Les Classifications automatiques

Classifier consiste à identifier des groupes d'individus et ce sur plusieurs dimensions. Il faut pour cela utiliser des méthodes spécifiques de classification, qui peuvent soit, *à posteriori*, intervenir après une analyse de type ACP (ou autre : AFC, etc.), soit *à priori* sans qu'il soit nécessaire ou souhaitable d'effectuer une analyse multivariée au préalable.

Parmi les nombreuses méthodes de classification, deux ont la préférence des biologistes : l'algorithme des centres mobiles (*K-means = clustering* en anglais) et les classifications hiérarchiques ascendantes (*hierarchical clustering ou méthode UPGMA*[1]). Il s'agit d'outils intégrés au "machine learning", parfois aussi à la "fouille de données" (*data mining*) qui jouent des rôles importants dans presque tous les domaines scientifiques (marketing, génétique et biologie *s.l.*, psychologie, sociologie, informatique (reconnaissance de forme), linguistique...).

Distances (ou dissimilarités) entre individus d'une population

Pour regrouper les individus qui se ressemblent (et bien séparer ceux qui ne se ressemblent pas), il faut choisir un "critère de ressemblance". Pour cela, on examine l'ensemble des informations dont on dispose concernant les individus (pression artérielle, température, taux de métabolisme, ... par exemple s'il s'agit de malades) notées (x_i, y_i, . . .) pour le *ième* individu, et on imagine que chaque individu est un point $M_i = (x_i, y_i, z_i, \ldots)$ de l'espace. S'il n'y a que deux variables relevées (x_i, y_i) on obtient ainsi un nuage Γ de points dans le plan usuel, chaque point M_i ayant pour coordonnées (x_i, y_i). Ce nuage $\Gamma = \{M_i,\ i = 1,\ \ldots\ ,\ n\}$ contient n points. La *distance euclidienne* (car nous nous situons dans la géométrie éponyme) entre deux individus M_i et M_j est, *par définition* (voir le théorème de Pythagore) :

$$d_2(M_i, M_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Elle est d'autant plus petite que les deux individus sont semblables (du point de vue des valeurs des deux critères retenus) et d'autant plus grande qu'ils sont différents.

On peut associer à chaque nuage d'individus une matrice $D = (d_{ij})_{0\le i\le n, 0\le j\le n} = d_2(M_i, M_j)$, dite *matrice des distances*. C'est une matrice à n lignes et n colonnes, à coefficients positifs, symétrique (puisque $d_2(M_i, M_j) = d_2(M_j, M_i)$) et nulle sur la diagonale (puisque $d_2(M_i, M_i) = 0$). Pour un nuage d'effectif n, il y a donc $\frac{n(n-1)}{2}$ distances à calculer. A coté de la distance euclidienne, on peut définir d'autres distances. Par exemple :

$d_1(M_i, M_j) = |x_i - x_j| + |y_i - y_j|$

$d_\infty(M_i, M_j) = \text{Max}\ \{|x_i - x_j|, |y_i - y_j|\}$

$d_{ward} = \frac{p_i p_j}{p_i + p_j} d_2(M_i, M_j)^2$, où p_i et p_j sont les poids respectifs de M_i et M_j

De la même façon que l'on calcule des distances entre points du plan, on peut calculer des distances entre des groupes (ou classes) de points. Supposons que le nuage $\Gamma = \{M_i,\ i = 1,\ \ldots\ ,\ n\}$ est composé de plusieurs classes $\Gamma_1, \Gamma_2, \ldots, \Gamma_n$. Pour mesurer la distance entre les deux classes Γ_l et Γ_m, il existe plusieurs façons de procéder. L'une des plus utilisée est la *distance au plus proche voisin*

$d(\Gamma_l, \Gamma_m) = \text{Min}_{x\in\Gamma_l,\ y\in\Gamma_m} d(x, y)$.

Une alternative est de calculer la distance euclidienne séparant *les centres de gravité* des deux classes. C'est ce que nous allons utiliser dans la méthode suivante.

[1] ***U**nweighted **p**air **g**roup **m**ethod with **a**rithmetic mean.* On préfère depuis peu, surtout en phylogénie, l'algorithme de *Neighbour Joining* ou de *Maximum de Vraisemblance* ou des A*lgorithmes Bayésiens.* Cet algorithme reste cependant utilisé dans le cadre de l'alignement de séquences.

