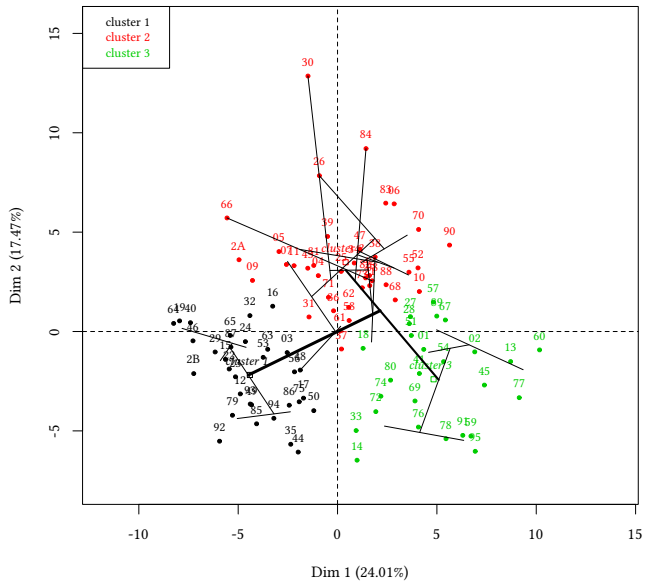
Classification et clustering

1. Dimensions et distance
2. La méthode des k-means
3. La classification ascendante hiérarchique



Dimensions et distances

On a un tableau de données multidimensionnel, et on souhaite classifier les individus.
Les objectifs peuvent etre multiples : analytique ou prédictif, levier d'action…
Cela revient à regrouper des individus en minimisant la distance à l'intérieur des groupes et maximisant la distance moyenne entre les groupes. D'où le lien avec la thématique de ce matin : la notion de distance est fondamentale.
Classifier, c'est avant tout établir une métrique.
Attention, plus la dimensionnalité est grande, plus le calcul des distances est coûteuse (d'un ordre de grandeur $O(n^2)$).

