

F. BENZÉCRI

## **Introduction à la classification automatique d'après un exemple de données médicales**

*Les cahiers de l'analyse des données*, tome 5, n° 3 (1980),  
p. 311-340

[http://www.numdam.org/item?id=CAD\\_1980\\_\\_5\\_3\\_311\\_0](http://www.numdam.org/item?id=CAD_1980__5_3_311_0)

© Les cahiers de l'analyse des données, Dunod, 1980, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme  
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

## INTRODUCTION A LA CLASSIFICATION AUTOMATIQUE D'APRÈS UN EXEMPLE DE DONNÉES MÉDICALES

[INT. CLASS. MED.]

par F. Benzécri (1)

N.B. Cet article fait suite à notre introduction à l'Analyse des Correspondances dont il utilise les principes géométriques et reprend l'exemple de base. Plus précisément, aux §§ 2, 3, 4, 5, nous supposons que le lecteur s'est familiarisé avec [INT. CORR. MED.] jusqu'au § 2.2.2 inclusivement, la fin du présent exposé faisant seule référence aux résultats d'analyse factorielle. Les §§ 2, 3, 4.1 et 5 peuvent être lus immédiatement après [INT. CORR. MED.] ; quant aux §§ 4.1.2 et 4.2 relatifs aux aides à l'interprétation, sans faire appel à des notions nouvelles, ces §§ sont d'une lecture plus difficile et s'adressent plutôt à des lecteurs qui ont la pratique de l'analyse des correspondances et notamment des calculs de contributions.

### 1 Qu'est-ce qu'une classification

1.1 Partition et hiérarchie des classes : Dans l'usage courant, le terme de classification est d'abord employé pour désigner le partage d'un ensemble d'individus en classes de telle sorte que tout individu appartienne à une classe et une seule (et qu'il n'y ait pas de classe vide !) : les mathématiciens appellent *partition* un tel système de classes. Mais le terme de classification sert aussi à désigner des systèmes emboîtés de classes, comme on en voit en sciences naturelles : les êtres vivants sont partagés en deux grands règnes, animal et végétal ; chacun de ces deux règnes est lui-même subdivisé ; les animaux sont partagés en vertébrés, arthropodes, mollusques... ; les vertébrés sont à leur tour subdivisés en mammifères, oiseaux, reptiles, batraciens et poissons, etc. . On parle alors de *classification hiérarchique*. On peut schématiser ainsi partition et classification hiérarchique, en considérant par exemple sept individus à classer :

---

(1) Docteur ès-sciences

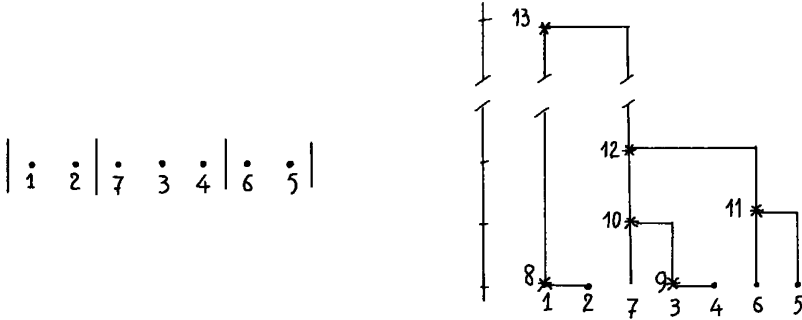


Figure 1 : partition et arbre hiérarchique .

à gauche, on a une partition de l'ensemble des sept individus en trois classes ; à droite un arbre de classification hiérarchique comprenant à sa base les sept individus à classer et, partant de ces individus , des branches se raccordant entre elles par des noeuds marqués par des astérisques et portant un numéro. Des noeuds partent à nouveau des branches qui se raccordent entre elles par d'autres noeuds, jusqu'au sommet, auquel est suspendu l'arbre tout entier. Un tel arbre définit un système emboîté de classes. En effet du sommet 13 partent deux branches ; à l'une, (A), se rattachent les individus 1 et 2 ; à l'autre, (B), les individus {7, 3, 4, 5}. Donc les deux branches A et B partant du sommet 13 définissent deux grandes classes. Mais chacune de ces classes se subdivise. La branche A, au noeud 8, se divise en deux branches reliées chacune directement (i.e. sans noeud intermédiaire) à un individu : l'une à l'individu 1, l'autre à 2. La branche B, au noeud 12, se divise en deux branches : A' tenant les individus 7, 3, 4 et B' tenant les individus 6 et 5. Donc la grande classe {7, 3, 4, 6, 5} suspendue à la branche B se divise en deux sous-classes {7, 3, 4} et {6, 5} au niveau du noeud 12, etc. .

A gauche de l'arbre, une échelle indique le *niveau* de chaque noeud ou le degré de généralité de la classe qu'il définit (de même qu'en sciences naturelles un ordre, e.g. les carnivores, est plus général qu'une espèce, e.g. le chat, nous dirons que le noeud 12 est d'un niveau plus élevé que le noeud 8). En coupant l'arbre à un niveau donné, on aura une partition de l'ensemble des sept individus en classes plus ou moins fines, selon le niveau. Par exemple en coupant sous le noeud 12, on obtient la partition de gauche en 3 classes {1, 2} , {7, 3, 4} et {6, 5}.

Cette latitude laissée au spécialiste de choisir lui-même le degré de généralité qui l'intéresse est un avantage de la classification hiérarchique produisant des arbres tels que celui que l'on a figuré ici.

On appelle cet arbre *hiérarchique* parce qu'un individu est rattaché au noeud le plus haut par l'intermédiaire de noeuds qui se succèdent le long d'une ligne continue de l'arbre à des niveaux de plus en plus élevés : e.g. le noeud 6 est relié au sommet 13 par l'intermédiaire des neuds 11 et 12.

Au contraire, dans une partition, que l'on peut aussi figurer comme un arbre (fig. 2), chaque individu est relié au sommet par un seul noeud intermédiaire :

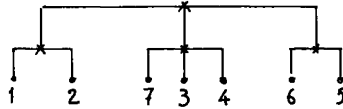


Fig 2: partition figurée par un arbre

Remarque : Dans l'arbre hiérarchique de la fig. 1, de chaque noeud partent exactement deux branches ; on dit que c'est un arbre *binnaire*. Ce n'est là qu'un effet de construction de l'algorithme qui a produit cet arbre : parmi tous les objets à classer, on groupe par paires ceux qui se ressemblent le plus puis on adjoint à une paire déjà formée un troisième objet ou une autre paire etc. et chaque opération d'agrégation produit un noeud. Mais quand on coupe l'arbre à un niveau donné, la partition obtenue ne se ressent pas du caractère binaire de l'arbre, puisqu'on ne tient pas compte des subdivisions opérées à un niveau inférieur à celui auquel on coupe l'arbre.

1.2 Qualités d'une classification : Une partition n'est intéressante que dans la mesure où les classes sont nettement individualisées, c'est-à-dire forment chacune un tout bien caractérisé et d'autre part sont bien distinctes les unes des autres. Par exemple, s'il s'agit de classer des objets étalés sur une table d'après leur forme ronde, ovale ou triangulaire, on réalisera trois classes en rapprochant les objets de même forme et en éloignant les classes les unes des autres :

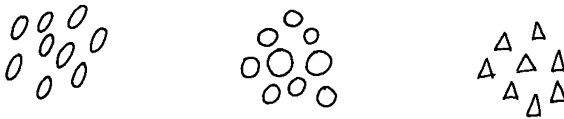


Figure 3

Dans la pratique, un ensemble d'individus se prête plus ou moins à la classification ; e.g. en cytologie, un lot de cellules peut présenter tout un continuum de formes depuis A (cytoplasme peu étendu et noyau volumineux) jusqu'à D (petit noyau, cytoplasme étendu):

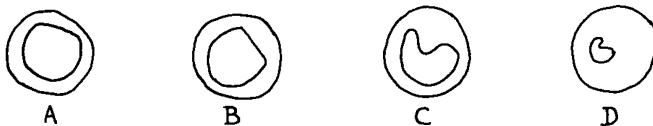


Figure 4

l'analyse des correspondances mettra en évidence à la fois les types et la dispersion des objets autour de ces types, mais aucune procédure automatique ne tranchera de façon satisfaisante au sein de l'ensemble pour définir des classes nettes.

On peut encore imaginer des situations où l'on ait le choix entre plusieurs façons de grouper les individus, comme nous le figurons ici en représentant la ressemblance par la proximité sur le graphique ; c'est en effet à ce mode de ressemblance qu'on se ramène toujours en analyse des données puisque les descriptions des objets à classer, quel qu'en soit le format initial, sont traduites en des représentations géométriques où à chaque individu est associé un point).

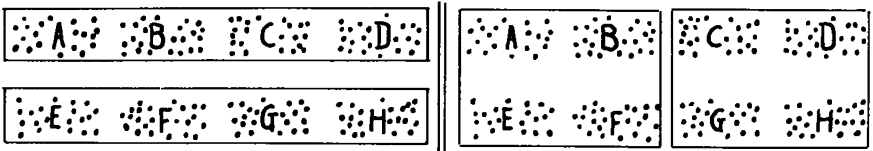


Fig.5. on a suggéré par des cadres les deux façons de grouper les classes A, B, C, D, E, F, G, H.

Une fois reconnues les classes A, B, C, D, E, F, G, H, assemblera-t-on A, B, C, D en raison du peu d'espace qui sépare A et B, B et C, C et D ; et de même E, F, G, H, pour ne constituer finalement que deux grandes classes allongées ? Une telle partition ne serait pas satisfaisante parce qu'entre un objet de A et un objet de D (qui seraient placés au sein d'une même classe) la distance est plus grande qu'entre un objet de A et un objet de E (lesquels seraient mis dans deux classes distinctes). Au contraire, si l'on groupe A, B, E, F d'une part, et C, D, G, H d'autre part, les deux classes obtenues seront relativement compactes, mais mal séparées (la largeur de l'espace entre B et C par exemple étant moindre que celle entre B et F). Les deux exigences : compacité des classes et séparation des classes entre elles sont ici contradictoires.

De toute façon, on doit chercher un critère de ressemblance entre individus et entre classes qui réalise au mieux la compacité des classes et la séparation des classes entre elles. Mais il faut savoir qu'il est des cas désespérés.

### 1.3 Deux procédés pour établir une classification : ascendant et descendant

Quand on classe des objets sur une table d'après leur forme ronde, ovale ou triangulaire, on sait à l'avance d'après quel attribut les classer. Partant du tout, on sépare les objets ronds des autres ; d'où deux grandes classes : celle des objets ronds et celle de ceux qui ne sont pas ronds ; dans cette dernière, on sépare les objets ovales des objets triangulaires. C'est le procédé *descendant* de classification : on divise le tout en deux parties selon un attribut ; puis chaque partie est subdivisée selon un nouvel attribut etc. .

Mais procéder ainsi, par voie descendante, suppose que l'on soit assuré d'avoir reconnu les paramètres (ou les caractères) auxquels il faut recourir pour définir les divisions supérieures de la hiérarchie (i.e. qu'on ait une vue juste de ce que, depuis Jussieu, les botanistes appellent *hiérarchie des caractères*. Or les caractères les plus frappants (e.g. pour un végétal l'opposition entre arbre et plante annuelle herbacée) ne sont pas toujours ceux qui finalement jouent le rôle principal (e.g. l'ébène et le robinier - arbres - sont rangés avec la luzerne et le pois de senteur au sein d'une même famille de l'ordre des légumineuses).

Si donc on n'est pas assuré d'avoir reconnu les paramètres pertinents, on procède plus sûrement de façon *ascendante* en partant des individus et en constituant, par agrégations successives, des classes de plus en plus grandes, selon des critères où entrent en jeu, à la fois, tous les traits servant à décrire les objets. L'arbre hiérarchique une fois constitué, on pourra s'appliquer à reconnaître ce qui caractérise les diverses classes. Si les individus que l'on a classés constituent un échantillonnage d'une certaine population, tout individu nouveau appartenant à cette population pourra être affecté aux classes déjà définies sur l'échantillon d'après les attributs principaux que l'on aura reconnu caractériser ces classes.

