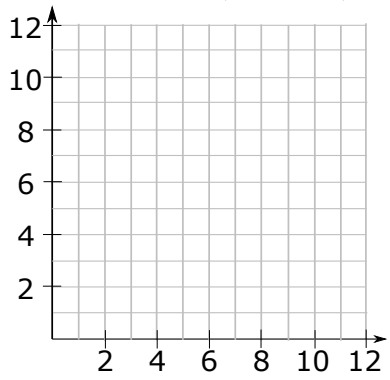
$$\begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 6 & 6 \\ 7 & 7 \\ 11 & 11 \\ 12 & 12 \end{pmatrix}$$

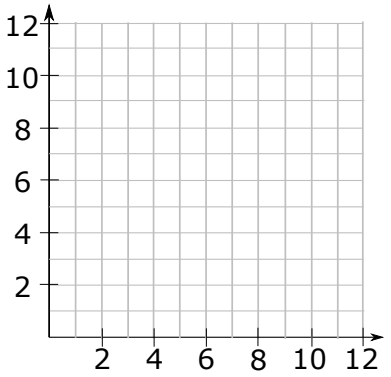
1. En appliquant l'algorithme des k-means, réaliser un clustering de ces données en $K = 3$ classes, en partant des deux initialisations différentes suivantes : (x_2, x_4, x_5) et (x_4, x_5, x_6).

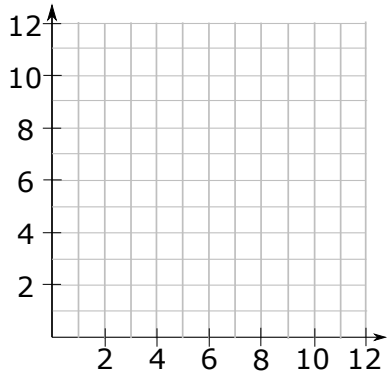
Initialisation (x_2, x_4, x_5)

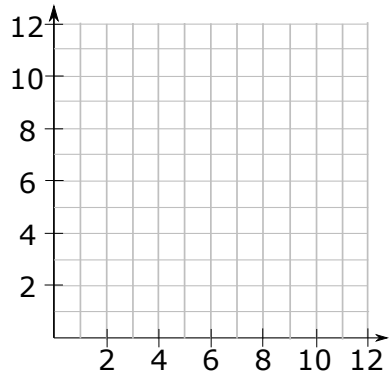


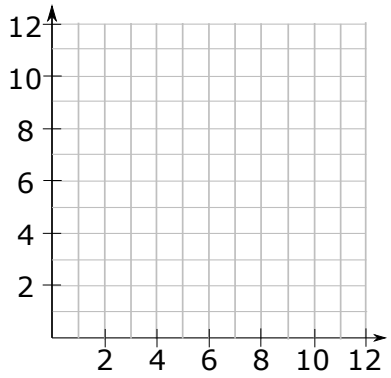
Initialisation (x_4, x_5, x_6)

