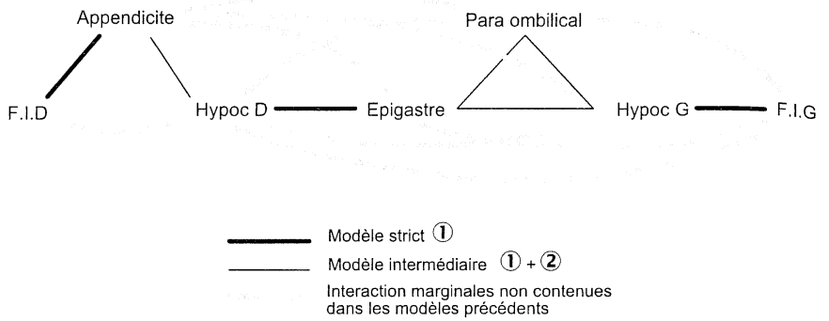


Fig. 2.2 — *Représentation graphique des interactions partielles et marginales d'ordre 2.*

Le réseau d'interactions est beaucoup plus dense et d'autant moins lisible. Il est, de plus, parasité par de nombreux facteurs de confusion. Ainsi, une douleur de la fosse iliaque droite se révèle liée marginalement à une douleur de l'hypochondre droit, mais non liée partiellement. Cela signifie possiblement qu'une crise d'appendicite aiguë génère souvent une douleur dans ces deux régions, mais que si l'on fixe la variable « crise d'appendicite », l'association disparaît (cette dernière variable serait donc un facteur de confusion).

La prise en compte des interactions partielles conduit donc à améliorer la lisibilité du schéma des relations associant les sept variables.

3.
Méthodes de classification, analyse en *clusters*

Les méthodes de classification, encore appelées analyses en *clusters* permettent de découper un jeu de données en sous-groupes homogènes. Un point important est que ce découpage est effectué *de novo* : les clusters ne sont pas définis à partir d'une variable externe (comme dans le cas des méthodes de segmentation ou de l'analyse discriminante) mais à partir de la structure même des données.

L'analyse en clusters permet, en théorie, de dégager des groupes homogènes de variables aussi bien que de sujets. S'il existe, en pratique, une différence, elle trouve son origine dans de simples considérations numériques : le nombre de variables étant le plus souvent bien inférieur au nombre de sujets, les méthodes coûteuses en temps de calcul seront généralement limitées au classement des variables.

Il existe une grande quantité de méthodes d'analyse en clusters. Loin d'être équivalentes, elles ont chacune leur terrain de prédilection : les classifications hiérarchiques sont particulièrement appréciées par les biologistes, alors qu'à l'opposé, les méthodes du type « nuées dynamiques » pourront plutôt intéresser l'économiste désireux de fragmenter un échantillon de plusieurs milliers d'agents.

> Une analyse en clusters ne requiert aucune condition de validité. Il est cependant crucial de noter que le mode de présentation des données (données brutes, centrées, normalisées, etc.) peut grandement influencer la forme des résultats ([1]).

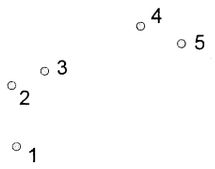
Classifications hiérarchiques ascendantes

En quelques mots : le terme de classification hiérarchique ascendante recouvre en fait un ensemble de techniques construites sur le même principe : le regroupement successif des points par ordre de proximité décroissante.

Prenons un exemple volontairement simpliste, le découpage de l'ensemble suivant ([2]) :

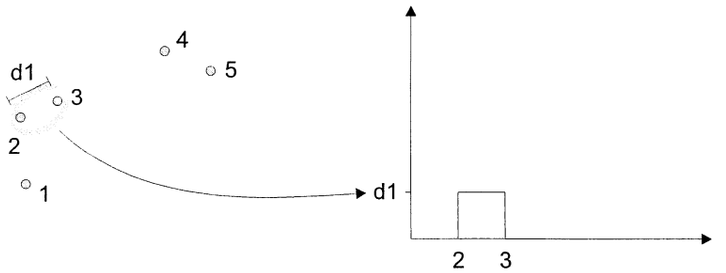
[1] Voir p. 313, chapitre « Analyse en composantes principales », pour une discussion du concept de normalisation de données.

[2] Exemple inspiré de Lebart, L., Morineau, A., Fenelon, JP. *Traitement des données statistiques, méthodes et programmes*, Dunod, Paris, 1982,



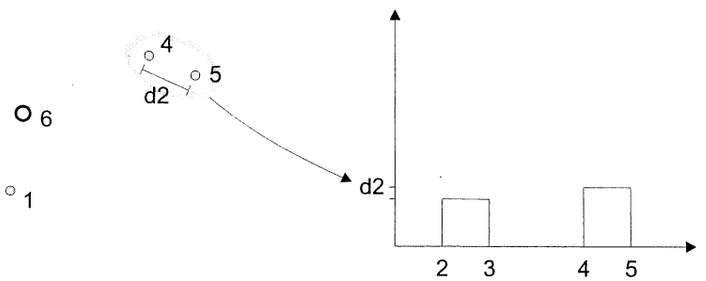
Dans un premier temps nous allons regrouper les deux points les plus proches : les points 2 et 3.

Pour garder une trace de ce regroupement, nous allons inscrire sur un graphique deux barres verticales, rejointes par une horizontale à la hauteur d1 correspondant à la distance qui sépare les points 2 et 3. Nous obtenons ainsi :

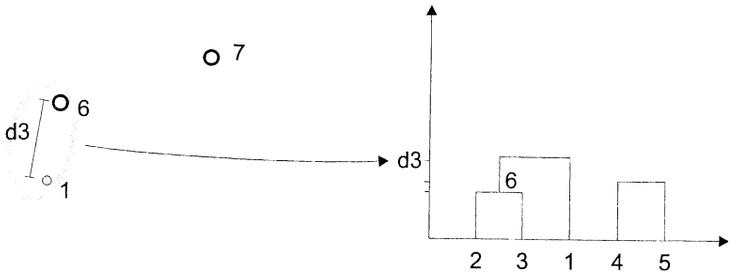


Une fois ce regroupement opéré, les points 2 et 3 sont remplacés par un nouveau point, placé au « centre » de 2 et de 3. Ce point (en réalité un *cluster* de points) porte désormais le numéro 6.