## 2 Classification ascendante hiérarchique selon l'inertie

Dans ce §, nous expliquons le critère communément utilisé pour édicter une hiérarchie de classes à partir d'un ensemble d'individus décrits initialement par les lignes d'un tableau de correspondance (cf. [INT. CORR. MED.] § 1). Au § 3, on expliquera comment lire les résultats d'une classification présentés par l'ordinateur ; et particulièrement les aides à l'interprétation, lesquelles, inspirées des calculs de contribution de l'a. des cor., peuvent suggérer finalement une définition conceptuelle des classes produites.

2.1 Représentation géométrique ; distance entre individus ; centre de gravité : La donnée de base sera pour nous un tableau de correspondance. Reprenons donc l'exemple donné dans [INT. CORR. MED.] :

|              | <i>pénicilline</i> | <i>tifomycine</i> | <i>tétracycline</i> | <i>érythromyc.</i> | <i>tiophénicol</i> | <i>gentalline</i> |
|--------------|--------------------|-------------------|---------------------|--------------------|--------------------|-------------------|
| TYPHOIDE     | 0                  | 4                 | 0                   | 0                  | 2                  | 0                 |
| SALMONEL. D. | 0                  | 2                 | 0                   | 0                  | 1                  | 0                 |
| AR. ORL      | 8                  | 0                 | 5                   | 3                  | 0                  | 3                 |
| PENUMOPATH.  | 7                  | 0                 | 5                   | 2                  | 0                  | 3                 |
| MENINGITE    | 2                  | 2                 | 0                   | 0                  | 0                  | 1                 |
| AF. V. URIN. | 4                  | 0                 | 2                   | 0                  | 0                  | 6                 |
| STAPHYLOCOC. | 3                  | 0                 | 1                   | 3                  | 0                  | 0                 |

Rappelons que ce tableau est supposé issu d'une enquête (fictive) auprès de médecins à qui l'on demanderait d'indiquer pour chaque cas de TYPHOIDE, SALMONELLOSE DIGESTIVE, AFFECTION O.R.L., PNEUMOPATHIE, MENINGITE, AFFECTION DES VOIES URINAIRES, STAPHYLOCOCCIE, les médicaments utilisés. On recense alors, dans les réponses des médecins, combien de fois chacune des sept maladies citées se rencontre avec chacun des six médicaments : *pénicilline*, *tifomycine*, *tétracycline*, *érythromycine*, *tiophénicol*, *gentalline*.

On se propose d'effectuer une classification des maladies et une classification des médicaments. Cela peut se faire selon diverses voies, mais comme on l'a annoncé (§ 1.3), on suivra ici une voie parallèle à l'analyse des correspondances.

Soit, par exemple, à classer les maladies. Comme en a. des c., celles-ci sont définies par leurs *profils* sur l'ensemble des six médicaments, c'est-à-dire par leurs lignes dans le tableau de correspondance, rapportées chacune à son total :

$$\begin{aligned} \text{TYPHOIDE} &= \{0, 4/6, 0, 0, 2/6, 0\} && (\text{total} : 1) \\ \text{STAPHYLOCOCCIE} &= \{3/7, 0, 1/7, 3/7, 0, 0\} && (\text{total} : 1) \end{aligned}$$

Ces sept profils sont considérés comme des points d'un espace euclidien à 6 dimensions :

à 6 dimensions parce que chaque point est défini par 6 nombres ;

*euclidien* parce que dans cet espace, la formule donnant le carré de la distance de deux points est, à des coefficients près, la somme des carrés des différences des coordonnées de même rang caractérisant les deux points (et se trouve être ainsi l'analogue de la formule donnant le carré de la distance dans le plan de la géométrie élémentaire où est vérifié le théorème de Pythagore).

Par exemple, la distance entre TYPHOIDE et STAPHYLOCOCCIE est égale à :

$$\begin{aligned} &(69/24) (0 - (3/7))^2 + (69/8) ((4/6) - 0)^2 + (69/13) (0 - 1/7)^2 + (69/8) (0 - (3/7))^2 \\ &+ (69/3) ((2/6) - 0)^2 + (69/13) (0 - 0)^2 ; \end{aligned}$$

où les coefficients (69/24), (69/8) etc. sont ceux de la *distance distributionnelle*, dite encore distance du  $\chi^2$ , (cf [INT. CORR. MED.] § 2.2.2) calculés sur le tableau de correspondance (e.g. 69/24 = total du tableau divisé par le total de la colonne *pénicilline*).

Dans l'espace à 6 dim. l'ensemble des sept maladies constitue un *nuage* de points pesants :

chaque maladie a pour *masse*, comme en a. des c., le total de sa ligne dans le tableau de correspondance, divisé par le total de ce tableau ; par exemple TYPHOIDE a pour masse 6/69.

Le nuage admet un *centre de gravité* (cf [INT. CORR. MED.] § 2.1) défini par ses coordonnées sur les six médicaments ; e.g. la coordonnée sur *pénicilline* du centre de gravité s'obtient en prenant la moyenne, pondérée par les masses, des coordonnées sur *pénicilline* des sept maladies. Le résultat de ce calcul (cf [INT. CORR. MED.] § 2.1) n'est autre que le profil de la ligne de marge du tableau de correspondance, ligne formée par les totaux des six colonnes.

Toute partie du nuage, e.g. les deux maladies : MENINGITE et AFFECTION DES VOIES URINAIRES, munies de leurs masses, admet de même un centre de gravité qui s'obtient en faisant la moyenne pondérée

coordonnée par coordonnée, des deux maladies, ce qui équivaut à prendre le profil de la ligne qui totalise les deux lignes du tableau de correspondance décrivant les deux maladies.

2.2 Inertie interclasse et inertie intraclasse d'une partition : Au § 1.2, nous avons défini la qualité d'une partition du double point de vue de la compacité des classes en elles-mêmes et de la séparation des classes entre elles. Quand les objets à classer sont les points pesants d'un nuage dans un espace multidimensionnel muni d'une métrique euclidienne, on peut mesurer la qualité d'une partition ainsi comprise par des formules mathématiques précises, objets du présent §.

Rappelons que l'inertie d'un point pesant par rapport à un centre est le produit de la masse de ce point par le carré de sa distance au centre et que l'inertie d'un ensemble de points pesants par rapport à un centre est la somme des inerties de ses points par rapport à ce centre (cf [INT. CORR. MED.] § 2.2).

Considérons une partition de l'ensemble des sept maladies représentées par leurs profils et munies de leurs masses dans l'espace euclidien à 6 dimensions (cf § 2.1) ; par exemple la partition en trois classes :

C1 = {TYPHOÏDE ; SALMONELLOSE}

C2 = {STAPHYLOCOCCIE ; AF. ORL ; PNEUMOPATHIE}

C3 = {AF. V. URINAIRES ; MENINGITE}

Chaque classe a une inertie par rapport à son propre centre de gravité qui mesure sa dispersion autour de son centre. Plus la classe est compacte, et plus son inertie par rapport à son centre est petite. Par exemple, la classe C1, qui est constituée de deux maladies ayant même profil, (cf INT. CORR. MED. § 1.5) puisque les deux lignes qui les définissent dans le tableau de correspondance sont proportionnelles, est représentée géométriquement par deux points confondus ; son centre de gravité est aussi confondu avec les deux points ; la distance au centre de gravité de chacun des deux points est nulle et l'inertie interne de la classe C1 est nulle. Donc à une compacité parfaite (tous les points de la classe confondus en un seul) correspond une inertie interne nulle.

Pour la partition, système de trois classes C1, C2, C3, on mesurera sa qualité quant à la compacité des classes par la somme des inerties des trois classes par rapport à leurs centres respectifs ; cette somme est appelée *inertie intraclasse* (ou inertie intérieure aux classes de la partition). Plus cette inertie intraclasse est petite, plus les classes sont, en un certain sens, compactes.

Pour mesurer la séparation des classes entre elles, on remplace chaque classe par son centre de gravité affecté de la masse totale de la classe : on obtient ainsi le nuage des centres des classes ; et le centre de gravité du nuage des centres coïncide avec le centre de gravité du nuage des maladies (dans la recherche du centre de gravité d'un ensemble de points pesants, on peut remplacer plusieurs points par leur centre de gravité affecté de la somme de leurs masses). L'inertie totale du nuage des centres des classes mesure la dispersion des centres autour du centre de gravité du nuage : c'est ce nombre unique que nous choisirons pour mesurer la séparation des classes entre elles. L'inertie totale du nuage des centres est appelée *inertie interclasse* (ou inertie entre les classes) de la partition. Plus cette inertie interclasse est grande, plus les classes sont, en un certain sens, séparées les unes des autres.



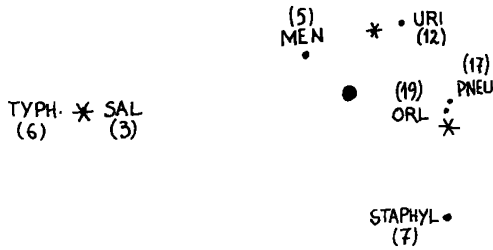


Figure 6 : Les maladies sont représentées par des points ; les centres des trois classes par des étoiles ; le centre de gravité du nuage des maladies, qui coïncide avec celui des trois centres, par un gros point ; on a indiqué entre parenthèses la masse de chaque maladie (comptée en soixante neuvièmes)

Il se trouve (nous ne le démontrerons pas ici) que la somme de l'inertie interclasse et de l'inertie intraclasse de toute partition d'un nuage de points est égale à l'inertie totale du nuage des points par rapport à son centre. Cette propriété a une conséquence remarquable : c'est que, traduites en termes d'inertie comme nous venons de le faire, les deux exigences : compacité interne des classes, séparation des classes entre elles, n'en font qu'une. En effet, rendre maxima la compacité des classes, c'est, pour nous, rendre minima l'inertie intraclasse ; et rendre maxima la séparation des classes, c'est rendre maxima l'inertie interclasse. Donc, puisque l'inertie interclasse et l'inertie intraclasse ont, quelle que soit la partition considérée, une même somme (l'inertie totale), augmenter l'inertie interclasse, c'est par le fait même, diminuer l'inertie intraclasse. D'où le critère d'agrégation selon l'inertie, objet du § suivant.

2.3 Mesure de la qualité d'une partition et critère d'agrégation selon l'inertie : Considérons les sept maladies à classer. Du strict point de vue de la qualité développé plus haut, la meilleure partition est celle qui a l'inertie intraclasse la plus faible et l'inertie interclasse la plus forte (ce qui est tout un). C'est donc la partition en sept classes comprenant chacune une maladie exactement ! en effet, pour cette partition, l'inertie intraclasse est nulle et l'inertie interclasse égale à l'inertie totale du nuage des maladies. Mais lorsque le spécialiste demande une classification, certes il la désire d'aussi bonne qualité que possible, mais il a d'abord d'autres exigences. Par exemple, il aura reconnu par sa propre intuition, éventuellement aidée par une analyse factorielle préalable, quatre classes qu'il voudrait cerner rigoureusement. Va-t-on pour le satisfaire considérer toutes les partitions possibles de ses individus en quatre classes pour choisir celle dont la qualité est la meilleure ? Ce serait une procédure trop longue car le nombre de combinaisons d'individus à considérer serait très grand. En revanche, procéder pas à pas à partir de la partition la plus fine (une classe par individu) en regroupant à chaque fois deux classes de façon à accroître le moins possible l'inertie intraclasse de la partition, est une méthode praticable qui conduit à une classification hiérarchique telle que nous l'avons décrite au § 1 et qui laisse au spécialiste tout un choix de partitions plus ou moins fines et de qualité acceptable.

Nous précisons dans ce § le critère utilisé pour choisir, à chaque pas, les deux classes à regrouper. Au § suivant, nous décrirons l'algorithme complet, de la base jusqu'au sommet.

Supposons qu'on soit parvenu à une partition des sept maladies (que nous désignons par des numéros : 1, 2, 3, 4, 5, 6, 7) en quatre classes ci-dessous définies :

$C_1 = \{1, 2, 3\}$   
 $C_2 = \{4, 5\}$   
 $C_3 = \{6\}$   
 $C_4 = \{7\}$  ;

les sept maladies sont représentées par des points dans l'espace à six dimensions.