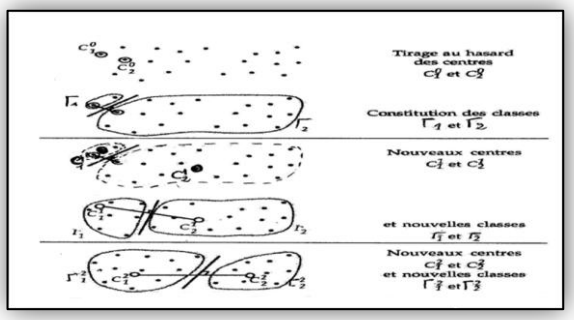
Méthodes des centres mobiles (= nuées dynamiques) ou k-means (Forgey (1965).

1. ALGORITHME

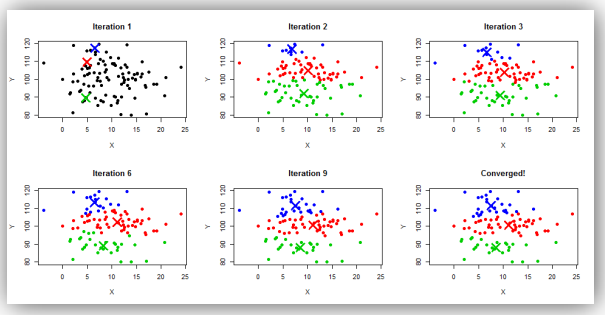
On représente les individus comme des points de l'espace ayant pour coordonnées des mesures (≥ 2 dim.) On cherche à regrouper autant que possible les individus les plus semblables (du point de vue des mesures que l'on possède) tout en séparant au mieux les classes ainsi formées. On choisit de procéder *de façon automatique*, c'est-à-dire un moyen de *faire apparaître*, uniquement à partir des mesures, des ressemblances et des différences à priori peu visibles. Cette idée, travailler automatiquement, à l'aide de l'ordinateur et *en aveugle*, est appelée *classification non supervisée*. L'inconvénient de cette méthode est qu'elle ne permet pas de découvrir quel peut être un nombre cohérent de classes, ni de visualiser la proximité entre les classes ou les objets. La méthode des centres mobiles s'applique lorsque l'on sait à l'avance combien de classes on veut obtenir. Appelons *k* ce nombre de classes. L'algorithme est le suivant, ayant choisi une méthode de calcul de distance entre individus :

Etape 0. Initialisation des classes : Pour initialiser, on tire au hasard *k* individus appartenant à la population, $C_1^0, C_2^0, \dots C_k^0$. L'indice indique qu'il s'agit des *k* centres initiaux.
Etape 1. Constitution de classes : On répartit l'ensemble des individus en *k* classes $\Gamma_1^0, \Gamma_2^0, \dots, \Gamma_k^0$ en regroupant autour de chaque centre C_i^0 $i = 1, \dots, n$, l'ensemble des individus qui sont plus proches du centre C_i^0 que des autres centres C_j^0 pour $j \neq i$ (au sens de la distance choisie).
Etape 2. Calcul des nouveaux centres : On détermine les centres de gravité G_1, G_2, ... , G_k des *k* classes ainsi obtenues et on désigne ces points comme les nouveaux centres $C_1^1 = G_1, C_2^1 = G_2, \dots, C_k^1 = G_k$.
Etape 3. Répétition des étapes 1 et 2 : on répète ces deux étapes jusqu'à la stabilisation de l'algorithme, c'est-à-dire jusqu'à ce que le découpage en classes obtenu ne soit (presque) plus modifié par une itération supplémentaire.

Le schéma ci-dessous illustre la méthode. Dans cette figure la distance choisie est la *distance euclidienne*. En effet, pour répartir les points du nuage en deux groupes, ceux qui sont les plus proches d'un point C_1^0 et ceux qui sont les plus proches d'un autre point C_2^0 au sens de la distance euclidienne, il suffit de tracer la médiatrice du segment $[C_1^0, C_2^0]$.



