UNIVERSITÉ DE NEUCHÂTEL – FACULTÉ DES SCIENCES

CLASFAC-APL

UN PACKAGE INTERACTIF D'ANALYSE STATISTIQUE MULTIVARIABLE

Thèse
présentée à la Faculté des Sciences
par

ANNE-CHANTAL MEZGER-VOIDE

diplômée en science
actuarielle
pour l'obtention du grade de
docteur ès sciences

1983

IMPRIMATUR POUR LA THÈSE

CLASFAC-APL: un package interactif d'analyse statistique multivariable

de M adame Anne-Chantal Mezger-Voide

UNIVERSITÉ DE NEUCHATEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury,

Messieurs les professeurs A. Strohmeier, P. Banderet et J. Kohlas (Fribourg)

autorise l'impression de la présente thèse.

Neuchâtel, le 3 mars 1983

Le doyen:

A. Aeschlimann

A Tony

CLASFAC-APL est un package (ensemble de programmes) auto-documenté et développé pour des utilisateurs dépourvus de connaissances informatiques. Il permet de traiter des tableaux de données par des méthodes d'analyse statistique multivariable. Les techniques implantées appartiennent à l'analyse factorielle et à la classification automatique.

Ce présent ouvrage est découpé en trois parties.

La première partie explique la façon d'utiliser le package. Elle présente sa structure, son contenu, les principes du dialogue, le cheminement dans le package.
Après une lecture attentive, l'utilisateur statisticien sera capable de commencer une première analyse en s'aidant des modules utilitaires MENU et HELP.

La deuxième partie a pour but de donner quelques connaissances statistiques à un utilisateur débutant. Elle précise également les techniques mathématiques qui ont été implantées et donne les options disponibles. En étudiant cette partie, l'utilisateur averti apprendra à se servir plus efficacement du package.

La troisième partie décrit quelques aspects de la réalisation informatique.

Le package a été écrit en APL. Au moment de la rédaction de ce texte deux versions existent pour le système VSPC d'IBM et l'ordinateur VAX/VMS de Digital Equipment Corporation. Une version pour le système CMS d'IBM est en préparation.

Pour les descriptions des manipulations sur terminal, nous nous sommes basés sur les systèmes d'IBM. A quelques modifications de détail près, ce manuel reste cependant valable pour VAX/VMS.

Plusieurs chercheurs non informaticiens de la Division économique et sociale de l'Université de Neuchâtel ont utilisé CLASFAC-APL pour effectuer des analyses statisti-

quea.
En aociologie, le package a été utiliaé dans diveraea recherches et travaux de thèaes. Une étude a portée sur la participation politique de la commune de Rougemont, une autre sur l'évolution dea effectifs des étudianta de l'Univeraité de Neuchâtel entre 1968 et 1977 et une troiaième avait comme objet l'étude socio-culturelle des agriculteurs du Val-de-Travers. En marketing, différentea étudea de marché ont été soumises à CLASFAC, notamment une étude sur les images de marques de montres et une autre aur le comportement d'achat. En économie appliquée, les résultats statistiques de l'étude "Méthode d'analyse des structures et de l'évolution de marchés régionaux de l'emploi industriel en Suisse: étude comparative de quelques régiona types" ont été obtenus avec CLASFAC.

L'utilisation du package par dea tiera nous a permis d'améliorer les dialogues et noua a montré en même temps que CLASFAC-APL correspondait à un réel besoin.