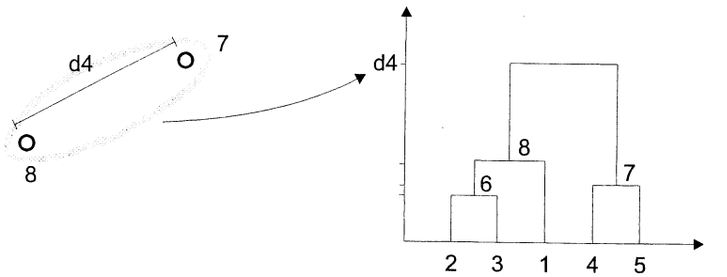
Quels sont maintenant les deux points les plus proches ? Il s'agit de 4 et 5. Nous allons, de nouveau, regrouper ces deux points et compléter notre graphique. Nous obtenons maintenant :



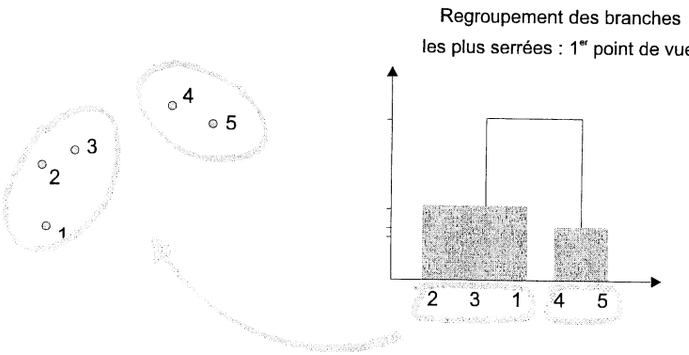
Les points 4 et 5 étant remplacés par l'élément numéro 7, le regroupement entre les points les plus proches se fait maintenant entre les points 1 et 6 ; nous obtenons :



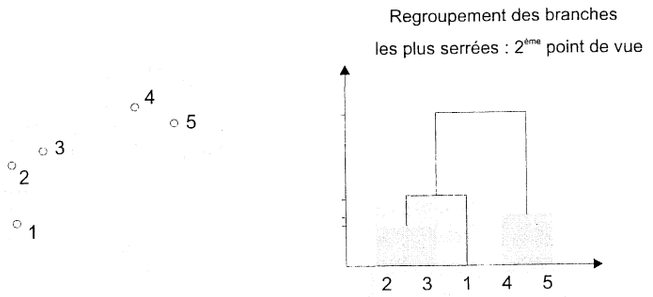
puis finalement :



Nos points sont représentés par un schéma arborescent, comment l'interpréter ? Il s'agit essentiellement de repérer visuellement les branches serrées les unes contre les autres. Il est ainsi possible de décider qu'un regroupement pertinent se fait de la façon suivante :



Dans ce cas le découpage final sera (1, 2, 3) et (4, 5). Il est aussi envisageable de trouver que le groupe (2, 3, 1) est trop hétérogène et de préférer le découpage :



On est en droit d'être choqué par une approche aussi subjective ; plusieurs questions méritent ainsi d'être posées :

1. S'il faut reconnaître visuellement les branches les plus ramassées, pourquoi ne pas regarder directement les données et les classer de la même façon ?

2. N'existe-t-il pas des méthodes objectives permettant de déterminer le nombre optimal de clusters ?

3. Un point technique peut aussi être soulevé : une fois le cluster 6 substitué aux points 2 et 3, comment fait-on pour calculer la distance d3 entre le point 1 et le cluster 6 ?

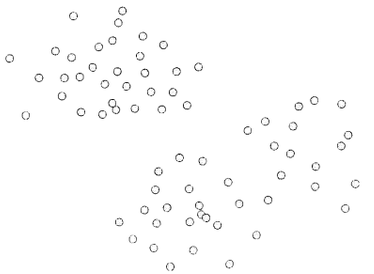
Répondons point par point :

1. N'est-il pas aussi simple de regarder directement les données ? Non, c'est l'exemple choisi qui est (volontairement) trop simple. C'est, en effet, parce que les points à classer sont représentés sur un plan que nous pouvons les appréhender visuellement et, éventuellement, les regrouper intuitivement.

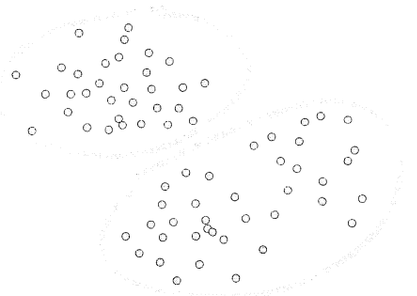
En pratique, il est exceptionnel de pouvoir représenter des données de la sorte : si p variables sont mesurées chez n sujets, classer les sujets revient à regrouper n points dans un espace de dimension p, alors que classer les variables revient à regrouper p points dans un espace de dimension n. Or la représentation mentale d'objets géométriques est impossible au-delà de 3 dimensions ; ainsi, la représentation schématique par arborescence permet-elle d'accéder à une information sinon inaccessible.

2. Existe-t-il une méthode pour déterminer le nombre optimal de clusters ? Oui, il existe *des* méthodes permettant de déterminer un nombre optimal de clusters, nous en aborderons quelques-unes dans l'exemple ci-après. Ces méthodes ne donnent cependant que des indications et n'aboutissent par ailleurs pas toujours aux

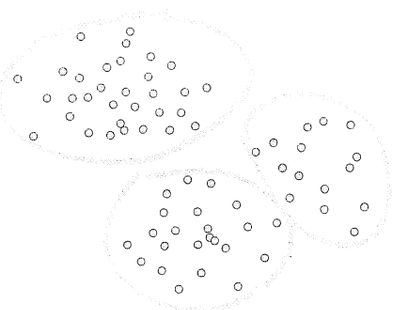
mêmes résultats. Cela est en fait compréhensible si l'on se rappelle que l'analyse en clusters est proche des techniques de reconnaissance de forme, qui comportent toujours une grande part de subjectivité. Ainsi, à la question : dans le nuage de points ci-dessous, combien y a-t-il de clusters ?



Certains diront deux :



D'autres diront trois :

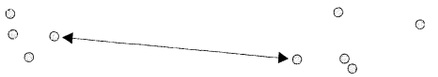


La réponse n'est pas d'ordre statistique, elle dépend implicitement de l'idée que l'on se fait *a priori* d'un cluster : doit-il être nécessairement circulaire, peut-il être elliptique ?

3. A la question : « Une fois le cluster 6 substitué aux points 2 et 3, comment fait-on pour calculer la distance d3 entre le point 1 et le cluster 6 ? », il n'est pas simple, en fait, de répondre.

Considérons plus généralement la question suivante : à partir de n_1 points (x_i) nous définissons le cluster c_1 ; à partir de n_2 points (y_j) nous définissons le cluster c_2. Comment définir la distance entre les clusters c_1 et c_2 ? Plusieurs réponses sont envisageables :

– La plus petite distance entre un x_i et un y_j.



$d(c_1, c_2) = \min[d(x_i,y_j)]$

– La plus grande distance entre un x_i et un y_j.

$d(c_1, c_2) = \max[d(x_i,y_j)]$

– La distance entre le centre de gravité de c_1 et de c_2.

$d(c_1, c_2) = d[\text{centre}(c_1), \text{centre}(c_2)]$

– La distance moyenne entre les x_i et les y_j.

$d(c_1, c_2) = \text{moyenne}[d(x_i,y_j)]$

– Et bien d'autres...

Cette remarque est plus qu'un point de détail, car la définition de la distance entre c_1 et c_2 va souvent modifier sensiblement la forme de l'arbre de classification.