D'après F. Diener, Cours de Mathématiques pour la Biologie



D'après Satyasheel and Shreyak Tiwari (2016)
Royal Holloway University London

Mais est-on sûr que cet algorithme conduit bien à une partition *meilleure* que celle qui était issue du tirage aléatoire initial des *k* centres ? Pour répondre à cette question, il faudrait préciser ce que l'on entend par « meilleure ».

Nous allons pour cela introduire la notion d'*inertie* d'un nuage de points.

2 INERTIE INTERCLASSE ET INTRACLASSE

On appelle *inertie totale* d'un nuage $\Gamma = \{M_i, i = 1, \dots, n\}$ la somme pondérée des carrés des distances de ses points au centre de gravité du nuage. Donc, si *G* désigne le centre de gravité de Γ, l'inertie totale de Γ est, si tous les points du nuage sont de même poids égal à 1/*n* , donnée par la formule :

$$I(\Gamma) = (1/n)\ (\ d_2(M_1, G)^2 + d_2(M_2, G)^2 + \dots + d_2(M_n, G)^2\ ) \qquad (1)$$

Notons que le centre de gravité est précisément le point *G* pour lequel cette somme pondérée est minimale. L'inertie "mesure" la dispersion du nuage, elle sera grande pour un nuage très dispersé et petite lorsque le nuage est constitué de points bien regroupés. Si le nuage Γ est composé de *k* classes $\Gamma_1, \Gamma_2, \dots, \Gamma_k$ (disjointes deux à deux), celles-ci seront d'autant plus

homogènes que les inerties de chaque classe, $I(\Gamma_1)$, $I(\Gamma_2)$, ... , $I(\Gamma_k)$, calculées par rapport à leurs centres de gravité G_1, G_2, ... , G_k respectifs, seront faibles. La somme de ces inerties est appelée *inertie intraclasse* (I_{intra}) :

$$I_{intra} = I(\Gamma_1) + I(\Gamma_2) + \ldots + I(\Gamma_k)$$

Les inerties des classes $I(\Gamma_1)$, $I(\Gamma_2)$, ... sont simplement calculées avec la formule (1) ci-dessus où l'on remplace le centre de gravité G par celui de la classe G_1, G_2, ... et le poids 1/n par celui de la classe. L'inertie totale d'un nuage n'est généralement pas égale à la somme des inerties des classes qui le composent, c'est-à-dire à l'inertie intraclasse (sauf dans le cas où les centres de gravité de toutes les classes sont confondus) car il faut prendre en compte également la dispersion des classes par rapport au centre de gravité du nuage. Il s'agit de l'*inertie interclasse* (I_{inter}) définie par

$$I_{inter} = p_1 d_2(G_1, G)^2 + p_2 d_2(G_2, G)^2 + \ldots + p_k d_2(G_k, G)^2 ,$$ où p_j désigne le poids total de la classe Γ_j.

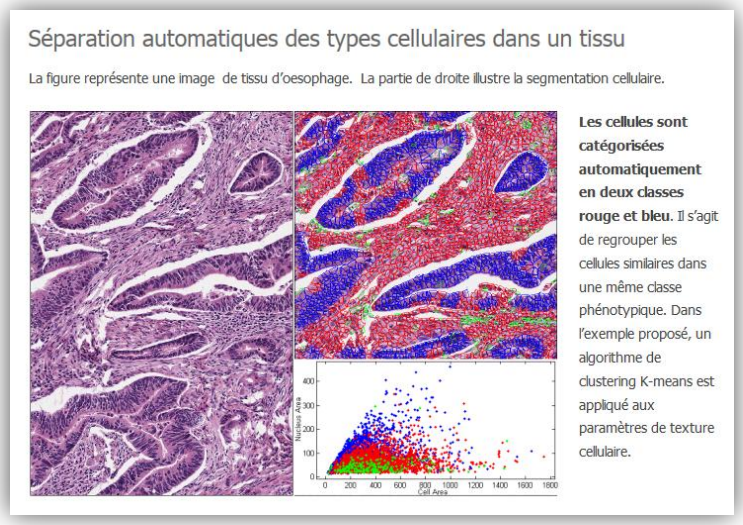
On montre le résultat suivant appelé *décomposition de Huygens :*