PREMIERE PARTIE : LES POSSIBILITES DE CLASFAC-APL

TROISIEME PARTIE : QUELQUES ASPECTS DE LA REALISATION INFORMATIQUE

LES POSSIBILITES DE CLASFAC-APL

1. INTRODUCTION

CLASFAC-APL est un package (ensemble de programmes) auto-documenté et développé pour des utilisateurs dépourvus de connaissances informatiques. Il permet de traiter des tableaux de données par des méthodes d'analyse statistique multivariable. Les techniques implantées appartiennent à l'analyse factorielle et à la classification automatique. Le package est programmé en APL.
L'exécution du package est interactive; l'utilisateur est guidé par un dialogue homme-machine écrit en français. A chaque question posée par l'ordinateur, l'utilisateur peut demander de plus amples informations en répondant par un point d'interrogation "?". Cette facilité s'appelle la fonction HELP.
L'utilisateur peut obtenir de nombreuses informations sur les possibilités offertes par le package en accédant à une documentation qui fait partie intégrante du package.
Le package est structuré en modules. Par l'intermédiaire du dialogue, l'utilisateur peut sélectionner le module qu'il veut exécuter. Une session commence toujours par l'appel à un module de définition de données et se poursuit par l'analyse du tableau de données par une ou plusieurs méthodes statistiques. Certaines méthodes statistiques créent des matrices-résultat qui peuvent être analysées à leur tour. Tel est le cas, par exemple, pour les techniques d'analyse factorielle qui créent les matrices des projections des observations ou des variables dans l'espace factoriel.
La connaissance des choix statistiques est contrôlée automatiquement par le package. Deux facilités, qui sont elles-mêmes des modules, aident l'utilisateur à sélectionner un module. Le MENU dynamique indique quelles sont les manipulations de données et les analyses statistiques qui peuvent être effectuées. Si l'utilisateur appelle MENU au début de l'exécution du package, ce dernier lui demandera d'appeler le module de définition des données. Le MENU répond donc à la question "Que puis-je faire ?".

Par la module TRACE, l'utilisateur peut obtenir un résumé de l'historique de la session. La facilité TRACE répond donc à la question : "Qu'ai-je déjà fait ?".
Finalement, il existe un mode de dialogue expert, appelé PSEUDO-BATCH, qui permet à l'utilisateur d'anticiper les questions futures du dialogue.

2. LA STRUCTURE DU PACKAGE

CLASFAC-APL est constitué de deux parties (cf. figure 1.1):
- la documentation du package qui se nomme DESCRIBE
- le package exécutable à proprement parler, appelé CLASFAC

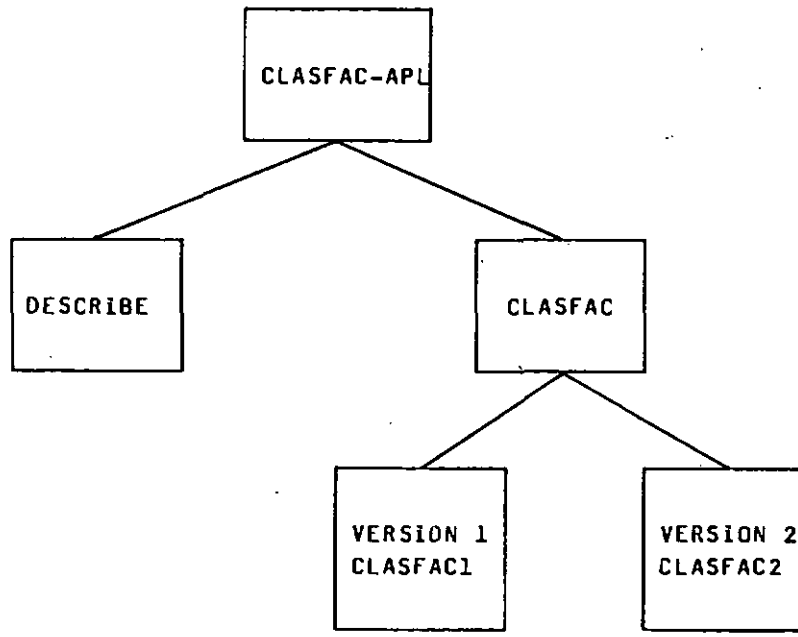


Figure 1.1

Le package exécutable CLASFAC est disponible en deux ver-

sions. Le contenu des deux versions est identique pour l'utilisateur. Elles se distinguent seulement par la grandeur des matrices de données qu'elles peuvent analyser pour une capacité de mémoire centrale donnée. La réalisation informatique est donc différente.
Dans la version 1, appelée CLASFAC1, le package est entièrement chargé en mémoire centrale. Cette version permet d'exécuter des analyses statistiques sur des tableaux de petite taille. Dans la version 2, appelée CLASFAC2, seulement un "résident permanent" et le module actif sont chargés en mémoire centrale. Il s'agit d'une version segmentée qui permet de traiter des matrices de plus grande taille.
Lors de l'exécution, CLASFAC2 est plus lente et plus onéreuse en temps de calcul que CLASFAC1.
Précisons cependant que les matrices qui peuvent être traitées par un package "interactif" sont nécessairement de plus petite taille que ce n'est le cas pour un package travaillant en mode "batch". Très souvent, ce désavantage est amplement compensé par la plus grande facilité d'utilisation du premier type de package.