Laquelle faut-il choisir en pratique ? Cette question est sans réponse univoque, chaque approche ayant ses avantages et ses inconvénients. En fait, quand on décide d'utiliser une méthode de classification hiérarchique ascendante, il est raisonnable d'utiliser plusieurs distances : une stabilité des résultats sera en faveur de l'existence réelle d'un découpage naturel des données, une instabilité la mettra en question.

En pratique : regrouper des variables en sous-ensembles est une attitude fréquente en médecine. Quand ces variables sont des symptômes, il s'agit d'ailleurs de la définition même d'un syndrome : « réunion d'un groupe de symptômes (ou de signes) qui se reproduisent en même temps dans un certain nombre de maladies. Puisqu'il peut avoir des origines diverses, le syndrome se distingue donc de la « maladie » due (en principe) à une cause spécifique » ([3]).

A partir de données sur la sémiologie des douleurs aiguës de l'abdomen, nous allons essayer d'objectiver des syndromes par une classification hiérarchique ascendante. Cela revient, en fait, à regrouper les symptômes qui sont les plus fréquemment associés. Nous disposons de 24 variables mesurées chez 4 541 patients de sexe masculin ([4]) :

- douleur provoquée à la palpation de l'hypochondre droit (hypoc_d) ;
- douleur épigastrique (epig) ;
- douleur de l'hypochondre gauche (hypoc_g) ;
- douleur de la fosse iliaque droite (fid) ;
- douleur de la fosse iliaque gauche (fig) ;
- douleur para-ombilicale (para_omb) ;
- irradiations lombaires (irr_lomb) ;
- irradiations descendantes (irr_desc) ;
- arrêt des gaz (arre_gaz) ;
- arrêt des matières (arre_mat) ;
- pollakiurie (pollakiu) ;
- hématurie (hematuri) ;
- antécédents de chirurgie abdomino-pelvienne (atcdchir) ;
- température (temp) ;
- agitation (agit) ;
- abolition de la respiration abdominale (respabdo) ;
- météorisme abdominal (meteoris) ;
- défense abdominale (defense) ;
- signe de Murphy (murphy) ;
- contracture abdominale (contract) ;
- tympanisme abdominal (tympanis) ;
- disparition des bruits hydro-aériques (BHA) ;

[3] *in « Dictionnaire des termes techniques de médecine »* par M. Garnier et V. Delamare.
[4] Données aimablement fournies par les associations de recherche en chirurgie.

– douleur à droite au toucher rectal (tr_droit) ;
– douleur dans le cul-de-sac de Douglas au toucher rectal (tr_doug).

Les données se présentent ainsi sous la forme :

obs	hypoc_d	epig	hypoc_g	fid	fig	ombil	...	tr_doug
1	0	0	0	1	0	1	...	0
2	0	0	0	0	1	0	...	0
3	0	0	1	0	1	0	...	0
...	...	...	...	...	...	...	...	...
4541	1	1	0	0	0	0	...	0

Les logiciels d'analyse en clusters sont souvent prévus pour classer des sujets. Il nous faut donc envisager dans un premier temps de transposer ce tableau (5). Nous obtenons finalement :

obs	1	2	3	4	5	6	...	4541
hypoc_d	0	0	0	0	1	0	...	1
epig	0	0	0	0	0	1	...	1
hypoc_g	0	0	1	0	0	1	...	0
...	...	...	...	...	...	...	...	...
tr_doug	0	0	0	0	0	0	...	0

La procédure CLUSTER du logiciel SAS nous donne les résultats suivants (6) :

Average Linkage Cluster Analysis ❶

Root-Mean-Square Total-Sample Standard Deviation = 0.961837
Root-Mean-Square Distance Between Observations = 91.66259

NCL	-Clusters Joined- ❷		FREQ	SPRSQ	RSQ	ERSQ	CCC ❹	PSF ❺	PST2 ❻	Norm RMS Dist ❼	Tie
23	METEORIS	TYMPANIS ❸	2	0.01748	0.983	.	.	2.6	.	0.6342	
22	HYPOC_G	FIG	2	0.01798	0.965	.	.	2.6	.	0.6431	
21	HYPOC_D	EPIG	2	0.02194	0.943	.	.	2.5	.	0.7104	
20	ARRE_MAT	ARRE_GAZ	2	0.02258	0.920	.	.	2.4	.	0.7206	
19	CL22	PARA_OMB	3	0.02674	0.893	.	.	2.3	1.5	0.7514	
18	CL21	CL19	5	0.03622	0.857	.	.	2.1	1.6	0.8029	
17	RESPABDO	CONTRACT	2	0.02836	0.829	.	.	2.1	.	0.8077	
16	CL23	BHA	3	0.03224	0.796	.	.	2.1	1.8	0.8103	
15	FID	TR_DROIT	2	0.02961	0.767	.	.	2.1	.	0.8252	
14	CL20	CL16	5	0.03840	0.728	.	.	2.1	1.6	0.8297	
13	CL17	TR_DOUG	3	0.03660	0.692	.	.	2.1	1.3	0.8913	
12	CL18	CL13	8	0.05858	0.633	.	.	1.9	2.1	0.9192	
11	CL14	ATCDCHIR	6	0.04423	0.589	.	.	1.9	1.6	0.9300	
10	TEMP	DEFENSE	2	0.03826	0.551	.	.	1.9	.	0.9381	
9	CL12	CL11	14	0.08366	0.467	.	.	1.6	2.6	0.9503	
8	IRR_LOMB	MURPHY	2	0.04003	0.427	.	.	1.7	.	0.9595	
7	IRR_DESC	POLLAKIU	2	0.04162	0.385	.	.	1.8	.	0.9783	
6	CL9	AGITE	15	0.04947	0.336	.	.	1.8	1.4	0.9958	
5	CL15	CL10	4	0.05267	0.283	.	.	1.9	1.6	0.9980	
4	CL7	HEMATURI	3	0.04493	0.238	0.216	1.08	2.1	1.1	1.0072	
3	CL8	CL4	5	0.05244	0.186	0.137	3.06	2.4	1.2	1.0317	
2	CL6	CL3	20	0.09055	0.095	0.056	4.55	2.3	2.4	1.0411	
1	CL2	CL5	24	0.09541	0.000	0.000	0.00	.	2.3	1.0614	