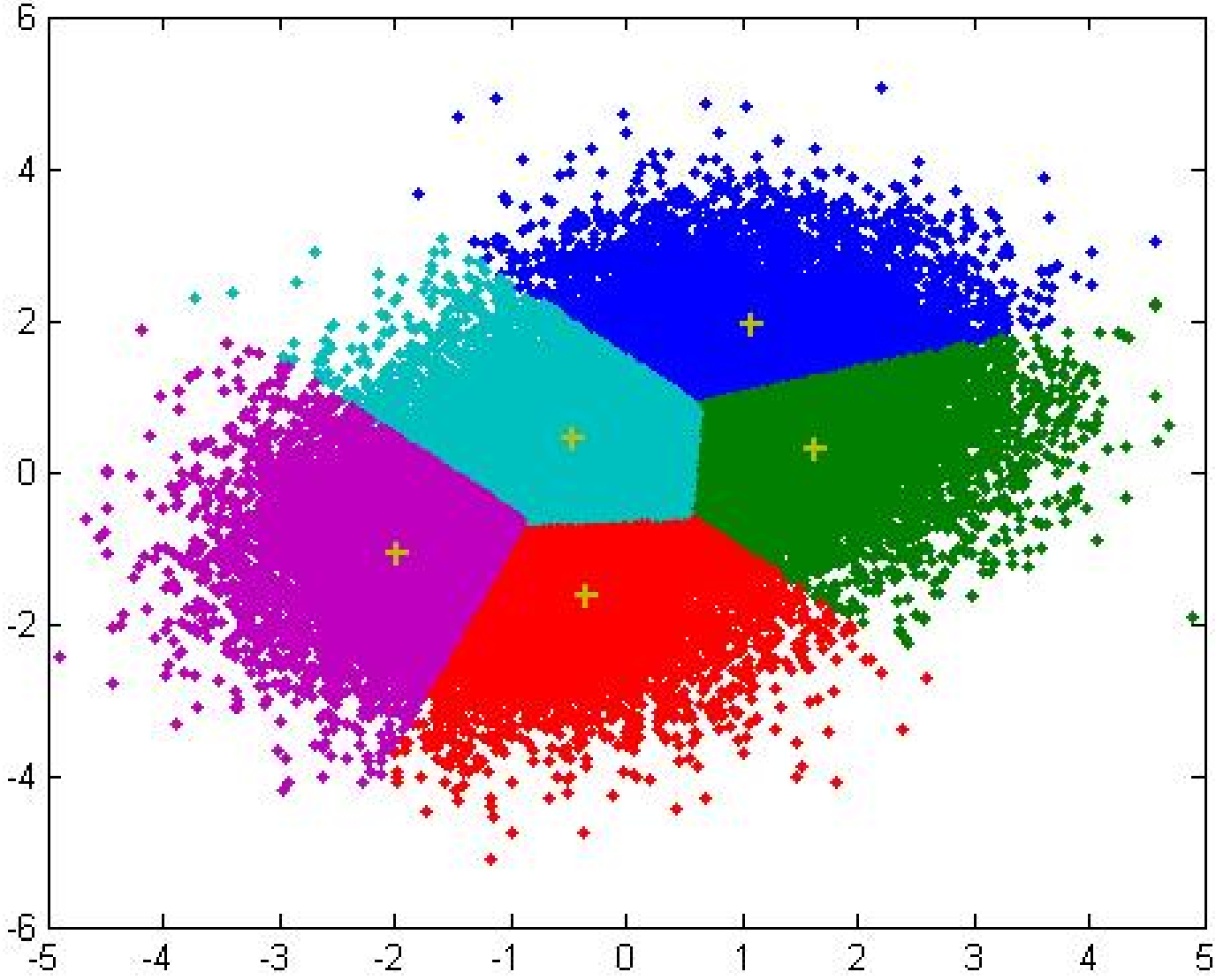
Classification hiérarchique et non-hiérarchique

Toutes les méthodes ici envisagées sont non paramétriques : elles ne font aucune hypothèse sur la distribution des individus au sein des classes.
Mais certaines sont hiérarchiques : elles emboîtent les classes les unes dans les autres ; d'autres ne le sont pas.

La méthode des k-means

Algorithme dont le point de départ est aléatoire et qui est itératif. Implique de calculer la matrice complète des distances (pas nécessairement euclidiennes). Porte généralement sur des variables continues.
L'idée est de partir de points moyens aléatoires, et d'assigner à chaque étape chaque observation à la classe la plus proche. On met à jour la moyenne, et on recommence.
La conséquence est que les solutions ne sont pas uniques – et pas nécessairement optimales non plus (optimum local plutôt que global).