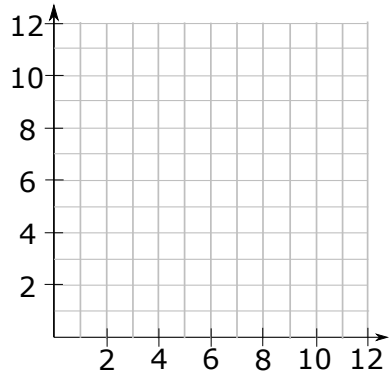


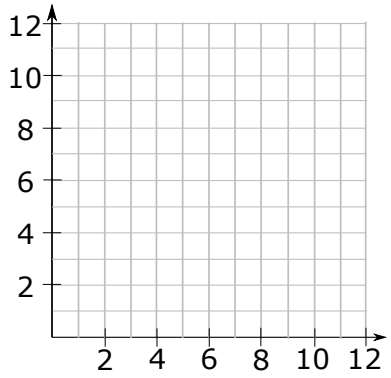

12
10
8
6
4
2
2 4 6 8 10 12

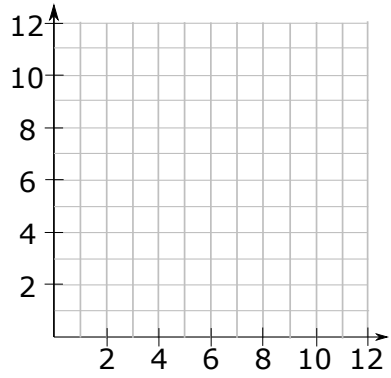

12
10
8
6
4
2
2 4 6 8 10 12

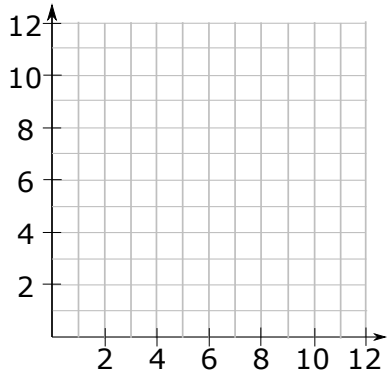

12
10
8
6
4
2
2 4 6 8 10 12

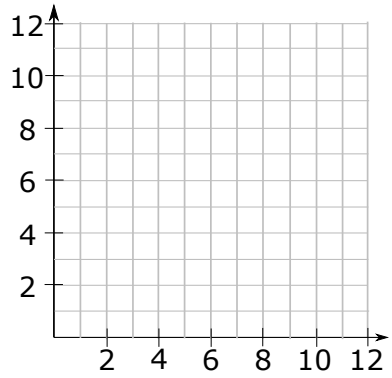

12
10
8
6
4
2
2 4 6 8 10 12

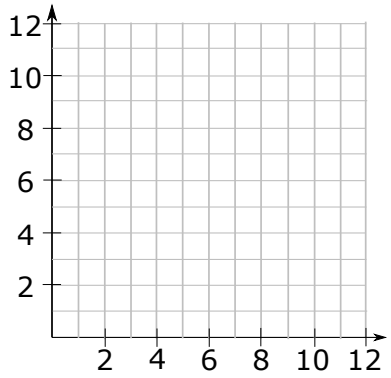

12
10
8
6
4
2
2 4 6 8 10 12


12
10
8
6
4
2
2 4 6 8 10 12


12
10
8
6
4
2
2 4 6 8 10 12


12
10
8
6
4
2
2 4 6 8 10 12

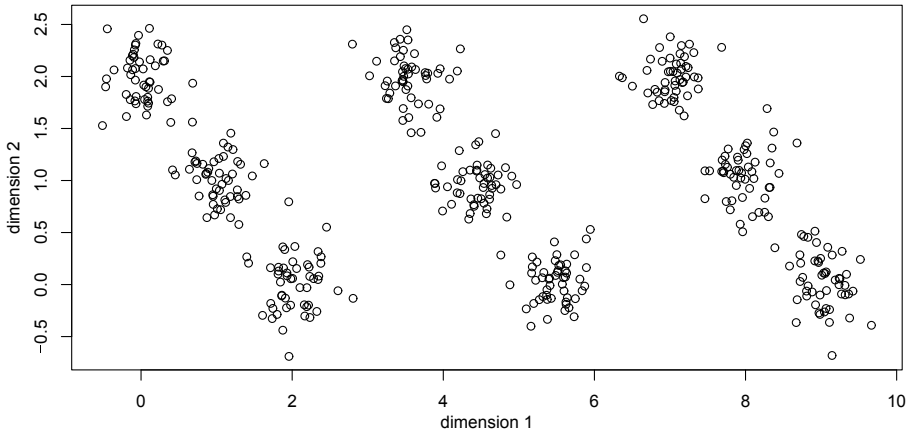

12
10
8
6
4
2
2 4 6 8 10 12


12
10
8
6
4
2
2 4 6 8 10 12

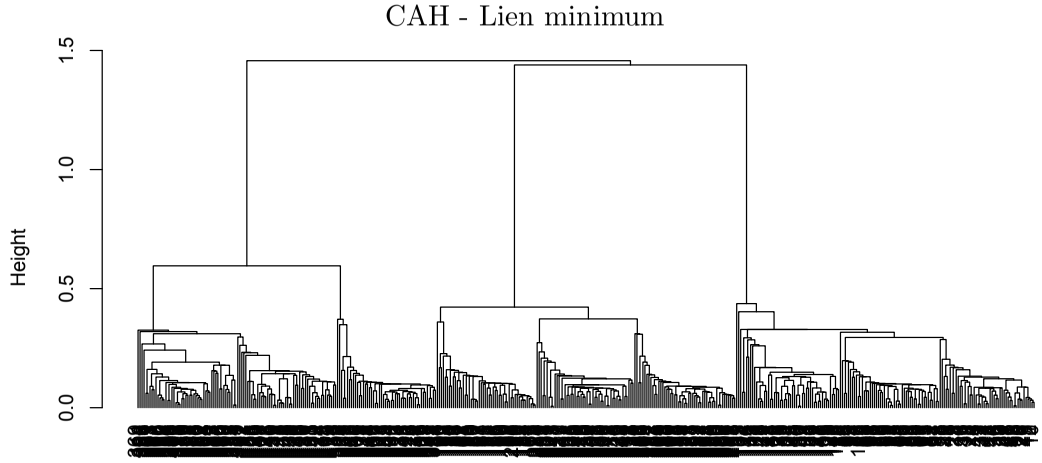
2. Proposer un critère pour choisir l'une de ces deux solutions. Calculer ce critère et en déduire la meilleure solution.