Ward's Minimum Variance Cluster Analysis ❽

NCL	-Clusters Joined-		FREQ	SPRSQ ❾	RSQ	ERSQ	CCC	PSF	PST2	Tie
23	METEORIS	TYMPANIS	2	0.017485	0.9825	.	.	2.6	.	
22	HYPOC_G	FIG	2	0.017979	0.9645	.	.	2.6	.	
21	HYPOC_D	EPIG	2	0.021940	0.9426	.	.	2.5	.	
20	ARRE_MAT	ARRE_GAZ	2	0.022577	0.9200	.	.	2.4	.	
19	CL22	PARA_OMB	3	0.026738	0.8933	.	.	2.3	1.5	
18	RESPABDO	CONTRACT	2	0.028364	0.8649	.	.	2.3	.	
17	FID	TR_DROIT	2	0.029609	0.8353	.	.	2.2	.	
16	CL23	BHA	3	0.032239	0.8031	.	.	2.2	1.8	
15	CL21	CL19	5	0.036223	0.7668	.	.	2.1	1.6	
14	CL18	TR_DOUG	3	0.036601	0.7302	.	.	2.1	1.3	
13	TEMP	DEFENSE	2	0.038263	0.6920	.	.	2.1	.	
12	CL20	CL16	5	0.038401	0.6536	.	.	2.1	1.6	
11	IRR_LOMB	MURPHY	2	0.040031	0.6136	.	.	2.1	.	
10	IRR_DESC	POLLAKIU	2	0.041615	0.5719	.	.	2.1	.	
9	ATCDCHIR	AGITE	2	0.042769	0.5292	.	.	2.1	.	
8	CL10	HEMATURI	3	0.044933	0.4842	.	.	2.1	1.1	
7	CL11	CL9	4	0.046419	0.4378	.	.	2.2	1.1	
6	CL17	CL13	4	0.052670	0.3851	.	.	2.3	1.6	
5	CL7	CL8	7	0.052699	0.3324	.	.	2.4	1.2	
4	CL15	CL14	8	0.058576	0.2739	0.2161	2.870	2.5	2.1	
3	CL6	CL5	11	0.078204	0.1957	0.1372	3.693	2.6	1.8	
2	CL4	CL12	13	0.082606	0.1131	0.0562	6.659	2.8	2.7	
1	CL2	CL3	24	0.113061	0.0000	0.0000	0.000	.	2.8	

Le type de distance entre clusters est précisé en ❶ (« average linkage » correspond à la distance moyenne entre chaque paire de points, on dit encore « la distance du lien moyen »).

La liste des variables regroupées à chaque étape est en ❷. Ainsi, les deux variables « météorisme » et « tympanisme » ❸ sont-elles regroupées en premier car ces deux signes sont les plus fortement associés. En ❹, ❺ et ❻, nous trouvons trois outils permettant (en théorie) de se faire une idée du nombre de clusters objectivement présents (*cf.* ci-après). Enfin, se trouve en ❼ la distance d séparant les deux variables venant d'être regroupées.

Comme précisé plus haut, il est utile de faire la même analyse à partir d'une autre définition de la distance entre clusters. En ❽ est utilisée la méthode de Ward (une des plus courantes). Les résultats se présentent sous la même forme, à l'exception de la distance entre variables agrégées qui est maintenant en ❾.

5 Quand cela est nécessaire, il est aussi très important de normaliser les données (de les centrer et de les diviser par leur écart type). En effet, si, contrairement aux autres variables, tr_doug et tr_droit avaient toutes deux été cotées en 1, 2, 3 et 4, n'importe quel algorithme les aurait classées artificiellement dans le même cluster du simple fait de leur similitude de cotation. Voir p. 313 chapitre « Analyse en composantes principales », pour une discussion plus approfondie de ce problème.

6 Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Pour interpréter facilement ces résultats, il est indispensable de les représenter sous forme d'une arborescence. Nous obtenons ainsi pour chaque méthode :

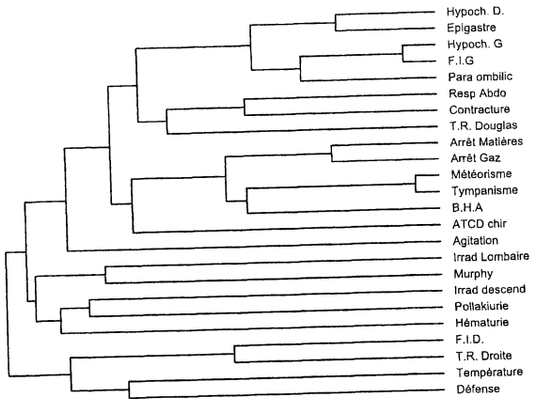


Fig. 3.1 — *Représentation graphique par arborescence d'une classification hiérarchique ascendante (méthode du lien moyen).*

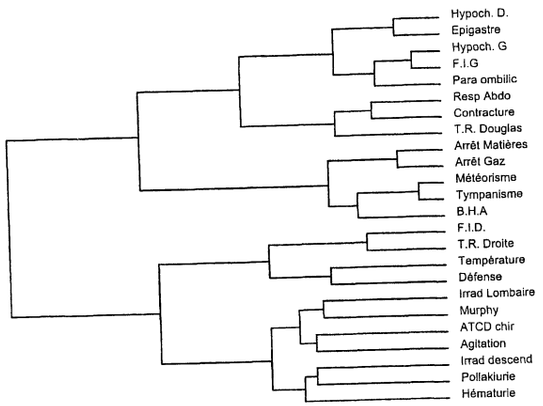


Fig. 3.2 — *Représentation graphique par arborescence d'une classification hiérarchique ascendante (méthode de Ward).*

Visuellement, que retrouvons-nous ? On peut envisager :

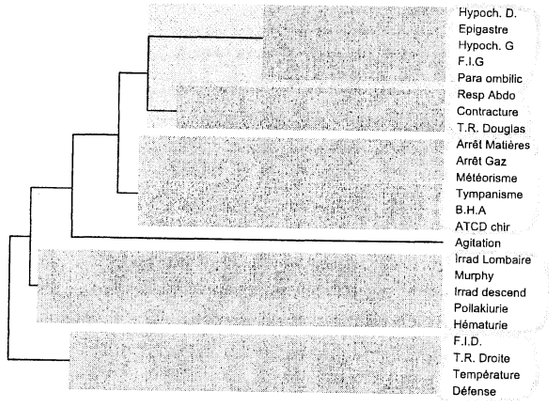


Fig. 3.3 — *Regroupement de symptômes à partir d'une classification hiérarchique ascendante (méthode du lien moyen).*

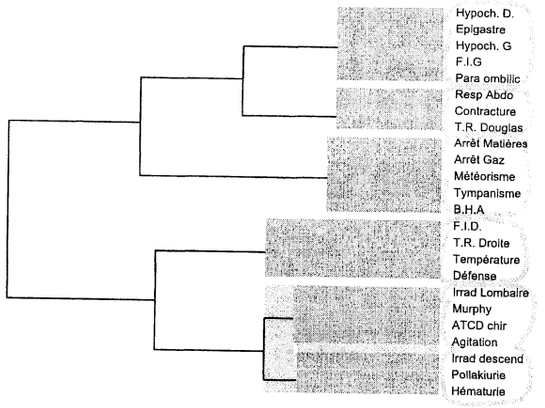


Fig. 3.4 — *Regroupement de symptômes à partir d'une classification hiérarchique ascendante (méthode de Ward).*

Dans les deux cas il y a 4 ou 5 syndromes. Trois d'entre eux sont en bonne adéquation avec des entités cliniques préexistantes :

- le syndrome occlusif (arrêt des matières, arrêt des gaz, météorisme, tympanisme, disparition des bruits hydroaériques et dans un des deux cas antécédents chirurgicaux) ;
- le syndrome appendiculaire (douleur provoquée à la palpation de la fosse iliaque droite, douleur à droite au toucher rectal, température, défense) ;
- la péritonite (abolition de la respiration abdominale, contracture et douleur dans le cul-de-sac de Douglas au toucher rectal).