3. LA STRUCTURE DE CLASFAC ET SON CONTENU

CLASFAC est formé de 18 modules indépendants de même niveau hiérarchique. Ils peuvent être regroupés en 3 types :

- les modules d'entrée-sortie
- les modules de méthodes statistiques
- les modules utilitaires (leur but est de faciliter l'utilisation du package)

Dans un premier temps, nous donnerons pour chaque module son contenu. La lettre qui suit la description d'un module est son abréviation. Lors du dialogue, cette lettre permet de sélectionner le module.
Dans l'exemple suivant, l'utilisateur appelle le module de définition de données :
QUEL MODULE ? /E/F/.../S/?/ : E
(Remarque : Le texte qui précède le signe ":" est écrit par l'ordinateur. La réponse de l'utilisateur se réduit au

caractère "E" à la fin de la ligne.)

Les modules d'entrée-sortie permettent :
- de définir la matrice des données /E/
- de définir des matrices auxiliaires /F/
- de définir des masques /M/
- d'imprimer une matrice /I/
- de représenter graphiquement un ou plusieurs nuages de points /D/
- d'initialiser des matrices /V/

Les modules de méthodes statistiques font appel aux techniques suivantes :
- l'analyse factorielle en composantes principales /A/
- l'analyse factorielle des correspondances /C/
- la projection de vecteurs dans un espace factoriel déjà défini /U/
- le partitionnement par la méthode KMEAN /P/
- le partitionnement par la méthode séquentielle adaptative (MSA) /B/
- la classification hiérarchique ascendante /H/
- la classification hiérarchique descendante polythétique par la méthode POLYDIV ou MIERKMEAN /X/
- la classification hiérarchique descendante monothétique par la méthode AID, appelée aussi "méthode de segmentation" /W/

Les modules utilitaires permettent :
- d'obtenir la trace automatique des opérations et des calculs déjà effectués (TRACE) /T/
- d'obtenir la liste des modules que l'utilisateur peut appeler compte tenu des opérations et des analyses déjà effectuées (MENU) /Q/
- de connaître la taille de la mémoire encore disponible /R/
- d'entrer une ou plusieurs lignes de commentaires /J/

La figure suivante illustre la structure de CLASFAC.

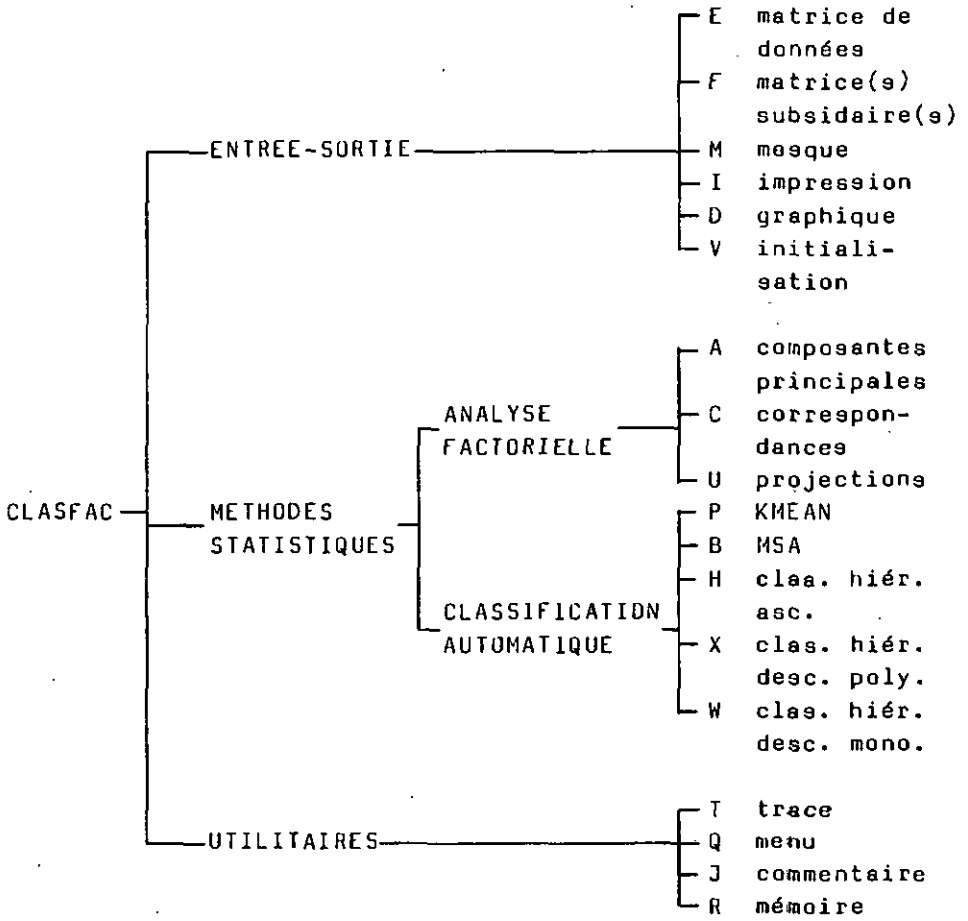


Figure 1.2

Rappel: Le point d'interrogation "?" ne fait pas appel à un module.
Il permet d'obtenir "à tout moment" des renseignements complémentaires.