Observons l'effet sur l'inertie intraclasse de cette partition, du regroupement des classes  $C_1$  et  $C_2$  en une seule :  $C_0$ . Dans la 1-ère partition (en 4 classes) l'apport de  $C_1$  et  $C_2$  à l'inertie intraclasse est la somme :

inertie interne de  $C_1$  + inertie interne de  $C_2$  ;

dans la 2-ème partition (après regroupement de  $C_1$  et  $C_2$  en  $C_0$ ), l'apport de  $C_0$  à l'inertie intraclasse est évidemment :

l'inertie interne de cette nouvelle classe  $C_0$ .

Or,  $C_1$  et  $C_2$  constituent une partition de  $C_0$  en deux classes ; à cette partition l'on peut appliquer le résultat énoncé à la fin du § 2.2 ; l'inertie totale, qui est ici l'inertie de la classe  $C_0$ , est égale à la somme de l'inertie intraclasse, qui est ici la somme des inerties internes de  $C_1$  et  $C_2$ , et de l'inertie interclasse, qui est ici très simple puisque le nuage des centres ne comprend que deux points. on a donc, en désignant par  $g_1$ ,  $g_2$ ,  $g_0$  les centres des classes  $C_1$ ,  $C_2$ , et  $C_0$  :

inertie interne de  $C_0$  = inertie interne de  $C_1$  + inertie interne de  $C_2$   
 + (masse de  $C_1$ ) $d^2(g_0, g_1)$  + (masse de  $C_2$ ) $d^2(g_0, g_2)$

Le regroupement des classes  $C_1$  et  $C_2$  en une seule  $C_0$  augmente donc l'inertie intraclasse de la partition  $\{C_1, C_2, C_3, C_4\}$  d'une quantité égale à :

(masse de  $C_1$ ) $d^2(g_0, g_1)$  + (masse de  $C_2$ ) $d^2(g_0, g_2)$  ;

expression que l'on peut transformer en tenant compte du fait que  $g_0$  est centre de gravité des points  $g_1$  et  $g_2$  affectés des masses de leurs classes respectives  $C_1$  et  $C_2$ , de façon à ne plus faire intervenir  $g_0$  dans la formule :

(masse de  $C_1$ ) $d^2(g_0, g_1)$  + (masse de  $C_2$ ) $d^2(g_0, g_2)$  =

((masse de  $C_1$ ) (masse de  $C_2$ ) / ((masse de  $C_1$ ) + (masse de  $C_2$ )) $d^2(g_1, g_2)$

C'est cette quantité (dont s'accroît l'inertie intraclasse d'une partition quand on regroupe deux de ses classes) que nous choisissons comme critère d'agrégation : étant donné une partition, on regroupera en priorité les deux classes  $q$ ,  $q'$  de cette partition rendant minima la quantité critère :

$\text{crit}(q, q') = (((\text{masse de } q) (\text{masse de } q')) / ((\text{masse de } q) + (\text{masse de } q')))$   
 $d^2(\text{centre de } q, \text{centre de } q')$  ;

quantité critère qui est égale, comme on le lit sur la formule, au produit de la distance au carré des centres des deux classes que l'on considère, par un coefficient qui se calcule sur les masses des deux classes (quotient de leur produit par leur somme) : c'est ce qu'on appelle inertie du dipôle constitué par les deux centres munis des masses de leurs classes respectives.

C'est ainsi qu'a procédé l'algorithme que nous avons fait fonctionner sur les sept maladies définies par le tableau de correspondance figurant au § 2.1, algorithme que nous décrivons au § suivant en nous aidant du listing de sortie.

2.4 L'algorithme de classification : L'algorithme part de la classification la plus fine qui soit où chaque maladie constitue une classe à elle seule : l'inertie intraclasse de cette partition est évidemment nulle et l'inertie totale du nuage des maladies coïncide avec l'inertie interclasse de la partition (pour une définition de l'inertie intracl. et intercl. cf § 2.2).

Le premier pas consiste à agréger deux classes de la partition de départ : on agrège en priorité les deux classes  $q$  et  $q'$  rendant minimum l'accroissement  $\text{crit}(q, q')$  de l'inertie intraclasse. Ce sont donc les classes {TYPHOÏDE} et {SALMONELLOSE} qu'on agrège en premier, car ces deux maladies étant représentées par le même point géométrique, l'inertie du dipôle qu'elles constituent est nulle et  $\text{crit}(\{TYPHOÏDE\}, \{SALMONELLOSE\}) = 0$  ; (pour la définition de  $\text{crit}$  cf § 2.3) ; on dit que l'agrégation se fait au niveau 0 :  $v_1 = 0$ . L'accroissement de l'inertie intraclasse est nul ; celle-ci reste donc nulle pour la nouvelle partition obtenue, et l'inertie interclasse reste égale à l'inertie totale du nuage des sept maladies, i.e. 1,1599 (valeur indiquée sur le listage de sortie, cf *infra* § 3).

Le deuxième pas prend pour base la dernière partition obtenue, laquelle est formée, d'une part de TYPHOÏDE, SALMONELLOSE, d'autre part des cinq autres classes coïncidant avec les cinq autres maladies. On agrège les deux classes  $q$  et  $q'$  de cette partition qui rendent minimum  $\text{crit}(q, q')$  : ce sont {PNEUMOPATHIE} et {AFFECTION O.R.L.}, pour lesquelles  $\text{crit} = 0,003$  (valeur indiquée sur le listage de sortie, cf § 3). La nouvelle partition ainsi obtenue : {TYPHOÏDE, SALMONELLOSE} ; {STAPHYLOCOCCIE} ; {AF. O.R.L., PNEUMOPATHIE} ; {AF. V. URINAIRES} ; {MENINGITE} a pour inertie intraclasse  $0 + 0,003 = 0,003$  et pour inertie interclasse : l'inertie du nuage des sept maladies diminuée de 0,003, c'est-à-dire  $1,1599 - 0,003 = 1,1569$ .

On procède ainsi par pas successifs. Chaque pas prend pour partition de base le résultat du pas précédent et agrège le couple des classes  $(q, q')$  de cette partition qui rend minimum  $\text{crit}(q, q')$ . La valeur de  $\text{crit}(q, q')$  pour les deux classes agrégées est appelée *niveau* de l'agrégation et mesure l'accroissement de l'inertie intraclasse au cours du pas considéré.

La classification ascendante hiérarchique des sept maladies comprend en tout six pas :

- 1° agrégation de {TYPHOÏDE} et {SALMONELLOSE} au niveau  $v_1 = 0$  ; inertie intraclasse de la partition obtenue = 0 ;
- 2° agrégation de {AF. O.R.L.} et {PNEUMOPATHIE} au niveau  $v_2 = 0,003$  ; inertie intraclasse de la partition obtenue =  $v_1 + v_2 = 0,003$  ;
- 3° agrégation de {STAPHYLOCOCCIE} et {AF. O.R.L., PNEUMOPATHIE} au niveau  $v_3 = 0,082$  ; inertie intraclasse de la partition obtenue =  $v_1 + v_2 + v_3 = 0,085$  ;
- 4° agrégation de {AF. URINAIRE} et {MENINGITE} au niveau  $v_4 = 0,103$  ; inertie intraclasse de la partition obtenue =  $v_1 + v_2 + v_3 + v_4 = 0,188$  ;
- 5° agrégation de {STAPHYL., AF. O.R.L., PNEUMOP.} et {AF. V. URI., MENING.} ; au niveau  $v_5 = 0,163$  ; inertie intraclasse de la partition obtenue =  $v_1 + v_2 + v_3 + v_4 + v_5 = 0,351$  ;
- 6° agrégation de {TYPHOÏDE, SALMONELLOSE} et {STAPHYLOC., AF.O.R.L., PNEUMOPATHIE, AF.URI., MENING.} au niveau  $v = 0,808$  ; c'est le dernier pas qui rassemble toutes les maladies en une seule classe. L'inertie intraclasse est alors égale à l'inertie totale du nuage et l'on a :

$$\begin{aligned} v_1 + v_2 + v_3 + v_4 + v_5 + v_6 &= 1,159 \\ &= \text{inertie totale du nuage des maladies.} \end{aligned}$$

On peut suivre cet algorithme sur l'arbre donné au § 1.1 : chaque agrégation correspond à un noeud : la 1° au noeud 8, la 2° au noeud 9, la 3° au noeud 10 etc. et le niveau auquel s'effectue l'agrégation est aussi le niveau du noeud correspondant. Comme on le voit les noeuds sont numérotés au fur et à mesure de leur création, à la suite des individus à classer : ici, les maladies sont numérotées de 1 à 7 et le premier noeud porte le n° 8.

On trouvera au § suivant comment se présentent les résultats de classification sur les listages de sortie de l'ordinateur.

### 3 Présentation des résultats sur les listages de sortie de l'ordinateur

Il y a diverses variantes de sortie d'ordinateur. Nous en décrivons une à titre d'exemple.

#### 3.1 Un premier tableau comprend six colonnes :

J : numéros des noeuds ; on les trouve rangés de bas en haut dans l'ordre où ils ont été créés (8, 9, 10, 11, 12, 13) ; ce numérotage commence à 8 parce que les sept premiers numéros sont dévolus aux sept maladies. A chaque noeud correspond une ligne du tableau : tous les nombres figurant dans cette ligne sont relatifs au noeud qui lui correspond.

I(J) : niveau du noeud J, i.e. niveau auquel s'effectue l'agrégation représentée par ce noeud (on dit aussi indice de niveau au lieu de : niveau). Rappelons que c'est l'accroissement de l'inertie intraclasses de la partition créée au pas précédent quand se fait l'agrégation considérée. On lit (en millièmes) :

pour le noeud 8 : niveau 0 (cf § 2.4 : 1° pas de l'algorithme)  
pour le noeud 9 : niveau 003 (2° pas)  
etc.  
pour le noeud 13 : niveau 808 (6° pas)

A(J) et B(J) : aîné et benjamin du noeud J. Chaque noeud J correspond à l'agrégation de deux classes qui elles-mêmes ont été constituées par les noeuds désignés par A(J) et B(J). On lit :

ligne du noeud 13 : A(J) = 8 ; B(J) = 12 ; ce qui signifie que le noeud 13 agrège les classes constituées aux noeuds 8 et 12.

ligne du noeud 12 : A(J) = 10 ; B(J) = 11 ; le noeud 12 agrège les classes constituées aux noeuds 10 et 11.

etc.

T(J) = I(J)/inertie totale du nuage des sept maladies ; l'inertie totale est elle-même donnée en tête du tableau : SOMME DES INDICES DE NIVEAU 0.11599E+01 ce qui se lit : 1,1599.

On a vu en effet que l'inertie totale du nuage des maladies était égale à la somme des niveaux auxquels se constituent les diverses classes de la hiérarchie : § 2.4, 6° pas. Les nombres qui figurent dans la colonne T(J) sont des millièmes. On lit :

ligne du noeud 13 : T(J) = 697 ; en effet,  $0,808/1,1599 = 0,697$

ligne du noeud 12 : T(J) = 141 ; en effet,  $0,163/1,1599 = 0,141$

etc.

C.A.H. A PARTIR D'UN TABLEAU DE COORDONNEES EUCLIDIENNES

SOMME DES INDICES DE NIVEAU 0.11599E+01

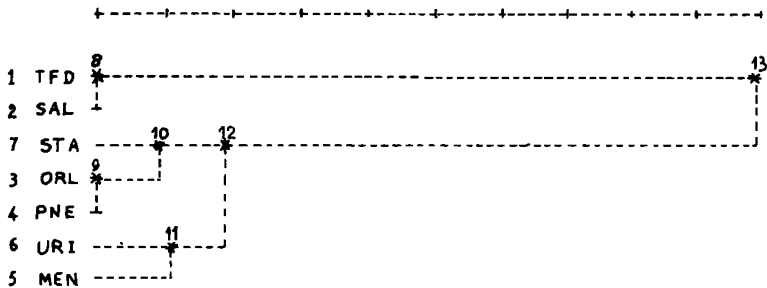
| J  | I(J) | A(J) | B(J) | T(J) | T(Q) |
|----|------|------|------|------|------|
| 13 | 808  | 8    | 12   | 697  | 697  |
| 12 | 163  | 10   | 11   | 141  | 838  |
| 11 | 103  | 6    | 5    | 89   | 927  |
| 10 | 82   | 7    | 9    | 71   | 998  |
| 9  | 2    | 3    | 4    | 2    | 1000 |
| 8  | 0    | 1    | 2    | 0    | 1000 |

l'indication  
E+01 signifie  
que le nombre  
donné est à mul-  
tiplier par 10<sup>1</sup>.