Théorème 1 *L'inertie totale d'un nuage de points* $I(\Gamma)$ *composé de différentes classes disjointes deux à deux est la somme de son inertie intraclasse et de son inertie interclasse* :

$$I(\Gamma) = I(\Gamma_1 \cup \Gamma_2 \cup \ldots \cup \Gamma_k) = I_{intra} + I_{inter}.$$

On a vu que lorsqu'un nuage est composé de plusieurs classes, si chacune est très bien regroupée autour de son centre de gravité, son inertie *intra*, qui est la somme des inertie de chaque classe sera petite. La partition d'un nuage en k classes, pour un nombre de classes fixé, sera d'autant *meilleure* que son inertie *intra* sera petite, ou, puisque son inertie totale reste la même quelque soit la partition, que son inertie *inter* est grande. Or on peut montrer justement que l'inertie *intra* classe ne peut que décroître lorsque l'on passe d'un regroupement en classes $\{\Gamma_1^i, \Gamma_2^i, \ldots, \Gamma_k^i\}$ au suivant $\{\Gamma_1^{i+1}, \Gamma_2^{i+1}, \ldots, \Gamma_k^{i+1}\}$ par une itération de l'algorithme des centres mobiles. Si cette décroissance était toujours stricte, puisque le nombre de partitions différentes d'un ensemble fini de points est lui-même fini (même s'il est gigantesque), on serait certain d'atteindre ainsi le minimum de l'*intra*.

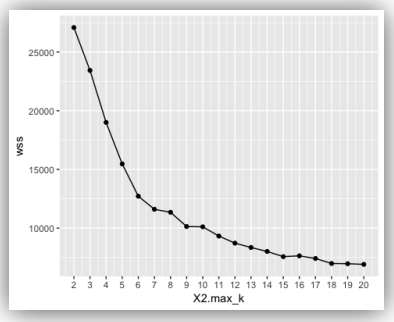
En pratique, la décroissance n'est pas toujours stricte et on n'est donc sûr de rien ! Mais cet algorithme est populaire car il est facile à utiliser et il suffit souvent de peu d'itérations pour avoir déjà une partition de qualité (proche de l'optimum).



© QUANTACELL, France.

3. Comment choisir K ?

Une technique pour choisir le meilleur k (lorsqu'on l'ignore) s'appelle la méthode du coude (elbow). Cette méthode utilise l'hétérogénéité au sein du groupe pour évaluer la variabilité (cluster$tot.withinss). En d'autres termes, vous êtes intéressé par le *pourcentage de la variance expliquée par chaque groupe*. On s'attend à ce que l'hétérogénéité diminue. Le défi est de trouver le k qui dépasse les rendements décroissants : à ce point, l'ajout d'un nouveau cluster n'améliore pas la variabilité des données car il reste très peu d'informations à expliquer. On effectue une série d'applications de kmeans() en faisant largement varier k. On récupère les valeurs de cluster$tot.withinss puis on trace les "withinss" en fonction de k (on peut aussi travailler sur la variabilité inter cluster$betweenss) :



Dans cet exemple, le K optimal est vraisemblablement 7 car au delà la variabilité diminue beaucoup plus lentement.
Un article très détaillé avec de nombreuses améliorations : https://www.guru99.com/r-k-means-clustering.html
Exercice : voir k-means.R

Méthode de la Classification Hiérarchique Ascendante (= UPGMA) (Sneath, 1957).

Pour classifier une population d'effectif *n* dont les individus sont numérotés 1, 2, ..., on considère cette population *comme la réunion de n classes à un seul élément* et on regroupe progressivement les classes deux à deux selon l'algorithme suivant :

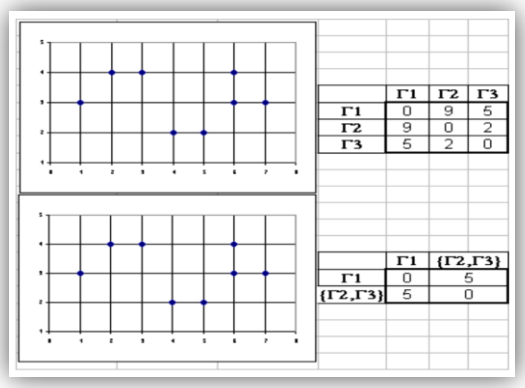
Etape 0 Calculer la matrice des distances **D** = (d*ij*) $_{0 < i \leq n,\ 0 < j \leq n}$
Etape 1 Remplacer les deux individus de distance minimale par une classe (à 2 éléments) Γ_1. La population compte alors *n*-1 classes (*n* - 2 classes à un élément et une à 2 éléments).
Etape 2 Calculer la perte d'inertie interclasse dû au regroupement précédent : on peut montrer qu'il s'agit exactement de l'écart de Ward des deux individus regroupés. On peut donc recommencer à l'étape 0 et remplacer "individus" par "classes" si nécessaire (et donc "distance entre individus" par "écarts entre classes"). On appelle cela l'*agglomération au plus proche voisin* . Aprés n-1 itérations, tous les individus sont regroupés en une classe unique.

