Préambule à la partie II

(présentation des programmes)

Les différents apsects méthodologiques étudiés tout le long de la partie I sont sous tendus par un ensemble important de programmes auxquels ont travaillé de nombreux chercheurs. De jour en jour ces programmes se développent avec l'apparition de situations nouvelles et s'organisent de manière de plus en plus cohérente. Nous avons annoncé dans l'introduction générale de l'ouvrage que nous préparons, sous forme séparée, un fascicule réservé à cette partie des travaux. Nous allons toutefois ici, avant d'entamer le coeur du sujet, donner un aperçu bref et général sur l'ensemble des programmes dont nous disposons.

Relativement à la *Classification*, on distingue actuellement

(a) une chaîne de programmes correspondante à la méthode de classification hiérarchique.

(b) un ensemble de trois programmes correspondants aux différentes stratégies de la méthode des "pôles d'attraction" (cf. chap. 8 § VIII).

La chaîne de programmes (a) comprend sous forme modulaire quatre étapes qui sont enchaînées :

(1) PROX dont le rôle est d'établir le tableau des proximités entre caractères descriptifs où l'ensemble des modalités d'un même caractère est soit totalement ordonné, soit sans structure (cf. chap. 2 § II - 2' et 3', § IV - 4).

Cette étape peut-être omise dans le cas où la structure des données ne lui correspond pas.

(2) ORDON qui a un double rôle

- gérer le système d'options selon que l'on organise l'ensemble des lignes ou des colonnes du tableau des données, en tenant compte de la structure de ce dernier par rapport au critère adopté pour la formation de l'arbre des classifications.

- établir l'information quant aux ressemblances entre éléments de l'ensemble à classifier dans tous les cas, sauf certains ayant un caractère spécifique tels que ceux traités dans PROX. De façon précise :

(α) on construit le tableau des indices de proximité entre

1- attributs descriptifs (cas des tableaux d'incidence)
2- variables numériques (cas des tableaux de mesures)
3- colonnes d'un tableau de contingence ou d'une juxtaposition de tels tableaux, pour une classification par A.V.L.
4- colonnes d'un tableau de contingence pour une classification par le critère de l'inertie expliquée.
5- colonnes d'un pseudo-tableau de contingence (i.e. tableau des indices bruts de proximité entre attributs) ou bien d'un tableau de "Burt" (i.e. tableau des indices bruts entre modalités des questions d'un questionnaire).

1'- individus décrits par des attributs descriptifs.
2'- individus décrits par des variables numériques.
3'- individus décrits par des variables partitions.
4'- individus décrits par des variables préordinales totales.
5'- lignes d'un tableau de contingence ou d'une juxtaposition de tels tableaux, pour une classification par A.V.L.
6'- lignes d'un tableau de contingence pour une classification par l'inertie expliquée.
7'- lignes d'un pseudo-tableau de contingence ou d'un tableau de "Burt".

Les tableaux d'indices de proximité 2,3,5 (entre variables ou bien entre colonnes d'un tableau de contingence (resp. pseudo-contingence) ; ainsi que les tableaux d'indices de proximité 2',3',4' et 5' (entre individus ou bien entre lignes d'un tableau de contingence (resp. pseudo-contingence)) conduisent à un arbre des classifications pour lequel le critère de formation des classes est A.V.L.

Les tableaux d'indices 4,6' et 7' mènent à un arbre des classifications pour lequel le critère de formation des classes est l'inertie expliquée conformément à la métrique du χ^2.

Enfin les tableaux d'indices 1 et 1' mènent au choix à un arbre des classifications où le critère est A.V.L. ou bien, celui de l'inertie expliquée, mais conformément à la métrique que suppose notre indice de proximité entre attributs (cf. chap. 2, partie I).

Pour une situation concernant la structure des données non encore intégrée à la chaîne, l'établissement des indices réduits doit se faire à partir d'un programme séparé et on entrera dans la chaîne à travers une option de l'étape ORDON ; c'est pas exemple le cas de la classification de variables totalement et strictement ordinales, ou bien totalement préordimales à "grand" nombre de modalités.

Suite au calcul du tableau des indices de proximité dont d'ailleurs, pour des raisons de symétrie, on ne stocke que la "moitié".

(β) on établit et on édite le tableau des valeurs croissantes de l'indice de neutralité de chacun des éléments de l'ensemble à classifier (cf. chap. 3 § II.3, formule (3)).

(γ) par un programme de tri rapide et efficace, on construit l'ordonnance (ordre total sur l'ensemble des paires) associée à l'indice de proximité (c'est une des fonctiosn les plus importantes de ORDON) (cf. chap. 2 § III, partie I).