Exercice 2 (7 points)

On se propose de classifier le jeux de données suivant ($n = 450$ points décrit par deux variables), constitué de 9 classes sphériques de mêmes proportions.

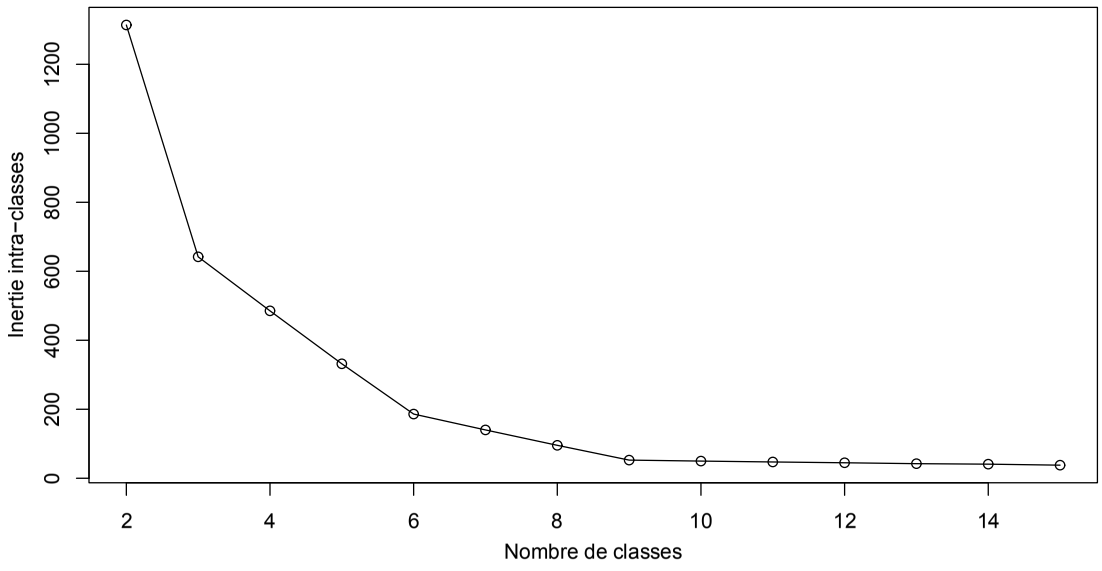


1. On lance tout d'abord l'algorithme de classification ascendante hiérarchique (CAH) sur ces données, en utilisant le critère d'agrégation du lien minimum. Les résultats obtenus sont les suivants.



Quel nombre de classes est suggéré par cette méthode ? Commentez.

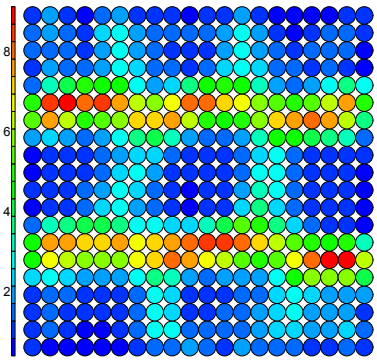
2. Ensuite, l'algorithme des k-means est lancé en faisant varier le nombre de classes de 2 à 15. La courbe d'inertie intra-classes obtenue en fonction du nombre de classes est la suivante :



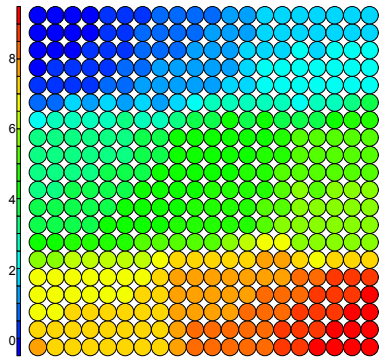
Expliquer pourquoi le fait de choisir le nombre de classes par minimisation de l'inertie intra-classes n'est pas judicieux. Quelle heuristique est généralement employée et quel nombre de classes nous suggère t'elle ici ?

3. Enfin, l'algorithme SOM a été lancé sur ces données, avec une grille 20×20. Les résultats obtenus sont donnés par les trois cartes suivantes.

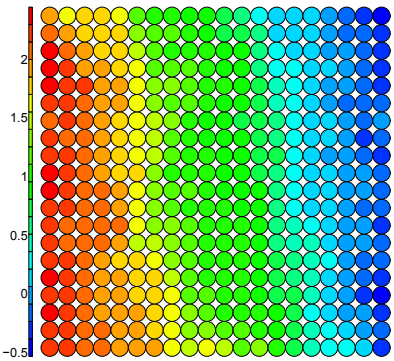
U-matrix



Component planes : variable 1



Component planes : variable 2



Etablir, en justifiant votre réponse, une correspondance entre les classes visibles sur la U-matrix et les vraies classes (figure initiale). *On pourra pour cela annoter les graphiques ci-dessus.*