On construit alors un arbre, appelé *dendrogramme*, de la façon suivante. On aligne sur l'axe horizontal des points représentant les différents individus et on les joint deux à deux, successivement, en suivant cet algorithme de classification hiérarchique ascendante (commençant par les plus proches, etc.). On poursuit ainsi jusqu'à regroupement de tous les individus en une classe unique. Pour plus de lisibilité, on pourra disposer les individus dans l'ordre dans lequel les regroupements ont été effectués. Le niveau (hauteur) de chaque nœud de l'arbre est, par exemple, choisi *proportionnel à la distance des deux classes regroupées*. On cherche ensuite à couper le dendrogramme au niveau où cela créé la meilleure répartition des points du nuage en classes bien distinctes entre elles. On peut comprendre qu'il ne sera pas optimal découper le dendrogramme à un niveau où le regroupement s'est fait entre deux classes assez proches mais qu'au contraire on cherche à couper là où les classes regroupées étaient les plus éloignées.

Exemple : Voici par exemple une étape dans la classification d'un nuage de 8 points du plan ayant pour coordonnées (1, 3), (2, 4), (3, 4), (4, 2), (5, 2), (6, 3), (6, 4), (7, 3). On les suppose déjà regroupés en trois classes :

Γ_1 = {(1, 3) ; (2, 4) ; (3, 4)}, Γ_2 = {(4, 2) ; (5, 2)} et Γ_3 = {(6, 3) ; (6, 4) ; (7, 3)}.