(3) POLAR est le programme qui définit la troisième étape de la chaîne ; son rôle est

(α) d'établir la représentation polonaise de l'arbre des classifications pour principalement deux critères de formation hiérarchique des classes. Le premier est celui de A.V.L. (cf. chap. 5 § III. 1, partie I) qui nous est dû et qui nous a toujours fourni les résultats les plus raffinés. Le second, plus classique est celui de l'inertie expliquée où on y comprend le traitement des tableaux de contingence et d'incidence. Cette partie est organisée autour de deux sous programmes, chacun reflétant un critère de type fixé.

(β) de calculer pour chacun des niveaux les valeurs des statistiques globale et locale des niveaux (cf. chap. 5 § IV) et d'établir la représentation polonaise de l'arbre condensé des classifications où on ne retient que les niveaux où apparaît un noeud significatif, repéré par un maximum local de la distribution le long de la suite des niveaux de la statistique locale (cf. chap. 5, § V). Cette distribution, ainsi que celle correspondante de la statistique globale sont édités avec un repérage des maximums locaux sur graphique BENSON.

(4) ARBRE est la quatrième et dernière étape de la chaîne dont le rôle est d'éditer directement l'arbre condensé ; mais de telle sorte que sur cette représentation on retrouve toute l'information concernant l'arbre initial. En effet, chacun des noeuds de l'arbre détaillé se trouve marqué à la position adéquate de l'arbre réduit, avec le numéro du niveau de l'arbre total où il s'est formé. D'autre part et c'est important, le noeud se trouve affecté du signe (*) s'il est significatif ; c'est-à-dire, s'il correspond à un maximum local de la distribution le long des niveaux de l'arbre de la statistique locale des niveaux (cf. chap. 5 § V, partie I).

Précisons que la structure de POLON est telle, qu'on peut l'utiliser pour n'importe quelle notion de distance ou proximité entre parties disjointes de l'ensemble à organiser en classes de proximité.

C'est sous forme séparée que A. Prod'homme, pour ses besoins de recherche (thèse de 3ème cycle) a adapté cette partie de la chaîne de programmes pour réaliser un algorithme de la vraisemblance des liens tenant compte d'une contrainte de connexité de nature quelconque, à caractère spatial ou statistique par exemple (programme A.V.L. - CONTIGU).

Signalons que pour répondre à des questions fines et précises dans l'interprétation des résultats, on dispose de deux prgrammes CROISE et RESP. Le premier, dû à Mme Hardouin (maître assistante), réalise le croisement d'une classification "nette" et d'une classification "floue" ou bien de deux classifications "floues" (cf. chap. 3 § III, partie I). Le second programme, dû à T. Chantrel (thèse de 3ème cycle) permet, à partir de la notion de degré de responsabilité d'un individu dans la formation d'une classe d'attributs (cf. chap. 3, § II. 1, partie I), de définir dans l'ensemble des sujets ayant une certaine caractéristique exogène, la partie de ceux, les plus responsables de la formation de tel ou tel profil d'attitude.

On peut enfin indiquer le programme de H. Rostam (dans le cadre d'une thèse de 3ème cycle en cours) qui permet de calculer l'indice entre un préordre total et une relation pondérée quelconque sur un même ensemble fini, conformément au développement du paragraphe IV. 5, chap. 2, partie I.

Cet indice est aussi utilsé comme critère pour la recherche du préordre total s'ajustant "au mieux" à la relation pondérée.

Parmi les chercheurs qui ont ravaillé sur la chaîne de programmes citons :

P. Achard (1968) (programmation de l'algorithme "lexicographique", construction de la représentation polonaise de l'arbre des classifications, premier programme de représentation graphique de l'arbre des classifications) ; I.C. Lerman (1970) (programmation du calcul des statistiques globale et locale des niveaux) ; N. Nicolaü (1971) (algorithme de tri pour l'établissement de l'ordonnance) ; Mme M.H. Bacelar Nicolaü (1971) (programme de construction de la représentation polonaise de l'arbre à partir des indices de comparaison de classes sous jacents à A.V.L. et à A.M.P.) ; I. Cohen (1974) (programme PROX) ; M. MOREL (1976) (optimisation de la place de mémoire centrale utilisée dans l'étape POLAR) ; C. Chauré (1977) (installation du critère de l'inertie expliquée dans l'étape POLAR et surtout le programme ARBRE de représentation graphique de l'arbre des classifications) ; T. Chantrel (1978) (reprise et optimisation du programme de tri pour la formation de l'ordonnance) ; P. Villoing (1979) (a tiré une version personnelle pour le CELAR de Bruz) et dernièrement, M. Raphalen (1980) a analysé la totalité de la chaîne, l'a purifiée et y a intégré de nombreuses options qui correspondent notamment à des programmes de J.Y. Lafaye et de B. Tallur de calcul d'indices, permettant aussi bien la classification des colonnes que des lignes en tenant compte, bien entendu, de la structure du tableau des données.