Nous trouvons aussi l'esquisse d'un syndrome urologique de type « colique néphrétique » : hématurie, irradiation lombaire, irradiation descendante, pollakiurie. L'agitation n'est retrouvée que dans un cas ; en revanche, le signe de Murphy (en faveur d'une cholécystite aiguë) est retrouvé dans les deux cas.

Enfin, il est possible de décrire un syndrome douloureux abdominal général aspécifique avec douleurs pan-abdominales : hypochondres droit et gauche, épigastre, fosse iliaque gauche et para-ombilicale.

En conclusion, le découpage sémiologique traditionnel est, dans l'ensemble, confirmé par cette analyse objective : les syndromes, tels qu'ils ont été construits à partir de l'expérience clinique, correspondent bien aux groupements de symptômes les plus compacts (les plus corrélés).

Nous avons choisi arbitrairement le nombre de clusters. Que propose le programme ? Les critères objectifs calculés sont conçus pour présenter un maximum local au voisinage du nombre optimal de clusters. Pour les interpréter facilement, on les représente graphiquement. Nous obtenons ainsi pour la méthode du lien moyen :

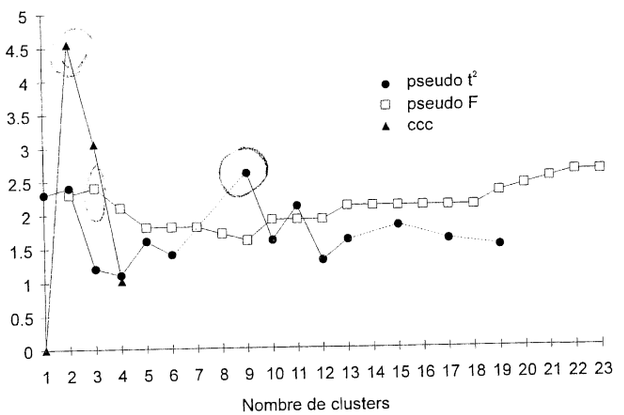


Fig. 3.5 — *Indices permettant de déterminer un nombre optimal de clusters correspondant à un maximum local des courbes (méthode du lien moyen).*

En ce qui concerne le score du pseudo F, un maximum local est obtenu pour n = 3 clusters, le pseudo t^2 propose lui n = 2, 5, 11 voire 9, le ccc (*cubic cluster criterion*) n = 2. La valeur n = 2 semble la plus proche d'un consensus, elle est néanmoins peu compatible avec la structure de l'arbre. Ces résultats sont donc à prendre avec prudence. Voyons ce que donne la méthode de Ward :

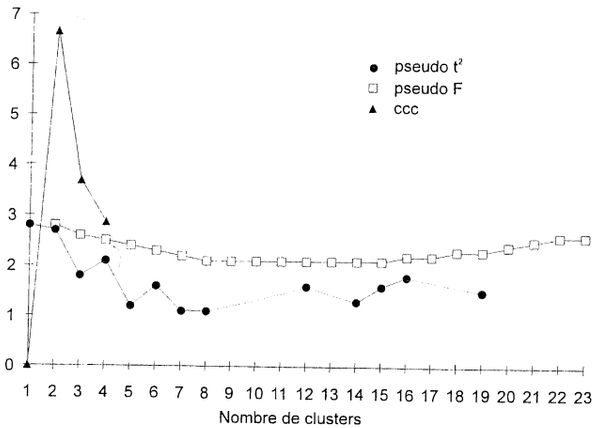


Fig. 3.6 — *Indices permettant de déterminer un nombre optimal de clusters correspondant à un maximum local des courbes (méthode de Ward).*

Le ccc penche toujours pour n = 2, le pseudo F pour aucune valeur particulière, et le pseudo t^2 pour n = 4 et n = 6. Il n'y a donc pas de solution évidente, on peut finalement conclure soit que le découpage en clusters n'est pas net, soit que ces méthodes ne sont pas efficaces.

La méthode des nuées dynamiques

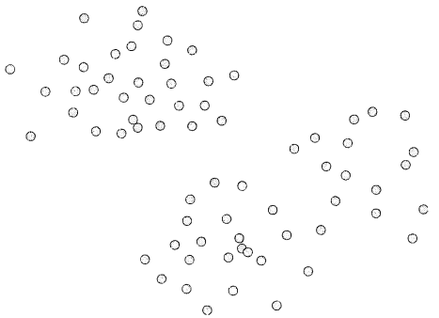
En quelques mots : si ce type d'approche rassemble plusieurs techniques (agrégation autour de centres mobiles, méthode des k-moyennes, nuées dynamiques proprement dites (7)), elles ne diffèrent cependant que par quelques détails que nous négligerons.

Leur principal avantage est sûrement la simplicité et la rapidité des calculs à mettre en œuvre. A l'opposé, elles présentent deux inconvénients majeurs : le découpage proposé est souvent grossier, et il est nécessaire de préciser *a priori* le nombre de clusters à rechercher. De ce fait, la méthode des nuées dynamiques est

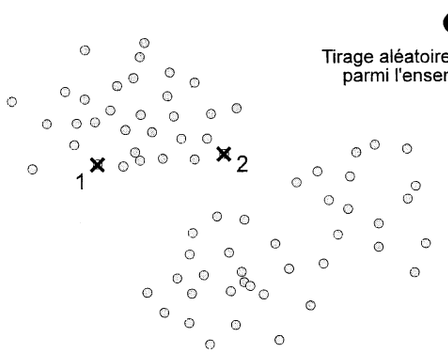
[7] Nous n'étudierons d'ailleurs pas ici la véritable méthode des nuées dynamiques. La nomenclature changeant d'un auteur à l'autre, le terme à la fois le plus français et le plus parlant a finalement été retenu.

le plus souvent utilisée, au début d'une étude, pour fragmenter un gros jeu de données en sous-ensembles plus maniables. A l'opposé, elle n'est, telle quelle, que de peu d'utilité pour classer des variables.