↑  
niveau du noeud  
↑  
n° du noeud

| J  | I(J) | HISTOGRAMME DES INDICES DE NIVEAU DE LA HIERARCHIE |
|----|------|--|
| 13 | 808  | *****  |
| 12 | 163  | *****  |
| 11 | 103  | *****  |
| 10 | 82   | *****  |
| 9  | 2    | *  |
| 8  | 0    | *  |

| J  | I(J) | A(J) | B(J) | IP(J) | DESCRIPTION DES CLASSES DE LA HIERARCHIE |
|----|------|------|------|-------|--|
| 13 | 808  | 8    | 12   | 7     |  |
| 12 | 163  | 10   | 11   | 5     | STA ORL PNE URI MEN                      |
| 11 | 103  | 6    | 5    | 2     | URI MEN                                  |
| 10 | 82   | 7    | 9    | 3     | STA ORL PNE                              |
| 9  | 2    | 3    | 4    | 2     | ORL PNE                                  |
| 8  | 0    | 1    | 2    | 2     | TFD SAL                                  |



on peut appeler  $T(J)$  *taux* d'accroissement de l'inertie intraclasse au noeud  $J$  ; on l'appelle aussi : contribution relative du noeud  $J$  à l'inertie totale du nuage.

$T(Q)$  cumule les  $T(J)$  de la ligne considérée et des lignes qui lui sont supérieures. On lit :

$$\text{ligne du noeud 13 : } T(Q) = T(13) = 697$$

$$\text{ligne du noeud 12 : } T(Q) = T(13) + T(12) = 697 + 141 = 838$$

etc.

Or,  $T(J)$  = niveau du noeud  $J$ /inertie totale ;

rappelons que l'inertie totale du nuage des sept maladies est égale à la somme des niveaux des noeuds 8, 9, 10, 11, 12, 13 :

$$I(8) + I(9) + I(10) + I(11) + I(12) + I(13) = \text{inertie totale}$$

et que l'inertie intraclasse de la partition obtenue en coupant l'arbre e.g. entre les noeuds 10 et 11 est égale à :

$$I(8) + I(9) + I(10) ;$$

par conséquent, la somme :

$$I(11) + I(12) + I(13)$$

est égale à l'inertie interclasse de cette même partition (cf § 2.2 : inertie totale = inertie interclasse + inertie intraclasse). En divisant cette inertie interclasse par l'inertie totale, on obtient  $T(Q)$  relatif à la ligne 11. Donc,  $T(Q)$ , à la ligne  $J$ , donne le taux d'inertie interclasse de la partition obtenue en coupant l'arbre entre les noeuds  $J$  et  $J-1$ .

Enfin, l'histogramme des indices de niveau de la hiérarchie donne pour chaque noeud une ligne d'étoiles de longueur proportionnelle au niveau de ce noeud ; et permet à l'utilisateur de juger visuellement de la décroissance des niveaux des noeuds quand on descend l'arbre. Nous y reviendrons au § 4.1.

3.2 Un deuxième tableau donne une description des classes de la hiérarchie. Il comprend six colonnes :

■  $J$ ,  $I(J)$ ,  $A(J)$ ,  $B(J)$  sont les mêmes colonnes que dans le tableau précédent.

■  $P(J)$  : donne pour chaque noeud le nombre d'individus qui lui sont reliés dans l'arbre ; autrement : le nombre d'individus de la classe que ce noeud définit dans la hiérarchie. Ces individus sont énumérés dans la colonne suivante intitulée :

#### ■ DESCRIPTION DES CLASSES

Par exemple on lit, ligne 10 (i.e. noeud 10) :

$$I(J) = 0,082 \text{ (niveau du noeud 10)}$$

$A(J) = 7$ ,  $B(J) = 9$  (le noeud 10 a été constitué par agrégation des noeuds 7 et 9)

$$P(J) = 3 \text{ (la classe définie par le noeud 10 comprend 3 maladies)}$$

DESCRIPTION DES CLASSES : STA, ORL, PNE qui sont les abréviations utilisées dans le calcul pour STAPHYLOCOCCIE, AFFECTION O.R.L., PNEUMOPATHIE.

3.3 L'arbre ou REPRESENTATION DE LA CLASSIFICATION HIERARCHIQUE : Sur le listage, l'arbre est représenté couché. A ceci près, nous l'avons



Les deux premiers tableaux du listage AIDE CAH FACOR sont relatifs, le premier aux classes supérieures de la hiérarchie (13 et 12), le second aux classes de la partition correspondant à la coupure faite sur l'arbre sous le noeud 12 (i.e. les classes 8, 10, 11). Ce sont les tableaux d'analyse des correspondances donnant facteurs, COR (ou cosinus carrés) et contributions CTR des éléments supplémentaires centres des classes 13 et 12 d'une part, 8, 10 et 11 d'autre part (cf [INT. CORR. MED.] §§ 2.4.3 et 3.1).

1° tableau :

FACTEURS POUR LES 2 CLASSES LES PLUS HAUTES DE LA HIERARCHIE

(TOUTES LES VALEURS SONT MULTIPLIÉES PAR 1000)

| CLASSE | AINE | BNJMN | POIDS | INR | QLT  | F1  | COR | CTR | F2 | COR | CTR |
|--------|------|-------|-------|-----|------|-----|-----|-----|----|-----|-----|
| 13     | 8    | 12    | 1000  | 0   | 0    | 0   | 0   | 0   | 0  | 0   | 0   |
| 12     | 10   | 11    | 870   | 91  | 1000 | 347 | 994 | 119 | 12 | 1   | 1   |
| CLASSE | AINE | BNJMN | POIDS | INR | QLT  | F3  | COR | CTR | F4 | COR | CTR |
| 13     | 8    | 12    | 1000  | 0   | 0    | 0   | 0   | 0   | 0  | 0   | 0   |
| 12     | 10   | 11    | 870   | 91  | 1000 | 23  | 5   | 10  | 5  | 0   | 1   |

**POIDS** : total des masses des maladies constituant la classe considérée.

**INR** : inertie du centre de la classe considérée, muni de la masse de sa classe, par rapport au centre du nuage, divisée par l'inertie totale du nuage.

**QLT** : qualité de la représentation des centres des classes dans l'espace des axes retenus (ici les 4 premiers axes factoriels :  $QLT = COR1 + COR2 + COR3 + COR4$ ).

**COR $\alpha$**  :  $(F\alpha)^2 / ((F1)^2 + (F2)^2 + (F3)^2 + (F4)^2 + (F5)^2)$ , ou cosinus carré de l'angle que fait le rayon joignant le centre du nuage au centre de la classe considérée avec l'axe  $\alpha$  ; on notera que nécessairement, pour toute classe autre que la cl. 13 pour laquelle l'angle en question n'est pas déterminé, cf infra,  $COR1 + COR2 + COR3 + COR4 + COR5 = 1$  : La qualité QLT n'est autre que la somme des COR $\alpha$  afférents aux axes retenus.

**CTR $\alpha$**  : inertie du centre de la classe considérée le long de l'axe  $\alpha$ , divisée par l'inertie du nuage le long de l'axe  $\alpha$ .

On remarquera que la classe 13 étant l'ensemble des sept maladies, son centre n'est autre que le centre du nuage des maladies, c'est-à-dire l'origine des axes factoriels. C'est pourquoi toutes ses coordonnées sont nulles :  $F1 = 0$  ;  $F2 = 0$  etc. ; les CTR sont donc nuls ; et, par convention on attribue également la valeur 0 à tous les COR ; son poids est 1 (rappelons que toutes les valeurs imprimées sont multipliées par 1000) : il en est ainsi de la masse de tout nuage défini par un tableau de correspondance ; son inertie INR par rapport à lui-même est nulle puisque la distance d'un point à lui-même est nulle. Quant au centre de la classe 12 on voit, en lisant la suite de ses cinq COR que seul le 1° facteur contribue à son écart de l'origine :  $COR1(12) = 0,994$ ,



c'est-à-dire presque 1 ce qui signifie que le centre de la classe 12 est presque sur l'axe 1 ; on peut voir que  $COR4(12)$  est nul alors que  $F4(12)$  ne l'est pas : cela est dû à l'imprécision des valeurs imprimées ; en effet,  $F4(12) = 0,005$  et  $COR4(12)$  est donc égal à  $(0,005)^2 / \text{carré du rayon joignant l'origine au centre de la classe 12}$  ; le carré du rayon s'obtient le plus facilement à partir de  $INR$  :

$$INR(12) = (\text{carré du rayon } (0,12) \times \text{masse de la cl.12}) / \text{inertie du nuage}$$

|                        |                        |            |
|------------------------|------------------------|------------|
| ↑                      | ↑                      | ↑          |
| 0,091                  | 0,870                  | 1,1599     |
| (cf tableau ci-dessus) | (cf tableau ci-dessus) | (cf § 3.1) |

d'où le carré du rayon  $(0,12) = (0,091 \times 1,1599) / 0,870 = 0,1213$

et  $COR4(12) = (0,005)^2 \times 0,870 / (0,091 \times 1,1599)$

$= 0,00002$  qui est imprimé 0 si l'on imprime que 3 chiffres après la virgule.

Comme le centre de la classe 12 est pratiquement sur l'axe 1, il se projette sur les autres axes au voisinage de l'origine :

$$F2(12) = 0,012 ; F3(12) = 0,023 ; F4(12) = 0,005 ; F5(12) = 0.$$

Pour utiliser ce résultat dans l'interprétation de la classe 12, il faut se reporter à l'analyse factorielle. Le tableau des facteurs sur l'ensemble des médicaments montre que le 1° axe est fait de l'opposition entre *tifomycine* et *tiophénicol* à sa partie négative ; et des autres médicaments qui ont tous un 1° facteur positif. Or le centre de la classe 12 est du côté positif de l'axe. On peut donc dire que ce qui rapproche les maladies constituant la classe 12 c'est de n'être pas traitées par *tifomycine* ni *tiophénicol*. Un retour au tableau des données confirme cette interprétation : en effet sur les 5 maladies de la classe 12 (AF. O.R.L. ; PNEUMOPATHIE ; MENINGITE ; AF. URIN. ; STAPHYLOCOCCIE) seule MENINGITE est traitée (2 fois) par *tifomycine*.

2° tableau :

FACTEURS POUR LES 3 CLASSES DE LA PARTITION (DEFINIE PAR LES DEUX NOEUDS LES PLUS HAUTS).

Ce tableau est semblable au précédent : seules les classes décrites diffèrent d'un tableau à l'autre. Pourtant on remarquera qu'on a calculé ici les totaux des colonnes relatives à l'inertie :  $INR$  et  $CTR$  : en effet, les trois classes considérées constituant une partition de l'ensemble des maladies, leurs centres constituent une représentation concentrée du nuage en trois points. Par cette concentration on perd en inertie tout ce qui est dû à la dispersion des classes 8, 10, 11 autour de leurs centres respectifs. C'est pourquoi le total de la colonne  $INR$  qui, dans le tableau des facteurs sur les sept maladies, vaut 1, vaut ici 0,838 ; la somme de chaque colonne  $CTR$  qui vaut également 1 dans le tableau des facteurs sur les maladies ne vaut pas exactement 1 ici pour l'axe 1, en diffère sensiblement pour l'axe 2

$$(\text{somme des } CTR1 = 0,937 ; \text{ somme des } CTR2 = 0,642)$$

et est très différente de 1 pour les axes suivants.