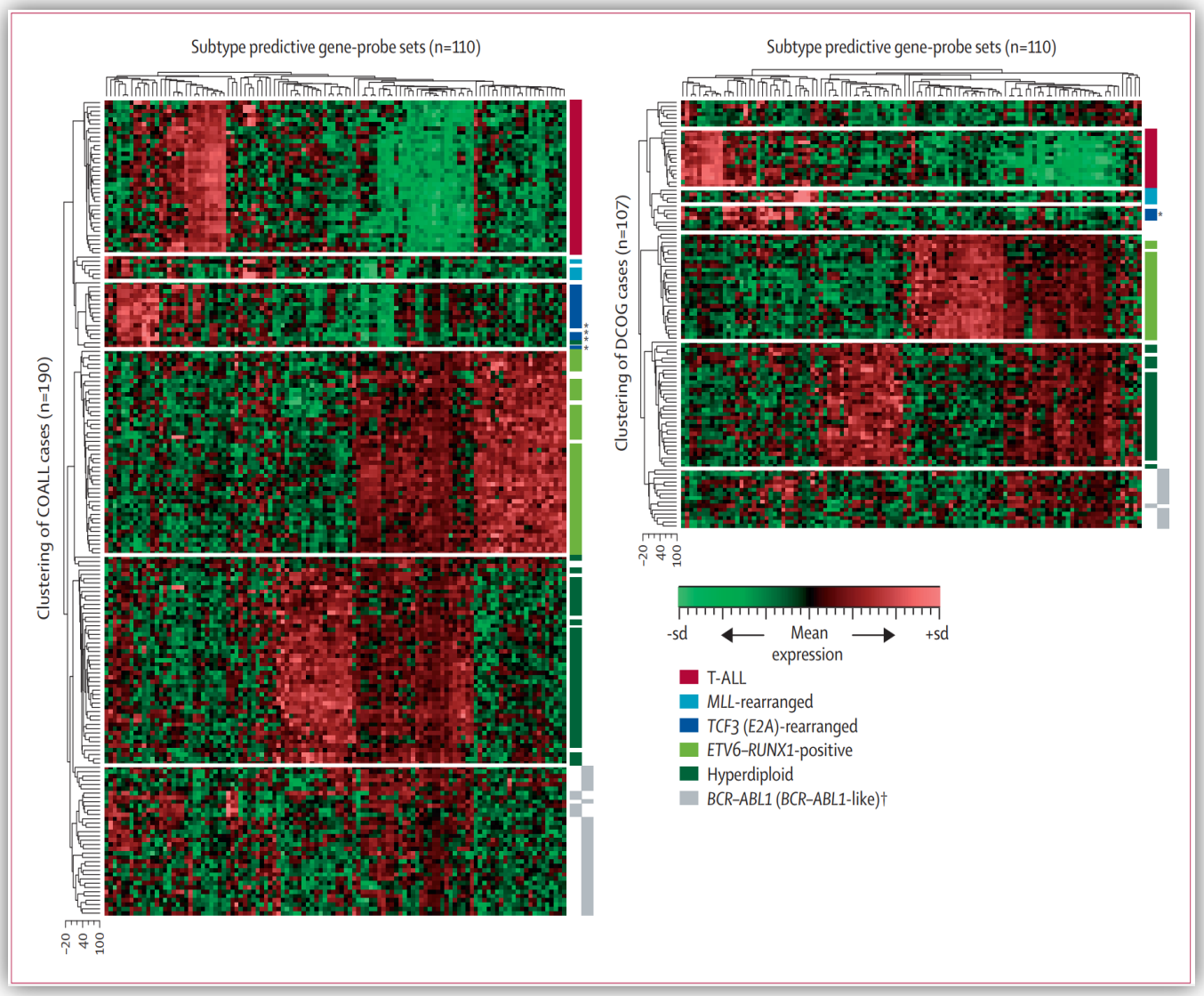
Le calcul des distances entre les trois classes (distance entre les points les plus proches) montre que les plus proches sont Γ_2 et Γ_3. On les agglomère donc à l'étape suivante puis on calcule la distance entre la nouvelle classe ainsi formée {Γ_2, Γ_3} et la classe restante Γ_1.



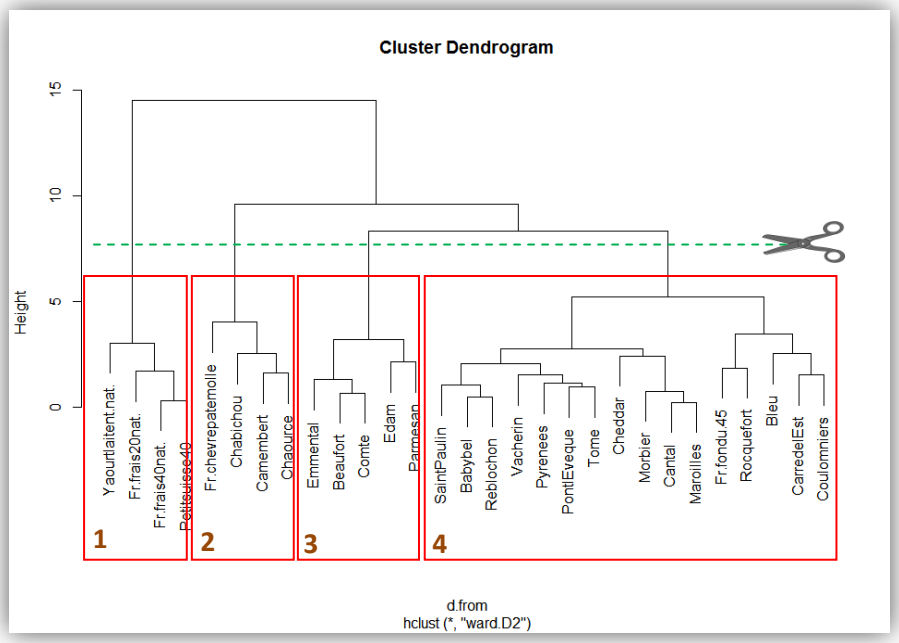
	Γ1	Γ2	Γ3
Γ1	0	9	5
Γ2	9	0	2
Γ3	5	2	0

	Γ1	{Γ2,Γ3}
Γ1	0	5
{Γ2,Γ3}	5	0

Après avoir ainsi aggloméré l'ensemble des classes en une seule, on peut tracer le dendrogramme résumant les agglomérations successives, en choisissant de mettre en ordonnée la distance des deux classes regroupées.