4. LE PRINCIPE OU DIALOGUE

Lors du dialogue, les questions posées sont de deux types :

- les questions d'aiguillage (question de type 1) qui permettent d'appeler un module ou de se brancher à l'intérieur du module courant.

  Exemples:

  QUEL MODULE ? /E/F/.../S/?/ :
  QUEL BRANCHEMENT ? /DIS/NBC/PTI/CEI/CLA/ :

- les questions qui demandent des données ou des paramètres à définir par l'utilisateur (question de type 2).

  Exemples:

  - NOMBRE DE VARIABLES ? :
  - ENTREZ DANS L'ORDRE LES NOMS DES VARIABLES ? :

Pour les deux types de questions, en répondant par un point d'interrogation "?", on appelle la fonction HELP qui affichera des explications supplémentaires relatives à la question posée et reposera la question courante.

Exemples:

1) FORME DES DONNEES ? /BRU/CEN/CRE/ : ?
   FORMES POSSIBLES POUR LA MATRICE DES DONNEES :
   /BRU/ : DONNEES BRUTES
   /CEN/ : DONNEES CENTREES
   /CRE/ : DONNEES CENTREES ET REDUITES
   FORME DES DONNEES ? /BRU/CEN/CRE/ :

2) - NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? : ?
   LE NOMBRE DE CLASSES POUR LE PARTITIONNEMENT
   DOIT ETRE COMPRIS ENTRE 1 ET 13.
   - NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? :

4.1. Question de type 1

La syntaxe d'une question de type 1 est la suivante :

INTITULE DE LA QUESTION ? /OP1/OP2/.../OPn/ : REP

Le texte qui précède le signe ":" est écrit par l'ordinateur.
OP1, OP2, ..., OPn sont les différentes réponses possibles à la question.
REP est la réponse donnée par l'utilisateur.

4.1.1. Choix d'un module

Quand la question est relative au choix d'un module, les réponses possibles sont formées d'un seul caractère. Les lettres qui caractérisent les différents modules ont déjà été définies au paragraphe précédent.
La lettre S, pour STOP, permet d'arrêter l'exécution du package.
Le caractère "?" permet d'obtenir des informations supplémentaires.

Exemple:
QUEL MODULE ? /E/F/.../S/?/ : A

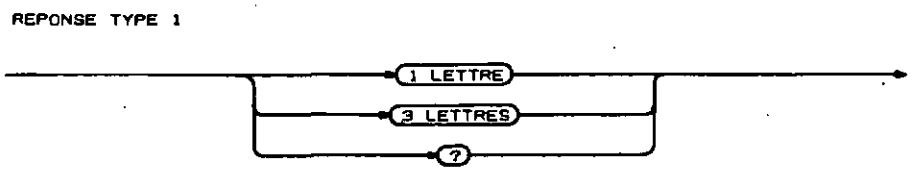
4.1.2. Branchement à l'intérieur d'un module

Si les réponses proposées, notées ci-dessus, OP1, OP2, ..., OPn, sont formées de 3 caractères, alors il s'agit d'une question qui permet de s'aiguiller à l'intérieur du module courant.
En choisissant une des réponses proposées, l'utilisateur se branche à l'intérieur du module courant. Par contre s'il répond par une seule lettre, il quitte prématurément le module courant et appelle un autre module.
Dans ce cas également, on peut obtenir des informations supplémentaires par "?".
La syntaxe des réponses est précisée à l'aide de diagrammes

syntaxiques. Un chemin à travers un tel diagramme définit une réponse syntaxiquement correcte. On distingue deux types de boîtes :

- les boîtes "arrondies", qui contiennent toujours des symboles terminaux, c'est-à-dire des réponses possibles de CLASFAC.
- les boîtes "rectangulaires", qui contiennent un nom qui est une référence à un autre diagramme. On retrouve ce nom en titre du diagramme correspondant.