## FACTEURS POUR LES 3 CLASSES DE LA PARTITION (DEFINIE A PARTIR DES 2 NOEUDS LES PLUS HAUTS)

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

| CLASSE | AINE | BNJMN | POIDS | INR | QLT  | F1   | COR | CTR | F2   | COR | CTR |
|--------|------|-------|-------|-----|------|------|-----|-----|------|-----|-----|
| 11     | URI  | MEN   | 246   | 96  | 1000 | 106  | 25  | 3   | 627  | 872 | 475 |
| 10     | STA  | 9     | 623   | 136 | 1000 | 442  | 773 | 139 | -231 | 210 | 163 |
| 8      | TFD  | SAL   | 130   | 606 | 1000 | 2314 | 994 | 795 | 83   | 1   | 4   |
|        |      |       | 1000  | 838 |      |      |     | 937 |      |     | 642 |
| CLASSE | AINE | BNJMN | POIDS | INR | QLT  | F3   | COR | CTR | F4   | COR | CTR |
| 11     | URI  | MEN   | 246   | 96  | 1000 | 187  | 78  | 176 | -107 | 25  | 99  |
| 10     | STA  | 9     | 623   | 136 | 1000 | -41  | 7   | 22  | 49   | 10  | 53  |
| 8      | TFD  | SAL   | 130   | 606 | 1000 | -156 | 5   | 64  | -34  | 0   | 5   |
|        |      |       | 1000  | 838 |      |      |     | 262 |      |     | 157 |

A partir de l'axe 3, l'inertie afférente à chaque axe est due à la dispersion de chaque classe autour de son propre centre, plus qu'à l'inertie (interclasse) du nuage des 3 centres : 8, 10, 11. On conclura donc que la qualité de la représentation simplifiée du nuage des maladies fournie par les centres des trois classes 11, 10 et 8 est bonne en général (total de la colonne INR = 0,838), est particulièrement satisfaisante sur l'axe 1, acceptable sur l'axe 2, mais non au-delà, ce qui, il est vrai est secondaire.

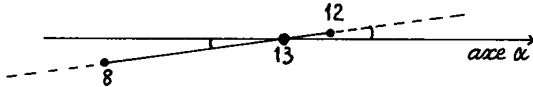
En parcourant les COR des trois classes 11, 10 et 8 on remarque une très forte valeur de COR1 (8) = 0,994 (cf *infra* Remarque). L'analyse factorielle a en effet déjà montré que la classe 8 (faite des deux maladies de même profil TYPHOIDE et SALMONELLOSE) en s'opposant à toutes les autres maladies par ses rencontres exclusives avec *tifomyicine* et *tiophénicol* créait l'axe 1 ; on le retrouve ici : fort facteur négatif, forte contribution CTR1 (0,795) et fort COR1. On remarque également une forte valeur de COR2 pour la classe 11 : COR2 (11) = 0,872 ; et, pour la classe 10, COR1 + COR2 = 0,983. Or, si le rayon joignant l'origine au centre de la classe 10 fait avec l'axe 1 un angle  $\theta_1$ , avec l'axe 2 un angle  $\theta_2$  et avec sa projection sur le plan  $1 \times 2$  un angle  $\theta$ , on a, par le théorème de Pythagore :  $\cos^2 \theta = \cos^2 \theta_1 + \cos^2 \theta_2$ , en sorte que COR1 + COR2 n'est autre que  $\cos^2 \theta$ , qui vaut donc 0,983, ce qui montre que le centre de la classe 10 est très proche du plan  $1 \times 2$  ; sur l'axe 1, la classe 10 s'oppose à la classe 8 (les 1-ers facteurs des deux centres ont des signes opposés) ; sur l'axe 2, la classe 10 s'oppose à la classe 11 (les 2-èmes facteurs des deux centres ont des signes opposés). Un retour au listage d'analyse des correspondances (ou l'examen d'un graphique représentant simultanément médicaments

et classes) permet d'expliquer ces oppositions par les emplois des médicaments avec les diverses maladies : sur l'axe 1, du côté négatif (donc avec la classe 8) on trouve, comme on le sait, *tifomycine* et *tio-phénicol* ; sur l'axe 2, du côté positif (donc avec la classe 11) on trouve *gentalline* et du côté négatif (donc avec la classe 10) on trouve *érythromycine*.

Ainsi, la confrontation des tableaux des facteurs de la partition et des médicaments permet d'expliquer la partition par les médicaments qui se rencontrent principalement avec les maladies de chacune des trois classes :

- cl. 8 = (TYPHOÏDE, SALMON.) : présence de *tifomycine* et *tio-phénicol*
- cl. 10 = (STAPHYL., PNEUMOPATHIE, AF. ORL) : présence de *érythromycine*
- cl. 11 = MENINGITE, AF. V. URINAIRES) : présence de *gentalline*.

*Remarque* : La ligne des COR de la classe 12 (1-er tableau) est identique à la ligne des COR de la classe 8 (2-ème tableau) :  $COR1(12) = COR1(8) = 0,994$  ;  $COR2(12) = COR2(8) = 0,001$  ;  $COR3(12) = COR3(8) = 0,005$  ;  $COR4(12) = COR4(8) = 0$  ;  $COR5(12) = COR5(8) = 0$ . Cela est dû au fait que les points 12 et 8 (centres des classes 12 et 8) sont alignés avec l'origine i.e. le centre de la classe 13 : en effet les deux classes 8 et 12 constituent une partition du nuage (classe 13) et le centre du nuage peut s'obtenir en prenant le centre de la classe 8, le centre de la classe 12, puis le centre de gravité des deux centres 8 et 12 affectés respectivement des masses des classes 8 et 12 : donc le centre de la classe 13 est aligné avec les centres 8 et 12 ; les deux rayons joignant l'origine aux centres 8 et 12 sont dans le prolongement l'un de l'autre ; et ils font nécessairement un même angle avec toute droite de l'espace, en particulier avec les axes. Or  $COR\alpha(8)$  et  $COR\alpha(12)$  pour un même axe  $\alpha$  sont les cosinus carrés des angles que font les rayons joignant l'origine aux centres 8 et 12 avec l'axe  $\alpha$  ; ces angles étant égaux, les  $COR\alpha$  sont égaux.



4.1.2 Les dichotomies successives : Tableau des facteurs pour les différences associées aux deux noeuds les plus hauts : Nous avons si peu d'individus à classer que nous n'avons retenu de la structure hiérarchique de l'arbre que les deux noeuds supérieurs. Dans la pratique on aura davantage de noeuds au dessus de la coupure effectuée dans l'arbre. Pour chaque noeud  $n$  on désire connaître avec précision la raison du partage de la classe associée au noeud  $n$  en ses deux sous-classes associées aux deux descendants  $A(n)$  et  $B(n)$  de  $n$ . Or, comme on peut le constater sur notre exemple, plus  $n$  est bas dans la hiérarchie et plus  $n$ ,  $A(n)$ ,  $B(n)$  sortent proches les uns des autres sur les graphiques d'analyse factorielle (e.g. le noeud 10 sort dans le même quadrant que 9 et STAPHYL.).

Pour savoir ce qui oppose les classes  $A(n)$  et  $B(n)$  en lesquelles se scinde la classe  $n$ , on examinera la direction du segment qui joint, dans l'espace à 6 dimensions, les centres des classes  $A(n)$  et  $B(n)$ . Comme nous l'avons déjà remarqué à propos des classes 8 et 12, descendantes de la classe 13, le segment qui joint les centres des classes  $A(n)$  et  $B(n)$  passe par le centre de la classe  $n$ , et le centre de la classe  $n$  est le centre de gravité des centres des classes  $A(n)$  et  $B(n)$  affectés chacun de la masse de sa classe ; les trois centres  $n$ ,  $A(n)$  et  $B(n)$  constituent ce qu'on appelle un dipôle ; par l'opposition des points  $A(n)$  et  $B(n)$  autour de  $n$ , le dipôle met en évidence l'opposition sous-jacente à la dichotomie de la classe  $n$  en les classes  $A(n)$  et  $B(n)$ . Mais l'image que nous donne de ces dipôles le graphique plan  $1 \times 2$  issu de l'analyse factorielle peut n'être pas fidèle ; et aucun

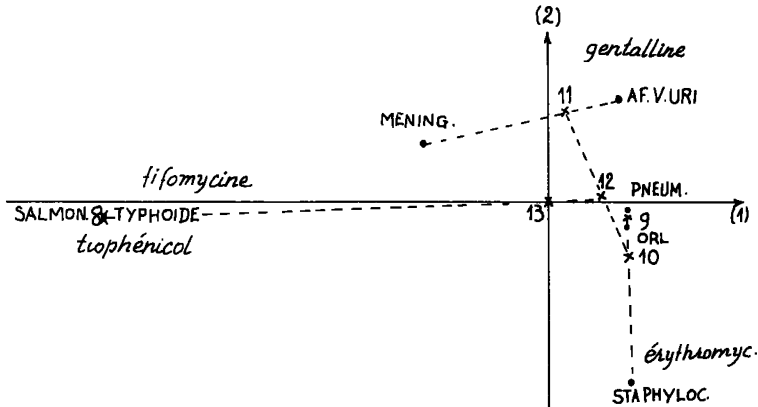


Figure : plan  $1 \times 2$  issu de l'analyse du tableau croisant les 7 maladies et les 6 médicaments. On n'a retenu que les médicaments contribuant le plus aux axes 1 et 2. Chaque segment tireté porte un noeud  $n$  et, à ses extrémités, les deux descendants  $A(n)$  et  $B(n)$  du noeud  $n$  : 8-12 descendants du noeud 13 ; 10-11 descendants du noeud 12 ; 9-STAPHYLOC. descendants du noeud 10 ; O.R.L. - PNEUMOPATHIE descendants du noeud 9.

graphique ne conviendra si le segment a une position oblique par rapport à plusieurs axes factoriels. C'est pourquoi on a édité le tableau des facteurs pour les différences associées aux noeuds retenus ; dans notre exemple les noeuds 13 et 12. On notera l'emploi du terme de noeud qui mieux que classe évoque l'idée d'une subdivision. Ce tableau donne les composantes  $D_1, D_2, \dots, D_5$  sur les 5 axes factoriels de chaque dipôle considéré : e.g. le dipôle 8-12 de centre 13 a pour composantes :

sur l'axe 1 :  $F_1(12) - F_1(8) = 0,347 - (-2,314) = 2,661$

(les valeurs 0,347 et -2,314 de  $F_1(12)$  et  $F_1(8)$  se lisent sur les tableaux 1 et 2 du § 4.1 précédent ; la valeur 2,661  $\approx$  2,662 de la différence  $F_1(12) - F_1(8)$  se lit sur le tableau des facteurs pour les différences à la ligne 13 car 13 est le centre du dipôle 8-12 ; cette composante est notée  $D_1$  (i.e. Différence sur l'axe 1) ;

sur l'axe 2 :  $F_2(12) - F_2(8) = 0,012 - (-0,083) = 0,095$   
composante notée  $D_2$  etc.

Avec chaque facteur, le tableau contient une colonne COD et une colonne CTD (les analogues de COR et CTR pour les points).

COD est le rapport du carré de la longueur du dipôle projeté sur l'axe considéré au carré de sa longueur réelle ; c'est encore le cosinus carré de l'angle que fait le dipôle avec l'axe considéré ; plus COD est voisin de 1 plus la direction du dipôle est voisine de celle de l'axe ; au contraire si le dipôle est presque perpendiculaire à l'axe considéré, COD est voisin de 0. Par exemple, le dipôle de centre 13 fait un très petit angle avec l'axe 1 :  $COD_1 = 0,994$  ; au contraire, il fait avec les 4 autres axes des angles voisins de  $90^\circ$  :  $COD_2 = 0,001$  ;  $COD_3 = 0,005$  ;  $COD_4 = 0$  ;  $COD_5 = 0$ . L'opposition entre les classes 8 et 12 est expliquée par l'axe 1 : on sait (cf. [INT. CORR. MED.] § 2.4.3) que c'est cette opposition même qui détermine l'axe 1 dans l'analyse factorielle ; ce fait réapparaît ici, colonne CTD (cf. *infra*).

