# La classification automatique

Le but de la classification, ou ***typologie***, encore appelée ***taxinomie*** ou ***taxonomie***, est de classer et regrouper les ensembles d'objets en sous-ensembles homogènes.

## La classification ascendante hiérarchique (CAH)

La ***classification ascendante hiérarchique*** (ou ***CAH***), présentée ici, procède par agrégations successives à partir des objets isolés; elle produit des classifications complètes, représentées par des arbres comparables à ceux des botanistes ou des biologistes (procaryotes, eucaryotes, invertébrés, vertébrés, oiseaux, reptiles, mammifères, etc.)

### Les tableaux de distances

La CAH traite les tableaux carrés donnant les distances mutuelles entre les objets d'un ensemble. Il peut s'agir d'un ***indice de dissimilarité***, donné à priori entre les objets, ou en un sens plus géométrique, d'une ***distance*** calculée entre les individus sur un tableau rectangulaire, tels ceux que traitent l'ACP ou l'AFC.

### Exemples de tableaux de distances

- Indice de dissimilarité entre partis politiques, obtenu en demandant à 100 personnes de noter de 1 à 10 l'éloignement entre différents partis.
- Distance "usuelle" entre les pays de l'Europe des 27, calculée sur un tableau quantitatif relevant un ensemble de variables socio-démographiques pour ces différents pays.
- Distance du chi-deux entre les départements calculée sur le tableau de contingence donnant les voix par candidat au premier tour des élections présidentielles de 2007.

### La méthode

Outre la distance entre les objets, la méthode nécessite que soit définie une distance entre groupes; parmi différentes possibilités, les plus fréquemment utilisées sont:

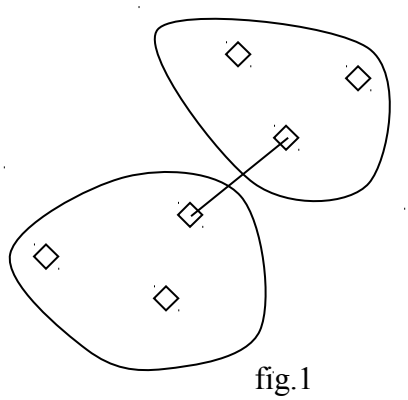


fig.1

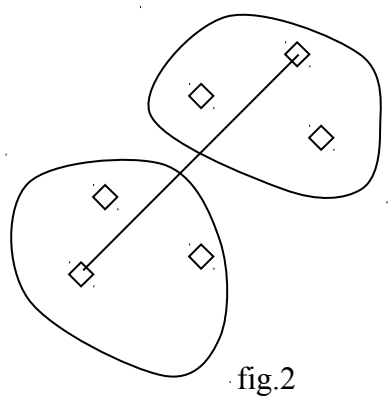


fig.2

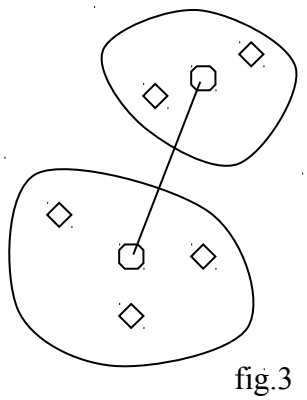


fig.3

- La plus petite distance entre les deux groupes (fig.1).
- La plus grande distance entre les deux groupes (fig.2).

- La distance des centres de gravité (éventuellement pondérés), dans le cas d'une distance géométrique calculée sur un tableau rectangulaire (fig.3).
- La moyenne des distances pour tous les couples d'objets entre les deux groupes.
- La variance du groupe obtenu par la fusion des deux groupes (à nouveau dans le cas d'une distance géométrique).

**Algorithme**

On considère au départ la partition de l'ensemble en tous les groupes individuels composés d'un seul élément, et on regroupe à la première étape les deux éléments les plus proches.

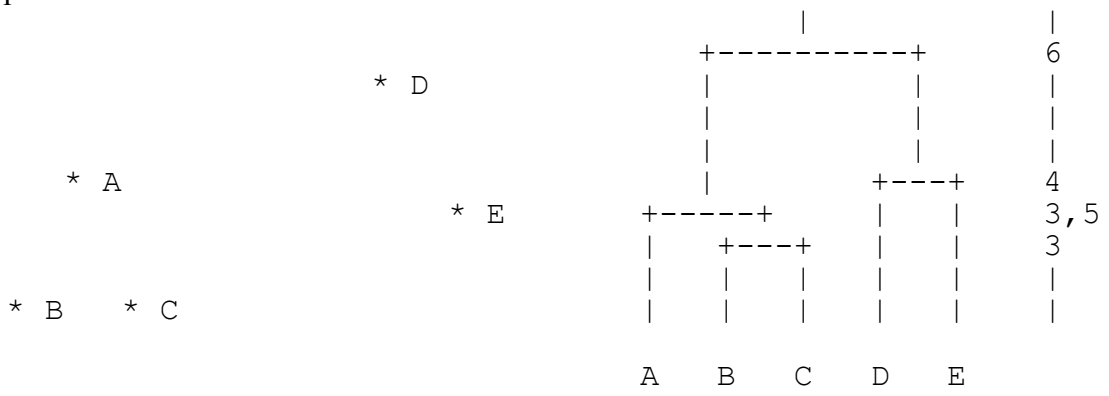
On itère le procédé, en regroupant à une étape donnée les deux groupes les plus proches, leur distance étant le niveau de cette agrégation.

L'algorithme est terminé quand on est parvenu à une seule classe qui regroupe la totalité des objets.