La réponse à une question de type 1 est résumée par le diagramme suivant:



Exemples:

```
QUEL BRANCHEMENT ? /NBC/PTI/CEI/CLA/ :   PTI
QUEL BRANCHEMENT ? /NBC/PTI/CEI/CLA/ :    A
```

Remarque:
Une question de type 1 qui demande la réponse OUI ou NON sera toujours une question de branchement à l'intérieur d'un module. Par commodité, on accepte les réponses abrégées O et N :

```
TEXTE DE LA QUESTION ? /O/N/ :
```

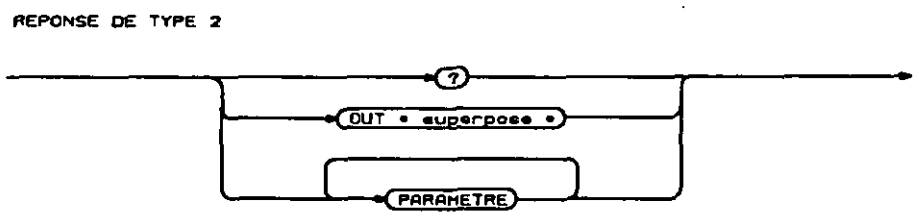
4.2. Question de type 2

Les questions de type 2 permettent de définir un ou plusieurs paramètres. Selon la question posée, ces paramètres sont des nombres ou des chaînes de caractères. Le texte de la question permet d'en décider.
La syntaxe d'une question de type 2 est :

- TEXTE DE LA QUESTION ? : PARAMETRE(S)

Le signe "-" signifie que la question posée se réfère à une

question de type 2. La question est posée en clair, l'utilisateur devra y répondre soit par une réponse appropriée, soit par "?". S'il y a plusieurs paramètres à entrer, il faut les séparer par des espaces.
Lorsque l'utilisateur répond à une question de type 2, il peut sortir du package, en introduisant la réponse "□" obtenue par la superposition des 3 caractères OUT. IL faut donc taper successivement O, <reculer>, U, <reculer>, T, <retour de chariot>.
Cette possibilité n'est à utiliser qu'en cas de "catastrophe". Pour relancer le package, l'utilisateur devra tout d'abord introduire le signe "→", puis effectuer un "retour de chariot" et, enfin, écrire GO.
Le diagramme syntaxique d'une question de type 2 est donc:



4.3. Les erreurs

En répondant aux différentes questions du dialogue, l'utilisateur peut commettre des erreurs. Dans ce paragraphe, nous donnerons pour chaque type d'erreur, leurs causes d'apparition et les mesures qui doivent être prises. Les types d'erreur sont classés par ordre croissant de gravité.

4.3.1. Fautes de frappe

Lorsque l'utilisateur a commis une faute de frappe avant qu'il n'ait appuyé sur la touche "retour de chariot", il doit reculer, à l'aide de la touche "←", jusqu'à la hauteur de la faute de frappe, appuyer sur la touche "ATTN" et récrire le texte à partir de la faute de frappe.

4.3.2. Erreur de transmission

Les messages "RESEND" ou "DATA LOST, REENTER" surviennent lors d'une difficulté de transmission entre le terminal et l'ordinateur. L'utilisateur devra réintroduire complètement la dernière ligne qu'il vient d'entrer, car elle est perdue.

4.3.3. Les réponses inappropriées

Pour les deux types de question, lorsque l'utilisateur apporte une réponse inappropriée, un message d'erreur apparaît et la question est reposée.

Exemple:

```
   QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV? : CRA
   ERREUR D'ENTREE
   QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/ :

   - NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? : ?
   LE NOMBRE DE CLASSES POUR LE PARTITIONNEMENT DOIT
   ETRE COMPRIS ENTRE 1 ET 13
   - NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? : 14
   ERREUR D'ENTREE
   - NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? :
```

Conformément à la définition d'une réponse de type 1 et 2, la détection d'une réponse inappropriée diffère selon le type de la question posée.