FACTEURS POUR LES DIFFERENCES ASSOCIEES AUX 2 NOEUDS LES PLUS HAUTS  
(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

| NOEUD | AINE | BNJMN | POIDS | IND | QLD  | D1    | COD | CTD | D2    | COD | CTD |
|-------|------|-------|-------|-----|------|-------|-----|-----|-------|-----|-----|
| 13    | 8    | 12    | 1000  | 697 | 1000 | -2662 | 994 | 915 | -.95  | 1   | 5   |
| 12    | 10   | 11    | 870   | 141 | 1000 | 336   | 122 | 23  | -.857 | 795 | 637 |
|       |      |       |       | 838 |      |       |     | 938 |       |     | 642 |

| NOEUD | AINE | BNJMN | POIDS | IND | QLD  | D3    | COD | CTD | D4   | COD | CTD |
|-------|------|-------|-------|-----|------|-------|-----|-----|------|-----|-----|
| 13    | 8    | 12    | 1000  | 697 | 1000 | -.179 | 5   | 74  | -.39 | 0   | 6   |
| 12    | 10   | 11    | 870   | 141 | 1000 | -.229 | 57  | 188 | 156  | 2.6 | 151 |
|       |      |       |       | 838 |      |       |     | 262 |      |     | 157 |

**POIDS**: total des masses des maladies de la classe correspondant au noeud considéré :  $n$ .

**IND**: inertie du dipôle  $A(n)$ - $B(n)$ , divisée par l'inertie totale du nuage.

**QLD**: qualité de la représentation du dipôle dans l'espace des 4 premiers axes factoriels.

**D $\alpha$** : différence des valeurs du facteur  $F_\alpha$  pour l'aîné et le benjamin du noeud  $n$ .

**COD $\alpha$** : cosinus carré de l'angle que fait le segment portant le dipôle  $A(n)$ - $B(n)$  avec l'axe  $\alpha$  (ou contribution relative de l'axe  $\alpha$  au noeud)

**CTD $\alpha$** : inertie du dipôle  $A(n)$ - $B(n)$  le long de l'axe  $\alpha$ , rapportée à l'inertie totale du nuage le long de cet axe (ou contribution relative du noeud à l'axe).

Il importe de noter que de même que dans le tableau de sortie usuel de l'a. des cor. la somme des  $COR_\alpha$  afférents à un individu vaut 1 (cf. § 4.1.1, 1-er tableau), de même ici la somme des  $COR_\alpha$  vaut 1 : dans notre exemple où il n'y a que 5 facteurs,  $COD1(13) + COD2(13) + COD3(13) + COD4(13) + COD5(13) = 1$ . Dans le cas présent les  $COD5$  sont nuls pour les noeuds 12 et 13 ; c'est pourquoi la qualité QLD vaut 1 avec les quatre axes retenus. Suivant le langage usuel on dira que  $COD_\alpha(n)$  est la contribution relative du facteur  $\alpha$  à la dichotomie (ou au dipôle) afférente au noeud  $n$ .

CTD est le rapport de l'inertie interne au dipôle (i.e. inertie des deux points pesants  $A(n)$  et  $B(n)$  par rapport à leur centre  $n$ ) à la valeur propre de l'axe i.e. à l'inertie du nuage tout entier projeté sur l'axe, par rapport au centre du nuage, origine des axes factoriels). Pour le dipôle de centre 13,  $CTD1 = 0,915$ , c'est-à-dire presque 1 : la dispersion du nuage le long de l'axe 1 provient presque entièrement de la projection sur cet axe du dipôle de centre 13 ; au contraire, les autres  $CTD_\alpha$  du centre 13 sont infimes (0,005 ; 0,074 ; 0,006 ; 0).

On a dit au § 2.4, 6° pas, que l'inertie totale du nuage était égale à la somme des niveaux des noeuds. En nous bornant à une représentation simplifiée du nuage en 3 classes : 8, 10 et 11, on a une inertie totale moindre, dont le quotient par l'inertie totale du nuage des maladies a été trouvé égal à 0,838 (somme de la colonne INR du tableau 2 du § 4.1). Ce même rapport (inertie réduite/inertie du nuage des maladies) se retrouve ici comme total de la colonne IND :

$IND(13) = \text{niveau du noeud 13/inertie totale du nuage des maladies};$   
 $IND(12) = \text{niveau du noeud 12/inertie totale du nuage des maladies};$   
 $IND(13) + IND(12) = (\text{niveau du noeud 13} + \text{niveau du noeud 12})/\text{inertie totale du nuage des maladies};$   
 $IND(12) = 0,697 ; IND(13) = 0,141 ; IND(13) + IND(12) = 0,838.$

Il en est de même sur chaque axe : la somme de la colonne  $CTR\alpha$  du tableau 2 du § 4.1 se retrouve comme somme de la colonne  $CTR\alpha$  du présent tableau des différences, et exprime, ici comme  $l\alpha$ , la qualité de la représentation simplifiée retenue pour le nuage en projection sur l'axe  $\alpha$ .

#### 4.2 Tableaux relatifs au système d'axes primitif : 1 axe par variable

Les tableaux décrits et commentés au § 4.1 ci-dessus visent à donner des classes et des dichotomies de la classification une explication par les facteurs issus de l'analyse du tableau de correspondance. Or, en général, et c'est le cas dans notre exemple, les premiers axes sont bien adaptés aux grandes divisions correspondant aux noeuds supérieurs de l'arbre, tandis que les dipôles de niveau inférieur ont des directions obliques par rapport à plusieurs axes, en sorte que leur explication requiert également plusieurs facteurs d'importance variable selon les valeurs des contributions  $COR\alpha$  ; et de même pour les écarts des centres des classes à l'origine. De toute façon, une interprétation d'une classification fondée sur les facteurs reste subordonnée à l'interprétation des facteurs eux-mêmes. Le spécialiste demandera parfois une interprétation directe fondée sur les variables de départ : dans notre exemple, les médicaments. On revient donc à l'espace des profils sur les 6 médicaments du § 2.1, rapporté aux 6 axes portant chacun le nom d'un médicament : l'axe *pénicilline*, l'axe *tifomycine* etc. et l'on fait dans ce système d'axes les mêmes calculs que ceux exposés au § 4.1 précédent relativement au système des axes factoriels. Dans la pratique, surgit ici une difficulté : les variables sont trop nombreuses pour pouvoir les regarder une à une comme on le fait avec cinq ou sept facteurs.

L'analyse factorielle du nuage des variables (dans notre exemple, les médicaments) permet de choisir les variables qui ont une part notable dans l'inertie totale du nuage : dans le tableau des facteurs sur l'ensemble des médicaments on choisira les médicaments qui ont une valeur INR (quotient de l'inertie du médicament par rapport au centre du nuage, par l'inertie totale du nuage) assez forte :

Reportons-nous à cette liste des valeurs :

|                      |       | INR |
|----------------------|-------|-----|
| <i>pénicilline</i>   | : PEN | 47  |
| <i>tifomycine</i>    | : TFM | 417 |
| <i>tétracycline</i>  | : TET | 54  |
| <i>érythromycine</i> | : ERY | 115 |
| <i>tiophénicol</i>   | : TIO | 250 |
| <i>gentalline</i>    | : GEN | 116 |

On peut, au vu de cette liste, décider d'ôter *pénicilline* et *tétracycline* et de se restreindre à l'espace à 4 dimensions rapporté aux 4 autres variables : *tifomycine*, *érythromycine*, *tiophénicol* et *gentalline* (on notera que ce sont ces variables qui ont été mises

en évidence au § 4.1 pour expliquer la partition de l'ensemble des maladies en les classes 8, 10 et 11. En fait, dans l'aide à l'interprétation présentée ici, les 6 variables ont été retenues. Le listage relatif aux coordonnées sur les axes médicaments porte le titre : AID. CAH. VACOR. CLASSIFICATION SUR LES LIGNES ;

VACOR indique que les coordonnées sont relatives aux *variables* du tableau de correspondance ;

LES LIGNES sont celles du tableau de correspondance, i.e. les maladies.

Ce listage comporte, comme le listage FACOR décrit au § 4.1, trois tableaux : les deux premiers (§ 4.2.1) donnent les coordonnées (sur les axes médicaments) des centres des classes et, avec chaque coordonnée, une valeur de COR et de CTR ; le troisième tableau (§ 4.2.2) concerne les dipôles.

4.2.1 Tableaux des variables pour les centres des classes : CAH, VACOR, tableaux 1 et 2

1° Tableau :

VARIABLES POUR LES CENTRES DES 2 CLASSES LES PLUS HAUTES DE LA HIERARCHIE  
(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

| CLASSE | AINE | BNJMN | POIDS | INR | QLT | TFM | COR | CTR | ERY | COR | CTR |
|--------|------|-------|-------|-----|-----|-----|-----|-----|-----|-----|-----|
| 13     | 8    | 12    | 1000  | 0   | 0   | 116 | 0   | 0   | 116 | 0   | 0   |
| 12     | 10   | 11    | 870   | 91  | 901 | 33  | 485 | 106 | 133 | 22  | 17  |
| CLASSE | AINE | BNJMN | POIDS | INR | QLT | TIO | COR | CTR | GEN | COR | CTR |
| 13     | 8    | 12    | 1000  | 0   | 0   | 43  | 0   | 0   | 188 | 0   | 0   |
| 12     | 10   | 11    | 870   | 91  | 901 | 0   | 359 | 130 | 217 | 35  | 27  |

**POIDS** : masse de la classe, i.e. somme des masses des maladies qui la constituent.

**INR** : inertie du centre de la classe (muni de la masse de la classe) par rapport au centre du nuage, divisée par l'inertie totale du nuage.

**QLT** : qualité de la représentation dans l'espace rapporté aux axes afférents aux variables retenues (cosinus carré de l'angle que fait le rayon joignant le centre du nuage au centre de la classe considérée, avec la projection de ce même rayon sur le sous-espace des variables retenues).

**TFM** : coordonnée du centre de la classe sur l'axe tifomycine.

**CORTFM** : cosinus carré de l'angle que fait le rayon joignant les centres du nuage et de la classe considérée, avec l'axe tifomycine. La qualité QLT n'est autre que la somme des CORv afférents aux variables retenues.

**CTR TFM** : Le nuage et les centres des classes étant projetés sur l'axe tifomycine, on calcule l'inertie du centre considéré par rapport au centre du nuage (en projection sur l'axe) et on la divise par l'inertie du nuage par rapport à son centre (en projection sur l'axe); c'est encore la contribution relative de la classe à l'inertie totale du nuage en projection sur l'axe tifomycine, la classe étant supposée concentrée en son centre.

■ POIDS et INR . On a ici les mêmes colonnes POIDS et INR que dans le 1-er tableau du § 4.1 puisque ces valeurs concernent le point considéré lui-même, sans référence à un système d'axes : POIDS est la masse totale de la classe ; INR est l'inertie du centre de la classe muni de cette masse par rapport au centre du nuage (i.e. le produit de sa masse par le carré de sa distance au centre du nuage), divisée par l'inertie totale du nuage.

■ QLT. Le calcul de la qualité de la représentation est expliquée avec les CORv.

■ {TFM, ERY, TIO, GEN} . Ces colonnes, les premières de chacun des quatre blocs de trois colonnes afférents aux variables, donnent les coordonnées des centres des classes dans le système d'axes adopté ici. Dans ce système, les coordonnées sont les composantes du profil (cf. [INT. CORR. MED.] § 1.5) ; par exemple (cf. tableau de correspondance § 2.1), à la classe 12 correspond la ligne :

| PEN | TFM | TET | ERY | TIO | GEN |
|-----|-----|-----|-----|-----|-----|
| 24  | 2   | 13  | 8   | 0   | 13  |

(ligne obtenue par cumul de celles afférentes, dans le tableau de correspondance, aux 4 maladies de la classe 12) ; d'où le profil :

| PEN   | TFM  | TET   | ERY  | TIO | GEN   |
|-------|------|-------|------|-----|-------|
| 24/60 | 2/60 | 13/60 | 8/60 | 0   | 13/60 |

c'est bien la valeur  $2/60 = 0,033$  qu'on lit pour la classe 12 dans la colonne TFM.

Il importe de noter qu'en prenant pour coordonnées les composantes du profil, on assigne à la classe 13 (classe qui coïncide avec le nuage tout entier) des coordonnées non-nulles qui sont les composantes du profil moyen du nuage des sept maladies (profil de la ligne de marge du tableau de correspondance).