Quant à l'avenir, notre première ambition est un important programme où on cherchera à prévoir la quasi totalité des situations possibles quant à la structure du tableau des données, pour établir la table des valeurs de l'indice de proximité entre variables descriptives ou bien entre objets ; plus généralement, entre colonnes ou bien entre lignes. Un tel programme modifierait la structure de la chaîne.

D'autre part, des chercheurs travaillent à des algorithmes de classification rapides, mettant en oeuvre nos indices et permettant le traitement de "gros" ensembles (i.e. de l'ordre de quelques milliers).

H. Leredde est l'auteur exclusif de l'ensemble des trois programmes relatifs à la méthode des "pôles d'attraction" (cf. chap. 8 § VIII) qui, dans l'état actuel, s'adresse à des tableaux de données où les variables descriptives sont de la première catégorie (attributs ou bien variables numériques).

Le premier programme (MPATS) est davantage orienté vers la classification, des paramètres de description et vers la construction de représentations euclidiennes, généralement autour des deux premiers pôles pour, notamment dégager des "sériations". L'affectation d'un élément à l'une des des classes en cours de formation se fait ici selon le critère de la plus grande proximité entre un point extérieur à la réunion des classes déjà constituées et l'une des classes en cours de formation.

Le deuxième algorithme (MPAGD) permet de déterminer la classification, classe après classe ; une même étape de l'algorithme est définie par la constitution d'une classe qu'on entraîne autour d'un pôle d'attraction par l'affectation à ce dernier de tous les éléments dont la distance est inférieure à un certain seuil δ qu'on détermine par un algorithme statis-

tique simple. Ici on travaille avec la distance associée à la métrique que suppose notre indice de proximité (cf. chap. 8 § II) et on adopte comme quantité critère le moment absolu d'ordre 2.

Pour le troisième algorithme (MPATD), le plus consistant, qu'il s'agisse de la classification des variables ou bien des objets ; on a un traitement unique, moyennant le remplacement des mesures brutes par celles, "centrées réduites" par rapport à la distribution sur l'échantillon (défini par l'ensemble des objets) des différentes variables de description. La quantité critère utilisée pour la détermination des pôles d'attraction ainsi d'ailleurs que pour la répartition autour de ces derniers des différents éléments de l'ensemble à classifier, est basée sur le moment d'inertie. Ici, on tente de choisir les différents pôles d'attraction aussi "éloignés" que possible les uns des autres.

Comme dans le cas de la classification hiérarchique, on associe à la suite des classifications (premier et troisième programmes), la suite des valeurs de deux statistiques de signification, ponctuée par l'édition de deux histogrammes. L'examen de ces deux distributions permet de retenir les "meilleures" classifications.

Il serait souhaitable de pouvoir introduire au niveau de ces programmes des options plus souples quant au choix de la métrique et pour une prise en compte d'une classe plus large de structures de données.

Pour terminer en ce qui concerne la classification, signalons l'ensemble des programmes sur la "classificabilité" (cf. chap. 3 § IV) que nous avons nous mêmes mis au point (1970). Ces programmes permettent d'une part de calculer la distribution $D(\omega)$ (cf. Chap. 3 § IV-2) qui caractérise le degré de classifiabilité de ω, dans un cas réel ou simulé et d'autre part, de calculer une telle distribution pour un tableau d'incidence aléatoire dans l'hypothèse N_1 d'absence de lien, où le nombre de composantes égales à 1 dans la description d'un même élément, est constant.

On dispose enfin d'un ensemble cohérent de programmes de recherche de l'"échelle hiérarchique" sous jacente à une classe d'items dont chacun définit un caractère à l'ensemble totalement ordonné des modalités (cf. chap. 9 § III). Ce sont les écarts e_1 et e_2 qui ont été programmés dans l'algorithme de construction de l'échelle (Programmation de C. Riso-Lévy (1969)).

Signalons que, de façon générale, nos programmes sont d'exécution rapide, mais demandent une place mémoire appréciable.

Nous allons aborder les différentes études ; chacune sera précédée d'une fiche technique portant les informations suivantes :

A- Types des tableaux de données traités et dimensions respectives.

B- Nature de l'ensemble dont on organise les liens entre éléments. S'agit-il de l'ensemble des variables de description ou de l'ensemble des objets. Types généraux d'analyses effectuées : Classification hiérarchique ou non hiérarchique, représentation euclidienne et analyse factorielle, recherche d'échelles d'attitude...

C- Types d'indices utilisés et mode de réduction globale des similarités (surtout dans le cas de la classification hiérarchique qui dominera

dans nos traitements).

D- Méthodes et programmes de formation de la structure de synthèse, en classes ou bien ordinale.

E- Aspects méthodologiques testés.

F- Auteurs du travail d'analyse et du texte.

G- Spécialistes ayant fourni les données et participé au dialogue et à l'évaluation des résultats.