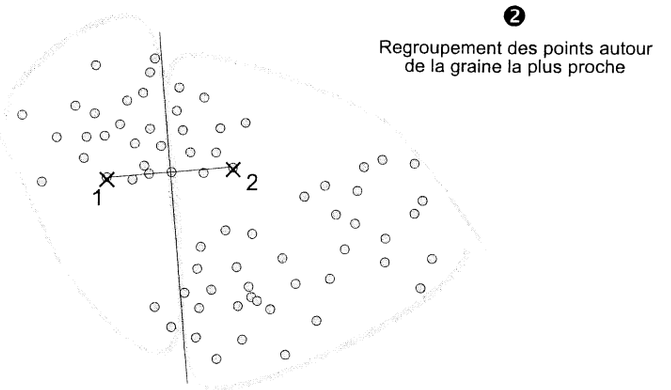
Nous allons voir sur quelques schémas le déroulement de l'algorithme de regroupement. Considérons l'ensemble de points suivant :



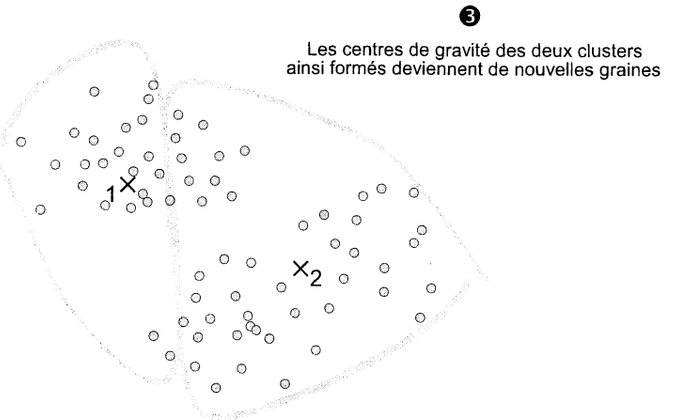
Nous désirons le fragmenter en deux parties les plus homogènes possible. Dans un premier temps, choisissons deux points au hasard que nous baptiserons « graines », elles serviront d'amorce au déroulement du programme :



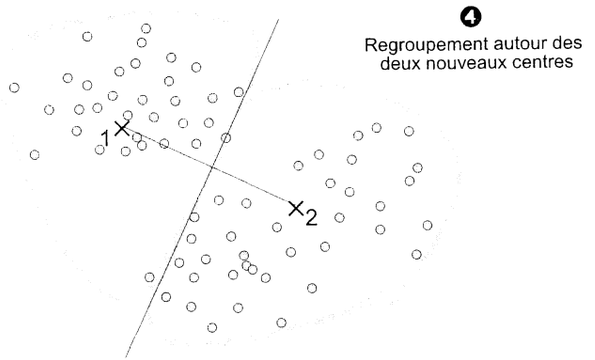
A partir de ces deux graines, découpons deux premiers clusters en agrégeant les points autour de la graine la plus proche :



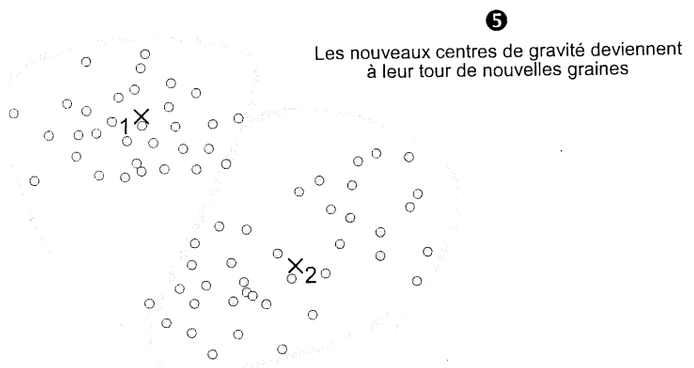
Recherchons maintenant le centre de gravité de ces deux clusters, et donnons-leur le statut de nouvelles graines :



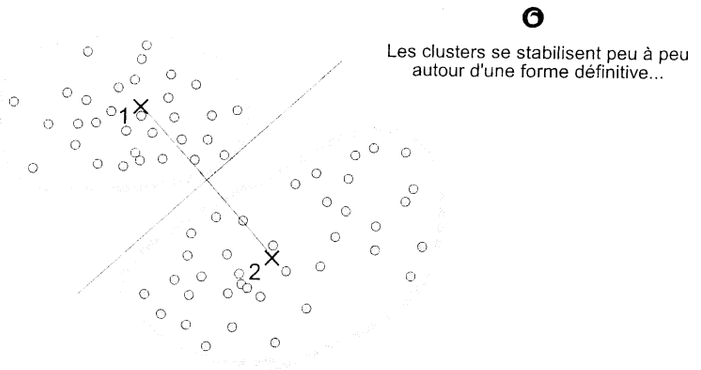
Un nouveau découpage en deux clusters peut ainsi se réaliser en rassemblant les points autour du centre de gravité le plus proche :



Une telle procédure est ainsi réitérée plusieurs fois :



jusqu'à ce que les *clusters* finissent par se stabiliser.



En pratique : voyons maintenant cette technique à l'œuvre sur de véritables données. Reprenons notre exemple sur les douleurs aiguës de l'abdomen. Nous avons montré dans la partie précédente que les symptômes se regroupaient de façon cohérente en syndromes, mais rien ne prouve que les patients puissent aussi se rassembler en catégories homogènes (un patient pouvant notamment présenter plusieurs syndromes douloureux).

Le nombre de patients étant élevé (4 541), pour des raisons calculatoires, seule la méthode des nuées dynamiques est envisageable. Un premier problème se pose aussitôt : une telle technique nécessite de fixer le nombre de clusters. Que choisir ici ? Deux solutions complémentaires sont envisageables : décider du nombre de catégories cliniquement cohérentes, ou alors, rechercher une solution comptant 2, 3 puis 4, 5, 6, 7, 8, 9 et 10 clusters pour finalement utiliser les différentes familles d'indices objectifs présentées plus haut afin de déterminer la solution optimale.

Dans le cas présent, la situation est un peu particulière puisque nous connaissons en outre la pathologie dont souffre chaque patient (diagnostic de référence), cinq diagnostics étant prépondérants : appendicite aiguë, cholécystite aiguë, colique néphrétique, occlusion, péritonite. Nous allons donc nous pencher dans un premier temps sur une solution en cinq clusters, nous verrons dans un second temps si ce nombre de clusters est le plus approprié.

La procédure FAST CLUST du logiciel SAS nous donne les résultats suivants pour n = 5 clusters ([8]) :

FASTCLUS Procedure: Replace=FULL Radius=0 Maxclusters=5 Maxiter=20 ❶
Converge=0.02

Minimum Distance Between Initial Seeds = 4.123106

Relative Change in Cluster Seeds ❷

Iteration	Criterion	1	2	3	4	5
1	0.6116	0.5294	0.5875	0.5139	0.5090	0.6163
2	0.3766	0.0997	0.1484	0.0796	0.0826	0.0550
3	0.3663	0.0434	0.0364	0.0758	0.0792	0.0267
4	0.3623	0.0244	0.0591	0.0708	0.0573	0.0298
5	0.3588	0.0222	0.0899	0.0709	0.0342	0.0188
6	0.3556	0.0347	0.0736	0.0359	0.0296	0.0318
7	0.3531	0.0262	0.0511	0.0255	0.0436	0.0262
8	0.3516	0.0145	0.0337	0.00935	0.0257	0.0111
9	0.3510	0.0290	0.00967	0.00285	0.0429	0.0347
10	0.3496	0.0204	0.00381	0.00380	0.0136	0.0153
11	0.3493	0.0105	0.00489	0.00148	0.00470	0.00584

Convergence criterion is satisfied.
Criterion Based on Final Seeds = 0.34929

Cluster Summary ❸

Cluster	Frequency ❹	RMS Std Deviation	Maximum Distance from Seed to Observation	Nearest Cluster	Distance Between Cluster Centroids
1	712	0.3689	2.8591	5	1.3179 ❺
2	384	0.4300	3.3212	3	2.1252
3	438	0.4191	2.8497	2	2.1252
4	411	0.3563	3.0450	5	1.6934
5	1200	0.2698	2.6692	1	1.3179

Pseudo F Statistic = 431.64 ❻
Approximate Expected Over-All R-Squared = 0.21936
Cubic Clustering Criterion = 92.800 ❼
WARNING: The two above values are invalid for correlated variables.