Si l'on désire connaître les composantes du rayon joignant le centre du nuage au centre d'une classe, il suffit de faire la différence, axe par axe, de la coordonnée du centre de la classe considérée et de la coordonnée du centre de la classe 13 : par exemple, les composantes du rayon joignant les centres du nuage et de la cl. 12 sont :

|                                |                            |
|--------------------------------|----------------------------|
| sur l'axe <i>tifomycine</i>    | : 0,033 - 0,116 = -0,083 ; |
| sur l'axe <i>érythromycine</i> | : 0,133 - 0,116 = 0,017 ;  |
| sur l'axe <i>tiophénicol</i>   | : 0 - 0,043 = -0,043 ;     |
| sur l'axe <i>gentalline</i>    | : 0,217 - 0,188 = 0,029 ;  |

on voit sur cette liste que c'est dans la direction de l'axe *tifomycine* que le centre de la cl. 12 s'écarte le plus du centre du nuage et cela, dans le sens négatif :  $TFM(12) - TFM(13) = -0,083$ . Un retour au tableau de correspondance montre en effet que le médicament *tifomycine* est principalement employé avec les maladies de la classe 8 qui sont précisément les maladies n'entrant pas dans la cl. 12.

■ CORv. Ces quantités sont définies comme les CORa en analyse des correspondances : CORv représente la part afférente à la variable v dans le carré de la distance du centre de la classe considérée c au centre du nuage (cl. 13) ; elle est encore appelée contribution relative de la variable v à l'écart de la classe c au centre du nuage. De façon précise on a pour la classe 12 :



$$d^2(13,12) = (1/f_{PEN}) (PEN(12) - PEN(13))^2 + \\ (1/f_{TFM}) (TFM(12) - TFM(13))^2 + \\ \dots \dots \dots \\ (1/f_{GEN}) (GEN(12) - GEN(13))^2 ;$$

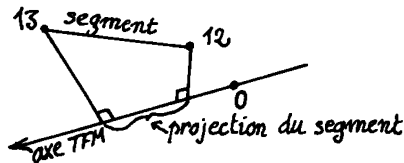
c'est la formule usuelle de la distance distributionnelle, où  $(1/f_{PEN})$  est l'inverse de la masse de *pénicilline* dans l'analyse de correspondance ; etc. cette distance a déjà été calculée d'après le tableau 1 de AIDE CAH. FACOR. (§ 4.1) ; on a :

$$d^2(13,12) = 0,1213.$$

Dans ce carré de distance la part relative afférente à la variable TFM est :

$$CORTFM(12) = (1/f_{TFM}) (TFM(12) - TFM(13))^2 / d^2(13,12) \\ = (69/8) \cdot (-0,083)^2 / (0,1213) \\ = 0,485.$$

En termes géométriques, le numérateur de la fraction CORTFM(12) peut s'interpréter comme le carré de la longueur de la projection sur l'axe TFM du segment (13 - 12) ; CORTFM(12) est le cosinus carré de l'angle que fait ce segment avec l'axe TFM : à la valeur 0,485 du cosinus carré correspond un angle d'environ 46°.



De même que dans le système des axes factoriels la somme des COR $\alpha$  afférents à tous les axes et relatifs à un point donné vaut 1, de même dans le système des axes primitifs la somme des COR $v$  afférents à toutes les variables et relatifs à un point donné vaut 1 (car, en bref, les COR $v$  sont les parts relatives afférentes aux  $v$  dans le carré de la longueur). Si on se borne à considérer une partie des variables (i.e. à projeter la figure sur le sous-espace engendré par quelques axes), le carré de la longueur n'est pas représenté exactement : la qualité de la représentation est la somme des COR $v$  afférents aux variables retenues : ici, on a pour 12 (i.e. pour la longueur du segment 13 - 12) une qualité QLT = 0,901. En ne retenant que les deux variables TFM et TIO pour lesquelles COR $v$  (12) prend ses plus fortes valeurs, on a une qualité de représentation de 0,844 :

$$CORTFM(12) + CORTIO(12) = 0,485 + 0,359 = 0,844.$$

D'autre part, sur les colonnes TFM et TIO on lit que les coordonnées du centre de la classe 12 sont inférieures à celles du centre du nuage (cl. 13) ; en fait, TIO(12) = 0, le *tiophénicol* n'étant jamais utilisé avec les maladies de la cl. 12. Donc, on retrouve le fait (déjà vu antérieurement) que la classe 12 se caractérise par un usage faible ou nul de *tifomycine* et *tiophénicol*.

Pour la classe 13 (dont le centre est celui du nuage) les COR $v$ (13) (comme les COR $\alpha$ (13) relatifs aux axes factoriels) sont de la forme indéterminée 0/0 ; on leur attribue par convention la valeur 0.

■ CTR<sub>v</sub>. Ces quantités sont définies comme les CTR<sub>α</sub> en a. des correspondances. En a. des correspondances, le CTR<sub>α</sub> d'un élément est la part relative de cet élément dans l'inertie  $\lambda_{\alpha}$  du nuage le long de l'axe. Ici, on considère la projection du nuage sur l'axe v. Le centre du nuage projeté est la projection du centre du nuage. On note INAV l'inertie du nuage le long de l'axe v, i.e. l'inertie du nuage projeté sur l'axe v par rapport à son centre. L'inertie totale du nuage se décompose ainsi sur les 6 axes de référence (comme elle se décompose aussi sur les axes factoriels) : de même que l'on a, dans le système des axes factoriels :

$$\text{inertie tot. du nuage} = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 ,$$

on a, dans le système des 6 axes médicaments :

$$\begin{aligned} \text{inertie tot. du nuage} &= \text{INA(PEN)} + \text{INA(TFM)} + \text{INA(TET)} + \text{INA(ERY)} \\ &+ \text{INA(TIO)} + \text{INA(GEN)} ; \end{aligned}$$

les INAV sont donc les équivalents des  $\lambda_{\alpha}$  des axes factoriels.

Or, la part relative e.g. de TFM dans l'inertie totale du nuage (part relative qui apparaît ici comme  $\text{INA(TFM)}/\text{inertie tot.}$ ) figure dans le listage d'analyse factorielle, tableau des facteurs sur les médicaments, sous le titre INR. On a donc le moyen de calculer simplement INAV : c'est le produit  $\text{INR}(v) \cdot \text{inertie tot. du nuage}$ .

Par exemple, pour le médicament TFM = *tifomycine*,

$$\begin{aligned} \text{INA(TFM)} &= \text{INR(TFM)} \cdot \text{inertie totale du nuage} \\ &= 0,417 \cdot 1,16 = 0,484 ; \end{aligned}$$

(la valeur 1,16 de l'inertie totale du nuage est prise sur le listage d'analyse factorielle, comme la valeur 0,417 de  $\text{INR(TFM)}$ .)

Calculons par exemple CTR<sub>TFM</sub> pour le centre de la classe 12. L'inertie du centre de la cl. 12 en projection sur l'axe TFM, par rapport à la projection sur cet axe du centre du nuage (13), est :

$$\begin{array}{ccc} \frac{(1/f_{\text{TFM}})}{\uparrow} & \underbrace{(\text{TFM}(12) - \text{TFM}(13))^2}_{\substack{\text{carré de la différence} \\ \text{des coordonnées des} \\ \text{centres 12 et 13 sur} \\ \text{l'axe TFM}}} & \cdot \frac{f_{(12)}}{\uparrow} \\ \text{inverse de la masse} & & \text{masse de la} \\ \text{de TFM coef. de la} & & \text{classe 12} \\ \text{dist. du } \chi^2 & & \\ = (69/8) & \cdot & (-0,083)^2 \cdot (0,870) = 0,052 \end{array}$$

(la masse  $f_{(12)}$  de la classe 12 est prise dans la colonne POIDS du listage de classification ; quant au carré de la distance entre les centres des classes 12 et 13 en projection sur l'axe TFM, il a déjà été calculé en expliquant CORTFM(12). D'où :

$$\text{CTR}_{\text{TFM}}(12) = 0,052/0,0484 = 0,106 \text{ (cf. listage).}$$

## 2° tableau

VARIABLES POUR LES CENTRES DES 3 CLASSES DE LA PARTITION (DEFINIE A PARTIR DES 2 NOEUDS LES PLUS HAUTS).

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

| CLASSE | AINE | BNJMN | POIDS | INR | QLT | TFM | COR | CTR  | ERY | COR | CTR |
|--------|------|-------|-------|-----|-----|-----|-----|------|-----|-----|-----|
| 11     | URI  | MEN   | 246   | 96  | 940 | 118 | 0   | 0    | 0   | 257 | 214 |
| 10     | STA  | 9     | 623   | 136 | 848 | 0   | 458 | 149  | 186 | 168 | 198 |
| 8      | TFD  | SAL   | 130   | 606 | 901 | 667 | 485 | 705  | 0   | 22  | 113 |
|        |      |       | 1000  | 838 |     |     |     | 854  |     |     | 525 |
| CLASSE | AINE | BNJMN | POIDS | INR | QLT | TIO | COR | CTR  | GEN | COR | CTR |
| 11     | URI  | MEN   | 246   | 96  | 940 | 0   | 96  | 37   | 418 | 587 | 484 |
| 10     | STA  | 9     | 623   | 136 | 848 | 0   | 172 | 93   | 150 | 50  | 59  |
| 8      | TFD  | SAL   | 130   | 606 | 901 | 333 | 359 | 870  | 0   | 35  | 182 |
|        |      |       | 1000  | 838 |     |     |     | 1000 |     |     | 725 |

Ce tableau est semblable au précédent : seules les classes diffèrent d'un tableau à l'autre. Comme au § 4.1.1, on a calculé les totaux des colonnes relatives à l'inertie, INR et CTRv. Ne disons rien de INR déjà vu au § 4.1.1. Quant à CTRv, on a par exemple pour total de la colonne CTRTFM : 0,854. Cela signifie qu'en projection sur l'axe TFM l'inertie du nuage des maladies concentré aux 3 centres des classes 11, 10, 8 atteint 85,4% de l'inertie totale ; ou encore que la part d'inertie sur l'axe TFM due à la dispersion des classes 11, 10, 8 autour de leurs centres respectifs n'est que de 14,6% ; ce qui est peu. Au contraire la dispersion interne aux classes reste grande sur l'axe ERY, car le total de la colonne CTRERY n'est que 0,525 : en retournant au tableau des données on voit en effet que la classe 10 est dispersée sur l'axe ERY : l'*érythromycine* jouant un rôle relativement plus important dans le traitement des STAPHYLOCOCCIES que dans le traitement des deux autres maladies de la classe 10 (AFFECTION ORL et PNEUMOPATHIE).

D'autre part, on constate que dans l'espace des quatre variables retenues la qualité de la représentation des trois classes 11, 10 et 8 est bonne, ce qui prouve que les variables ont été bien choisies (cf. 4.2 début).

On retrouve en particulier ici le fait maintes fois signalé que la classe 8 se caractérise par l'emploi de *tifomycine* et *tiophénicol* :  $CORTFM(8) + CORTIO(8) = 0,844$  ce qui prouve que le centre de la classe 8 est proche du plan de ces deux médicaments ;  $CTR(8) = 0,705$ , ce qui signifie que le centre de la cl. 8 représente 70% de l'inertie du nuage le long de l'axe *tifomycine* ;  $CTR(8) = 0,870$ , i.e. le centre de la cl. 8 représente 87% de l'inertie du nuage le long de l'axe *tiophénicol*. Comme au § 4.1.1 on remarque que les CTRv ont les mêmes valeurs pour les classes 8 et 12 : en effet, les centres de ces classes sont opposés sur un segment passant par le centre du nuage, et les CTRv pour 8 et 12 ne sont autres que les cosinus carrés des angles formés par ce segment avec les axes v.

Quant à la séparation de la classe 12 en les classes 10 et 11, elle apparaîtra plus nettement au § 4.2.2 sur le tableau relatif aux dipôles.

4.2.2 Les dichotomies successives : Tableau des variables pour les différences associées aux noeuds les plus hauts : Ce troisième tableau édité par le programme AID. CAH. VACOR. est l'analogue (relativement aux axes des variables) de celui (relatif aux axes factoriels) expliqué au § 4.1.2. Ici comme là, le but est de savoir ce qui oppose les classes A(n) et B(n) en lesquelles se scinde une classe n.