La méthode des k-means



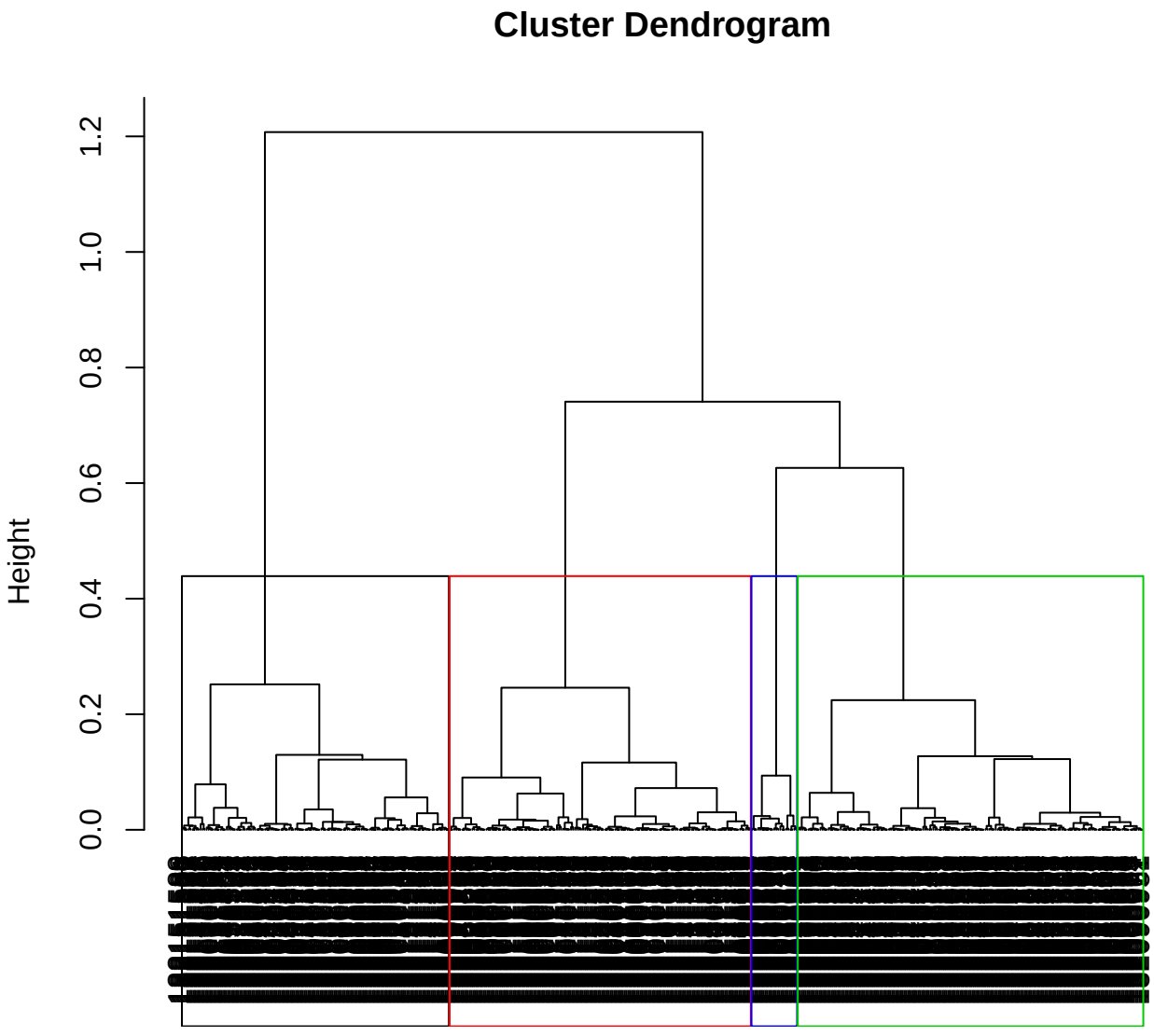
La méthode des k-means

On évalue la qualité d'une classification par k-means en comparant la variance intra et la variance inter. On recherche une variance intra aussi faible que possible et une variance inter ausi élevée que possible.
Mais ça ne part pas de l'idée qu'il existe une classification préalable, sous-jacente. Par contre l'utilisateur doit fixer (le cas échéant arbitrairement) le nombre de classes recherchées.

La classification ascendante hiérarchique (CAH)

La CAH part d'une situation où chaque individu constitue une classe à lui tout seul. À chaque étape de l'algorithme, on fusionne les deux classes les moins dissimilaires (les moins distantes, si on est dans un espace euclidien). On itère l'algorithme depuis la situation initiale jusqu'à la constitution d'une classe unique contenant tous les individus.
La CAH est l'ensemble de ces classifications successives. Il revient à l'utilisateur de décider où « couper » ses classes dans l'arbre hiérarchique de toutes les classifications, représenté par un dendogramme.

La classification ascendante hiérarchique (CAH)



La classification ascendante hiérarchique (CAH)