Cluster Means ❽

Cluster	HYPOC_D	EPIG	HYPOC_G	FID	FIG	PARA_OMB
1	0.61376	0.30618	0.08848	0.15449	0.09691	0.14747
2	0.99479•	0.99219•	0.94792•	0.95833•	0.88542•	0.96094•
3	0.27169	0.16438	0.07763	0.47260	0.19635	0.24886
4	0.40633	0.17762	0.04866	0.63017	0.09489	0.11436
5	0.03917	0.01417	0.00417	0.97667•	0.04750	0.03750

Cluster Means

Cluster	IRR_LOMB	IRR_DESC	ARRE_MAT	ARRE_GAZ	POLLAKIU	HEMATURI
1	0.38764•	0.08006	0.21348	0.16713	0.06180•	0.02107•
2	0.15885	0.04948	0.53906	0.55208	0.04167	0.00260
3	0.23516	0.08219	0.62100•	0.58676•	0.04338	0.00457
4	0.20438	0.06326	0.18978	0.14112	0.05109	0.00730
5	0.10750	0.09417•	0.13667	0.08250	0.04250	0.00833

[8] Les données et les syntaxes sas et R de cet exemple sont disponibles sur le site Internet du livre.

Cluster Means

Cluster	ATCDCHIR	TEMP	AGITE	RESPABDO	METEORIS	DEFENSE
1	0.39747	0.48315	0.19101•	0.13062	0.02528	0.32584
2	0.38281	0.72917	0.19010•	0.52865•	0.51771	0.33854
3	0.44292•	0.76484	0.12329	0.24429	0.93333•	0.39498
4	0.20438	0.90195•	0.07543	0.14112	0.08516	0.58637•
5	0.06833	0.64500	0.06333	0.05833	0.01417	0.43417

Cluster Means

Cluster	MURPHY	CONTRACT	TYMPANIS	BHA	TR_DROIT	TR_DOUG
1	0.28792•	0.04213	0.11657	0.22331	0.03792	0.06882
2	0.01563	0.28385•	0.56250	0.73177•	0.11458	0.42188•
3	0.08447	0.02740	0.76484•	0.73516•	0.13927	0.13699
4	0.17762	0.04623	0.09732	0.15085	0.33577	0.10706
5	0.01417	0.00750	0.03417	0.08917	0.54917•	0.10500

En ❶ sont rappelés les paramètres utilisés par l'algorithme. En ❷, nous trouvons le détail des itérations : les modifications de chaque graine sont précisées, ces modifications vont, comme prévu, en s'amenuisant pour devenir négligeables à la 11e itération.

En ❸ nous trouvons une présentation générale des clusters, avec notamment pour chacun d'entre eux : en ❹ l'effectif qu'il regroupe et en ❺ la distance du cluster le plus proche. En ❻ et ❼ nous retrouvons deux indices permettant d'objectiver (tant que faire se peut) un nombre optimal de clusters.

Enfin en ❽, les moyennes de chaque variable sont répertoriées pour chacun des clusters. Ces variables étant codées en 0 et 1, leur moyenne correspond à la proportion de patients présentant chaque élément sémiologique. Ce sont ces résultats qui vont permettre d'étiqueter les clusters.

Comment interpréter les données obtenues ? Pour chaque variable, nous avons coché par un « • » le cluster qui regroupe le plus de patients (il y a parfois des *ex aequo*) ; par exemple, pour la variable hypoc_d, le cluster 2 regroupe le plus de patients (moy = 0,99479), pour la variable irr_lomb c'est le 1 (moy = 0,38764)... Nous obtenons ainsi :

- le cluster 1 : irradiation lombaire, hématurie, agitation et signe de Murphy ; symptomatologie évocatrice d'une crise de colique néphrétique (à l'exception du signe de Murphy, dont nous avons déjà vu qu'il posait problème) ;
- le cluster 2 : douleur pan-abdominale, agitation, abolition de la respiration abdominale, contracture, arrêt des bruits hydroaériques, douleur dans le cul-de-sac de Douglas au toucher rectal ; ces symptômes évoquent fortement une péritonite ;
- le cluster 3 : arrêt des matières, arrêt des gaz, météorisme abdominal, tympanisme, arrêt des bruits hydroaériques ; on pense ici à une occlusion intestinale ;
- le cluster 4 : température et défense... Ce qui manque quelque peu de spécificité, mais serait plutôt en faveur d'une crise d'appendicite aiguë ;

– le cluster 5 : douleur à la palpation de la fosse iliaque droite, irradiation descendante et douleur à droite au toucher rectal ; ceci est en faveur d'une crise d'appendicite aiguë.

Que donne maintenant le croisement de ces 5 clusters avec les 5 diagnostics dont on peut disposer parallèlement. Par la routine PROC FREQ, nous obtenons :

TABLE OF CLUSTER BY DIAG

CLUSTER(Cluster) DIAG

Frequency / Percent / Row Pct / Col Pct	appendic	cholecys	neph_col	occlus	perit_pr	Total
1	61	362•	108•	98	83	712
	1.94	11.51	3.43	3.12	2.64	22.64
	8.57	50.84	15.17	13.76	11.66	
	3.95	58.58	55.10	20.29	27.21	
2	71	20	4	142•	147•	384
	2.26	0.64	0.13	4.52	4.67	12.21
	18.49	5.21	1.04	36.98	38.28	
	4.60	3.24	2.04	29.40	48.20	
3	88	77	20	215•	38	438
	2.80	2.45	0.64	6.84	1.21	13.93
	20.09	17.58	4.57	49.09	8.68	
	5.70	12.46	10.20	44.51	12.46	
4	238•	138•	8	8	19	411
	7.57	4.39	0.25	0.25	0.60	13.07
	57.91	33.58	1.95	1.95	4.62	
	15.42	22.33	4.08	1.66	6.23	
5	1085•	21	56	20	18	1200
	34.50	0.67	1.78	0.64	0.57	38.16
	90.42	1.75	4.67	1.67	1.50	
	70.32	3.40	28.57	4.14	5.90	
Total	1543	618	196	483	305	3145
	49.06	19.65	6.23	15.36	9.70	100.00

Le premier cluster apparaît comme un mélange de coliques néphrétiques et de cholécystites, ce qui explique la présence fréquente d'un signe de Murphy. Le deuxième cluster est une association de péritonites et d'occlusions. Le troisième est constitué presque exclusivement d'occlusions. Le quatrième est un mélange d'appendicites et de cholécystites et finalement le dernier rassemble principalement des crises d'appendicites aiguës.