Exemple. Clustering of ALL subtypes by gene-expression profiles. Hierarchical clustering of patients from the COALL (left) and DCOG (right) studies with 110 gene-probe sets selected to classify paediatric ALL. Heat map shows which gene-probe sets are over-expressed (in red) and which gene probe sets are under-expressed (in green) relative to mean expression of all gene-probe sets (see scale bar). *Patients with *E2A*-rearranged subclone (15–26% positive cells). †Right column of grey bar denotes *BCR–ABL1*-like cases. *In* : A subtype of childhood acute lymphoblastic leukaemia with poor treatment outcome: a genome-wide classification study. Monique L Den Boer *et al*., *The Lancet Oncol* (2009), 10:125-134.



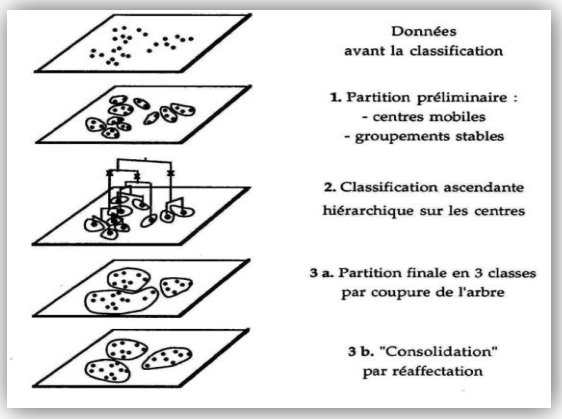
Exercice (cah.R). On coupe l'arbre à un niveau non trivial (ni trop haut ni trop bas) à un saut d'agrégation le plus grand possible. Remarquez que si l'on coupait le dendrogramme au plus haut, on obtiendrait une partition en deux groupes : les fromages frais et les autres. On distingue les fromages frais (1), à pâte molle (2), à pâte dure (3). Le 4ème groupe rassemble un peu tous les types.

3 METHODES MIXTES

L'algorithme des centres mobiles a cependant deux défauts principaux :

1) Il exige de l'utilisateur de choisir à l'avance le nombre de classes de la partition, ce qui est parfois difficile.
2) Ensuite, on constate que la partition finale peut varier sensiblement en fonction du choix des centres initiaux. Cela vient du fait que, si l'inertie intra décroit effectivement à chaque itération, ce n'est pas forcément vers le minimum recherché mais parfois vers un *minimum local* qui n'est pas du tout optimal. En pratique, comme le déroulement de l'algorithme est généralement rapide, on n'hésite pas à l'exécuter plusieurs fois avec des choix différents des centres initiaux et on compare les partitions obtenues pour ne retenir que celle dont l'inertie intra est minimale, ou, si aucune n'est clairement minimale, la partition qui revient le plus souvent (groupements stables).

Au delà de la classification hiérarchique ascendante et de la méthode des centres mobiles, il existe beaucoup d'autres méthodes (par exemple des méthodes stochastiques comme les réseaux de neurones). Mais l'utilisateur privilégie souvent, lorsque le nombre d'individus est très grand et qu'il est alors difficile de choisir d'avance le nombre de classes (voir aussi la méthode du elbow), une classification mixte comme indiquées sur la figure suivante :



Si l'on a des milliers, voire des dizaines de milliers d'individus à classifier, on commence par les répartir en un (trop) grand nombre de classes (par exemple k = 100) par la méthode des centres mobiles. Puis, on ne retient que les centres des classes (avec leur poids qui sera proportionnel au nombre d'individus dans chaque classe) $\{(C_1^n, p1), (C_2^n, p2), \dots \{(C_{100}^n, p100)\}$ et on effectue une classification hiérarchique ascendante *sur ces centres*. Une partition est alors obtenue par coupure du dendrogramme que l'on choisit aussi judicieusement que possible (par exemple *au plus grand saut*) pour avoir le *bon* nombre de classes. On peut alors calculer leurs centres de gravité et finalement allouer chaque individu au centre le plus proche, ce qui *consolide* la partition.