**Hiérarchies, partitions**

La suite de partitions "emboîtées" résultant d'une CAH est une ***hiérarchie de parties***, il est commode de la représenter par un ***arbre indicé***, chaque ***nœud*** portant l'indication de son niveau d'agrégation.

Exemple




Si on souhaite une partition en un nombre donné de classes, il suffit de tronquer l'arbre de la hiérarchie au niveau convenable.

## Recherche directe de partitions

Si, comme on vient de le voir, la CAH peut fournir des partitions d'effectif désiré par simple troncature de l'arborescence obtenue, on peut estimer le procédé coûteux, surtout en cas de données très abondantes.

**Relation fondamentale, critère**

On se limite désormais à des objets figurés par des points dans un espace vectoriel de dimension convenable muni de la distance usuelle, et on considère une partition donnée *a priori* de ces objets en p classes. On prend les notations suivantes:

- $y_{ik}$ est le k-ième objet de la classe i
- $n_i$ est l'effectif de la classe i
- $y_{i.}$ est le centre de gravité de la classe i
- $y_{..}$ est le centre de gravité de l'ensemble

Dans ces conditions, on a la relation:

$$\sum_{i,k} (y_{ik} - y_{..})^2 = \sum_i \sum_k (y_{ik} - y_{i.})^2 + \sum_i n_{i.}(y_{i.} - y_{..})^2$$

qu'on énonce encore

Variation Totale = Variation ***Intraclasse*** + Variation ***Interclasse***

ou

T = W + B (pour "Total", "*Within*" et "*Between*").

Une typologie est d'autant meilleure que les classes sont ramassées (faible variation intraclasse) et qu'elles sont nettement séparées (forte variation interclasse).

Leur somme étant constante, et égale à la variation totale qui ne dépend pas de la partition, il est équivalent de minimiser la variation intraclasse ou de maximiser la variation interclasse. Pour un nombre de classes donné, il n'est malheureusement aucunement garanti que la partition obtenue par troncature de la CAH optimise ce critère...

Une grande variété d'indicateurs ont été définis pour examiner les partitions, certains dérivent de la relation fondamentale.

La variance interclasse B peut ainsi se décomposer selon les différents axes de coordonnées (théorème de Pythagore):

$$B = \sum B_j$$

ce qui permet d'estimer l'influence de l'axe j par le ratio:

$$c(j) = B_j / B$$

**Méthodes de type « nuées dynamiques »**

Le critère précédent inciterait, pour un nombre de classes donné, à examiner toutes les partitions possibles. L'*explosion combinatoire* du nombre de cas à traiter rend une telle procédure impraticable.

Une famille de méthodes itératives, sélectionnant alternativement partitions et représentants privilégiés, ont l'ambition de parvenir rapidement à un résultat. Ces algorithmes supposent fixé a priori le nombre de classes désiré.

**Méthode des centres mobiles**

On se donne une partition de départ (éventuellement choisie en fonction d'idées préconçues), et chaque classe est représentée par son centre de gravité. On construit alors une nouvelle partition en affectant chaque point au plus proche des centres précédents, et on itère le procédé... Le nombre de classes peut décroître en chemin, et on s'arrête rapidement sur une partition stable.

**Méthodes des nuées dynamiques**

On représente cette fois les classes non plus par leur centre de gravité, mais par certains de leurs points (issus par exemple d'un sous-ensemble présélectionné de représentants privilégiés) appelés ***noyaux***, deux règles d'affectation permettent alors de construire itérativement la nouvelle partition... Sous des hypothèses simples, le procédé converge également assez vite vers une partition stable finale.

La faiblesse des méthodes précédentes tient au fait que la partition finale obtenue dépend non seulement de l'effectif demandé, mais aussi de la partition de départ choisie. Pour ces raisons, il est d'usage de multiplier les essais en faisant varier le nombre de classes demandé ainsi que la partition initiale.

## Conclusion

Plus simples encore dans leurs principes que les méthodes factorielles, les techniques de classification demandent toutefois beaucoup de calculs dès que les données atteignent une certaine taille. Elles n'ont pu se développer qu'avec l'apparition des ordinateurs.

On utilise en général concurremment ces deux types de méthodes.

----===ooOoo===----

19.11.2013

# Appendice mathématique

Démonstration de la relation fondamentale :

$$\sum_{i,k} (y_{ik} - y_{..})^2 = \sum_{i}\sum_{k} (y_{ik} - y_{i.})^2 + \sum_{i} n_i.(y_{i.} - y_{..})^2$$

On se limite tout d'abord au cas où les observations $y_{ik}$ sont simplement des éléments de **R** et on construit le vecteur y obtenu en les mettant bout-à-bout, on construit par ailleurs les dummy variables de même dimension indicatrices des différents groupes.

La relation peut être interprétée simplement comme la relation d'analyse de la variance associée à la régression de la variable y décrivant les observations $y_{ik}$ sur les p variables indicatrices des classes de la partition.

En effet, de même que la régression d'une série sur la seule constante donne sa moyenne (la somme des carrés des résidus en étant la variance, au facteur 1/n près), la valeur ajustée commune à toutes les observations d'un même groupe dans la régression précédente est la moyenne : $y_{i.}$ , de ce groupe.

Par suite la quantité $\sum n_i.(y_{i.} - y_{..})^2$ est la variation expliquée de la régression, alors que $\sum\sum (y_{ik} - y_{i.})^2$ en est la variation résiduelle, ce qui établit la propriété.

Le cas où les observations sont dans $\mathbf{R}^p$ se traite simplement en ajoutant les relations telles que la précédente, satisfaites composante par composante.

----===ooOoo===----

(20.03.2012)