Quels enseignements tirer de ces résultats ? Globalement, la symptomatologie douloureuse abdominale regroupe les patients en classes proches de celles générées par les diagnostics, mais il existe vraisemblablement des formes cliniques trompeuses : cholécystites et appendicites peuvent parfois s'exprimer sur un mode proche (cluster 4, penser aux appendicites sous-hépatiques), une occlusion peut prendre le masque d'une péritonite (ou *vice versa*, cluster 2), enfin,

certaines cholécystites semblent prendre l'allure d'une crise de colique néphrétiques (cluster 1).

Nous avons choisi un découpage en cinq clusters parce qu'il était suggéré par l'existence de cinq catégories diagnostiques, mais ce nombre est-il corroboré par les indices objectifs que nous propose le programme ? Pour répondre à cette question, le programme SAS, PROC FASTCLUS va successivement envisager les valeurs de n = 2, 3, 4, ..., 10 clusters. Finalement, nous obtenons :

```
N=2                          Pseudo F Statistic =   677.48
          Approximate Expected Over-All R-Squared =  0.09765
                       Cubic Clustering Criterion =   59.063
N=3                          Pseudo F Statistic =   520.06
          Approximate Expected Over-All R-Squared =  0.17003
                       Cubic Clustering Criterion =   46.492
N=4                          Pseudo F Statistic =   394.83
          Approximate Expected Over-All R-Squared =  0.19928
                       Cubic Clustering Criterion =   46.237
N=5                          Pseudo F Statistic =   431.64
          Approximate Expected Over-All R-Squared =  0.21936
                       Cubic Clustering Criterion =   92.800
N=6                          Pseudo F Statistic =   344.57
          Approximate Expected Over-All R-Squared =  0.23555
                       Cubic Clustering Criterion =   84.482
N=7                          Pseudo F Statistic =   293.97
          Approximate Expected Over-All R-Squared =  0.24892
                       Cubic Clustering Criterion =   81.940
N=8                          Pseudo F Statistic =   307.69
          Approximate Expected Over-All R-Squared =  0.26018
                       Cubic Clustering Criterion =  116.307
N=9                          Pseudo F Statistic =   295.86
          Approximate Expected Over-All R-Squared =  0.26995
                       Cubic Clustering Criterion =  133.123
N=10                         Pseudo F Statistic =   279.73
          Approximate Expected Over-All R-Squared =  0.27839
                       Cubic Clustering Criterion =  144.622
```

Traçons sur un schéma les valeurs du *cubic clustering criterion* (ccc) et du pseudo F. Nous obtenons :

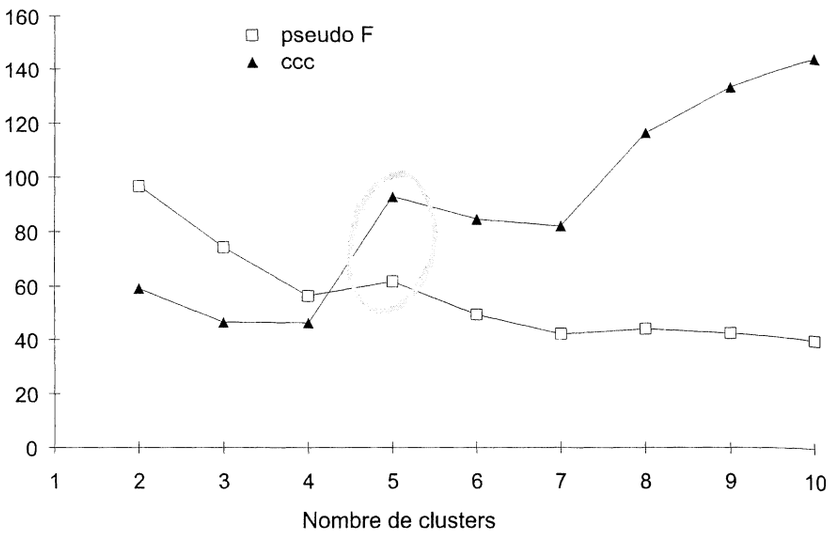


Fig. 3.7 — *Indices permettant de déterminer un nombre optimal de clusters (ce nombre correspond à un maximum local des courbes).*

Les deux courbes présentent chacune un seul maximum local, ce dernier est obtenu dans les deux cas pour n = 5.

4.
Méthodes de segmentation, CART

Les méthodes de segmentation, dont la méthode « CART » (*classification and regression tree*) est la plus utilisée, peuvent être considérées comme un compromis entre les méthodes de classification hiérarchiques et les modèles de régression.

Nous retrouvons en effet la dichotomie variable à expliquer / variables explicatives, mais, contrairement à un modèle de régression, l'objectif est ici de segmenter l'échantillon au moyen des variables explicatives, de façon que les segments obtenus soient le plus homogènes possible relativement à la variable à expliquer.

En médecine, un premier exemple d'application de ces méthodes pourrait être la constitution de catégories pronostiques en cancérologie, comme les classifications TNM ([1]). Ainsi, le choix optimal des seuils choisis pour les variables T, N et M pourrait-il reposer, au moins pour partie, sur une méthode de segmentation.

Un deuxième exemple porterait sur la constitution de groupes homogènes de patients en termes de coût de prise en charge. Une telle catégorisation pourrait ensuite être utilisée par les économistes de la santé pour proposer une rémunération des soins par types de patients traités et non par actes réalisés.

En quelques mots

La méthode CART procède par itérations successives. Sur un principe voisin de celui rencontré dans les méthodes de classification hiérarchique (voir p. 275), l'échantillon étudié est découpé dans un premier temps en deux sous-groupes homogènes, puis chaque sous-groupe ainsi obtenu est à son tour segmenté en deux parties, etc.

Sur un plan technique, il faut cependant définir formellement ce qu'est un « sous-groupe homogène » de sujets.

Prenons un exemple comprenant une variable à expliquer Y et une variable explicative X, toutes deux quantitatives (fig. 4.1). Si l'on examine les valeurs x_i prises par X, on ne constate aucun sous-groupe homogène de sujets. Il en est d'ailleurs de même pour les y_i, valeurs prises par Y.

[1] Si l'on prend l'exemple du cancer de la prostate, cette classification caractérise le stade de la maladie suivant que la tumeur (T) est cliniquement inapparente, limitée au tissu prostatique, envahissant la capsule ou fixée aux tissus avoisinants ; suivant qu'il y ait des ganglions régionaux envahis ou pas (N) ; et, enfin, suivant le niveau d'envahissement métastatique (M). En fonction du stade TNM le pronostic, et éventuellement la prise en charge thérapeutique, seront différents.