D'abord notons que dans le système des variables d'origine, comme dans le système des axes factoriels, le total d'une colonne CTD afférente aux deux noeuds les plus hauts est identique (aux imprécisions de calcul près) au total de la colonne CTR afférente aux trois classes

#### VARIABLES POUR LES DIFFERENCES ASSOCIEES AUX 2 NOEUDS LES PLUS HAUTS

(TOUTES LES VALEURS SONT MULTIPLIEES PAR 1000)

| NOEUD | AINE | BNJMN | POIDS | IND | QLD | TFM  | COD | CTD  | ERY  | COD | CTD |
|-------|------|-------|-------|-----|-----|------|-----|------|------|-----|-----|
| 13    | 8    | 12    | 1000  | 697 | 901 | 633  | 485 | 811  | -133 | 22  | 130 |
| 12    | 10   | 11    | 870   | 141 | 877 | -118 | 129 | 44   | 186  | 323 | 395 |
|       |      |       |       | 838 |     |      |     | 855  |      |     | 525 |
| NOEUD | AINE | BNJMN | POIDS | IND | QLD | TIO  | COD | CTD  | GEN  | COD | CTD |
| 13    | 8    | 12    | 1000  | 697 | 901 | 333  | 359 | 1000 | -217 | 35  | 209 |
| 12    | 10   | 11    | 870   | 141 | 877 | 0    | 0   | 0    | -272 | 425 | 515 |
|       |      |       |       | 838 |     |      |     | 1000 |      |     | 724 |

**POIDS** : total des masses des maladies constituant la classe correspondant au noeud n considéré.

**IND** : inertie du dipôle A(n)-B(n), rapportée à l'inertie totale du nuage.

**QLD** : qualité de la représentation du dipôle dans l'espace des axes afférents aux variables retenues.

**TFM** : différence des coordonnées (i.e. composantes de profil) sur l'axe TFM pour les classes A(n) et B(n) :  $TFM(A(n)) - TFM(B(n))$ ; e.g. pour le noeud 12 :

$$TFM(A(12)) - TFM(B(12)) = TFM(10) - TFM(11) = 0 - 0,118 = -0,118.$$

**CODTFM** : cosinus carré de l'angle que fait le segment joignant les centres des classes A(n) et B(n) (i.e. portant le dipôle A(n)-B(n)) avec l'axe TFM (ou contribution relative de la variable TFM au noeud n). La qualité QLD n'est autre que la somme des COD<sub>v</sub> afférents aux variables retenues.

**CTDTFM** : inertie du dipôle A(n)-B(n) le long de l'axe TFM, rapportée à l'inertie totale du nuage des maladies le long de cet axe (ou contribution relative du noeud n à la dispersion du nuage le long de l'axe TFM)

de la partition définie à partir de ces deux noeuds ; e.g. on a :

$$\text{CTRTFM}(11) + \text{CTRTFM}(10) + \text{CTRTFM}(8) = 0 + 0,149 + 0,705 = 0,854$$

$$\text{CTDTFM}(13) + \text{CTDFTM}(12) = 0,811 + 0,044 = 0,855.$$

Les informations portées sur la ligne du noeud 13 ne sont pas nouvelles pour nous (pas plus que ne l'étaient celles du tableau du § 4.1.2); par exemple :

$$\begin{aligned} \text{CTDTFM}(13) &= \text{CTRTFM}(12) + \text{CTRTFM}(8) \\ 0,811 &= 0,106 + 0,705 ; \end{aligned}$$

l'inertie du dipôle (A(13) - B(13)) = (12 - 8) est la somme des inerties des deux points : centre de la cl. 12 et centre de la cl. 8, par rapport au centre du dipôle, qui n'est autre que le centre du nuage.

De même, les cosinus carrés peuvent se calculer aussi bien entre l'axe TFM et le segment (12 - 8) qu'entre l'axe TFM et l'un des deux segments (13 - 12) ou (13 - 8) qui sont des fragments de (12 - 8) ; donc :

$$\text{CODTFM}(13) = \text{CORTFM}(12) = \text{CORTFM}(8) = 0,485.$$

La seule information intéressante (dans le cas très simplifié que nous avons choisi pour le présent exposé) est relative à la subdivision de 12 en 10 et 11. On voit immédiatement sur les colonnes COD que le dipôle (10 - 11) est principalement expliqué par les variables GEN et ERY qui, à elles-deux expliquent à peu près les 3/4 (74,8%) de l'écart des deux centres (10 - 11) entre eux :

$$\text{CODGEN}(12) + \text{CODERY}(12) = 0,425 + 0,323 = 0,748 ;$$

les colonnes GEN et ERY donnent immédiatement le sens de cette différence (i.e. en plus ou en moins) :

$$\text{GEN}(A(12)) - \text{GEN}(B(12)) = \text{GEN}(10) - \text{GEN}(11) = - 0,272 ;$$

$$\text{ERY}(A(12)) - \text{ERY}(B(12)) = \text{ERY}(10) - \text{ERY}(11) = 0,186 .$$

Donc, prédominance de *gentalline* avec la classe 11 (AFFECTION des VOIES URINAIRES et MENINGITE) et prédominance de *érythromycine* avec la classe 10 (STAPHYLOCOCCIE, AF. ORL, PNEUMOPATHIE). On peut vérifier ces faits sur le tableau de correspondance (plus précisément, les emplois de *gentalline* dans la classe 11 sont liés à AF. des VOIES URINAIRES, non à MENINGITE).

##### 5 Éléments supplémentaires ; agrégation autour de centres fixes ou variables

Dans la structure déterminée par l'analyse des données d'après un ensemble d'individus de base, il est fréquemment nécessaire d'insérer *a posteriori* des individus nouveaux. A cette nécessité répond, en analyse factorielle, l'adjonction d'éléments supplémentaires : on sait que, par des calculs très simples (éventuellement réalisés sur une calculatrice de poche), ces éléments sont projetés sur les axes factoriels, à la construction desquels ils n'ont pas contribué. En classification automatique on doit de même pouvoir insérer à peu de frais les individus nouveaux dans des classes préétablies : nécessité particulièrement pressante en médecine et biologie, où une classe pourra définir une maladie, un type de cellule... ; l'insertion d'un individu nouveau apportant un diagnostic ou un indice de diagnostic.

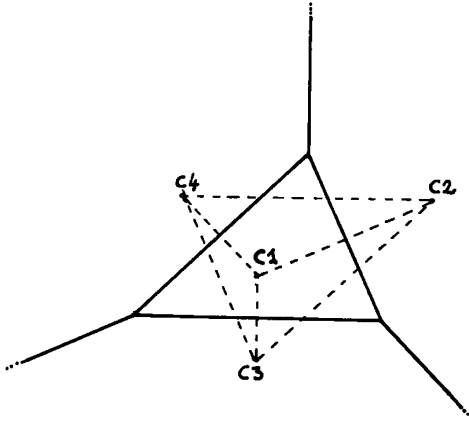
Il est possible d'insérer un individu nouveau dans une classification hiérarchique : sans exposer cette construction (pour laquelle nous renvoyons à Jambu et Lebeaux ; Dunod 1979) nous nous bornerons à indiquer brièvement comment un individu nouveau peut être affecté à l'une ou l'autre des classes  $c$  d'une partition (sur ce terme cf. ci-dessus § 1) : ce sera l'occasion de présenter le principe de base de l'algorithme d'agrégation autour de centres variables (algorithme des *nuées dynamiques* de E. Diday).

L'interprétation d'une classification ascendante hiérarchique aboutit en effet, communément, à choisir un niveau auquel on coupe l'arbre, pour ne retenir que les noeuds supérieurs à ce niveau ; les classes dépendantes de ces noeuds constituent une partition, comme on l'a expliqué au § 4.0. Placer un individu nouveau dans une de ces classes, c'est, par le fait même, situer celui-ci dans la hiérarchie des noeuds supérieurs, seule retenue en définitive.

Dans le cadre de la géométrie euclidienne, qui est celui de la présente introduction toute entière, une solution très simple s'offre : affecter l'individu nouveau  $is$  à celle des classes (de la partition) du centre  $c$  de laquelle il est le plus proche. Le coût en calcul est peu important : il suffit de calculer les distances de  $is$  à tous les centres  $c$  retenus (lesquelles ne sont jamais plus, dans la pratique, que quelques dizaines). Même s'il est fastidieux d'effectuer sans ordinateur un tel calcul, le plus petit ordinateur y suffit, et en un temps minime ; ce qui est essentiel dans des applications biologiques, où e.g. (cf. Flandrin et coll.) on devra pour le cas d'un seul malade, affecter à des classes une centaine de cellules.

Nous ferons ici quelques remarques complémentaires.

A) Attacher tout individu à la classe du centre de laquelle il est le plus proche revient à assigner à chaque classe  $c$  un domaine polyédral dont les faces sont des portions de plan (plus exactement d'hyperplan) médiateur des segments joignant les centres de la classe aux centres des classes limitrophes. La figure ci-jointe, dessinée dans le cas où l'espace ambiant n'est qu'un plan (deux dimensions) fait voir ce que nous voulons dire : dans ce cas simple les domaines des classes sont simplement polygonaux !



Sur la figure les domaines des classes sont délimités par un trait continu fort, et les segments joignant les centres des classes sont en trait tireté fin.

B) Les domaines ainsi délimités ne coïncident pas exactement avec ceux des classes construites sur les individus de base par l'algorithme de C.A.H. : e.g. un individu de la classe  $c_2$ , peut éventuellement tomber dans le domaine de la classe  $c_1$ . Toutefois, selon notre expérience, cette modification de la partition est peu importante et sans inconvénient pratique.

C) On peut dire plus. Nous sommes partis d'une partition : e.g. , sur la figure en quatre classes dont les centres sont numérotés de  $c_1$  à  $c_4$ . L'affectation des individus au centre le plus proche aboutit à redécouper ces classes. Par le fait même on a pour les classes nouvelles de nouveaux centres de gravité :  $c'_1, c'_2, \dots, c'_4$ . On effectue un nouveau découpage en affectant les individus à celui des centres  $c'_n$  dont il est le plus proche ; d'où de nouveaux centres de gravité  $c''_1, \dots, c''_n, \dots$  ; auxquels on peut affecter une fois de plus les individus, etc.

On peut dire qu'il y a alors *agrégation des individus autour de centres variables*. Comme l'a montré E. Diday, l'intérêt de cette procédure itérative est d'une part qu'en fait elle s'arrête rapidement (e.g. après 5 itérations...) ; et d'autre part qu'à chaque pas la nouvelle partition obtenue est meilleure que celle dont elle est issue en ce sens que l'inertie intraclasse a diminué (et donc l'inertie inter-classe augmenté ; cf. § 2.2).

D) Dans la méthode des *nuées dynamiques* de E. Diday, les centres  $n$ , sont initialement choisis de façon quelconque et il est de règle de faire plusieurs essais ; d'où plusieurs partitions ; dont on combine les résultats pour définir des classes stables (dites *formes fortes* par E.D.) qui sont en bref des classes d'individus qui à chaque essai ont été rangés ensemble. Ici, et c'est selon nous préférable, le système des centres  $c_1, \dots, c_n, \dots$ , est choisi d'après les résultats d'une classification ascendante hiérarchique, certes coûteuse à établir s'il y a un grand nombre d'individus, mais se prêtant à une interprétation sûre. Eventuellement si l'on part d'un tableau de correspondance  $I \times J$  comportant un très grand nombre d'individus (ensemble  $I$ ), on pourra faire une C.A.H. sur l'ensemble  $J$  des variables, puis placer les centres des classes d'une partition de l'ensemble  $J$  des variables dans l'espace rapporté aux axes factoriels (e.g. aux 10 premiers axes) et agréger autour de ces centres les individus  $i$  (eux-mêmes placés dans ce même espace, selon le principe de la représentation simultanée propre à l'a. des c.). Cette méthode (utilisée d'abord par J.P. Bordet) a l'avantage d'économiser les calculs d'une C.A.H. sur une ensemble  $I$  de cardinal élevé.