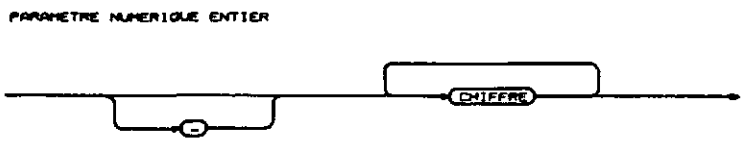
Réponse de type 1

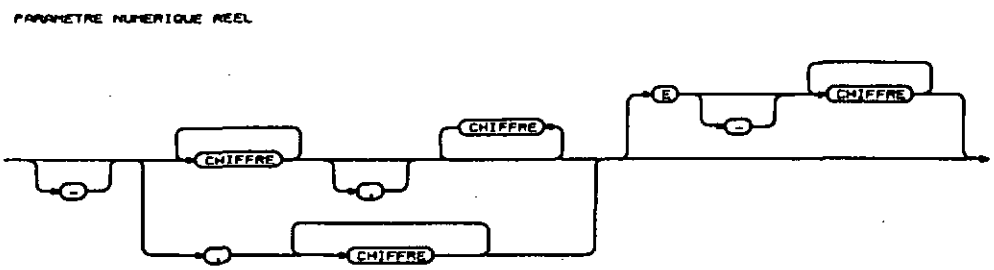
Pour les questions de type 1, une réponse inappropriée correspond à une réponse qui n'appartient pas à la liste des réponses possibles. Pour une question relative au choix d'un module, l'ensemble des réponses possibles est soit une lettre soit un "?".
Cet ensemble est complété par les réponses possibles de 3

lettres pour les questions d'aiguillage à l'intérieur d'un module.

Réponse de type 2

Pour les questions de type 2, la détection d'une réponse inappropriée dépend du type du paramètre (numérique ou alphanumérique) et de la question posée.
Pour un paramètre numérique, une réponse est inappropriée si sa syntaxe est incorrecte ou si le paramètre ne répond pas correctement à la question posée. Un paramètre numérique est formé uniquement des caractères suivants: [0, 1, 2, 3, 4, 5, 6, 7, 8, 9, ., ¯, E]. La syntaxe d'un nombre est résumée dans les diagrammes syntaxiques suivants:





Remarquons que le signe "¯" représente un nombre négatif.
Les réponses suivantes sont des réponses inappropriées:

```
  0.2A5          ABC      ..25          -.
caractère     caractère  erreur de     erreur de
illicite      illicite   syntaxe       syntaxe
```

En plus de la vérification de la syntaxe du paramètre numérique entré, un contrôle de sa validité est effectué. Les conditions de validité de la réponse varie d'une question à une autre.
Ces contrôles permettent de vérifier, selon les cas, la cohérence statistique du paramètre et la compatibilité du nouveau paramètre par rapport aux paramètres qui ont déjà été définis. Par exemple, à la suite de la question : "- NOMBRE DE CLASSES POUR LE PARTITIONNEMENT ? :", le programme contrôle, si la réponse apportée est un nombre entier compris entre 1 et le nombre d'objets à classer.
Pour les paramètres alphanumériques, une réponse est inappropriée si elle ne vérifie pas certaines conditions de validation propres à la question. Si, par exemple, les noms des observations sont TOTO, TITI, BOUL, BILL et que l'utilisateur désire désigner l'observation TOTO, mais écrit TOTI, on aura le dialogue suivant :

```
- NOM DE L'OBSERVATION ? : TOTI
ERREUR D'ENTREE
LE NOM TOTI N'APPARTIENT PAS A LA LISTE : TOTO TITI BOUL BILL
- NOM DE L'OBSERVATION ? :
```

Dans cet exemple, le programme contrôle si le nom entré, TOTI, appartient à la liste des noms des observations.

4.3.4. "Au secours"

L'utilisateur peut se trouver dans des situations particulières qui nécessitent l'interruption immédiate du package. Dans ce paragraphe, deux possibilités de sortie "en catastrophe" sont décrites.

L'ordinateur est en train d'exécuter (ATTN)

Entre deux questions, l'ordinateur exécute des calculs numériques et des manipulations de données. Si ces opérations utilisent peu de temps de calcul, les questions du dialogue se succèdent sans attente. Par contre, une opération telle que la diagonalisation d'une matrice nécessite beaucoup plus de temps. L'utilisateur doit donc attendre un moment avant que la question suivante n'apparaisse. Si l'utilisateur constate durant cette attente qu'il a commis une erreur grave, il peut sortir en "catastrophe" du package en appuyant deux fois sur la touche "ATTN". L'exécution est alors suspendue. Pour arrêter définitivement l'exécution du package, il faut ensuite appuyer sur la touche "→". Le package peut alors être relancé en son début par la commande GO.
Le terme "erreur grave" recouvre toutes les erreurs qui ne peuvent être détectées que par l'utilisateur, car aucune réponse n'était ni syntaxiquement ni logiquement fausse.
L'exemple suivant illustre une utilisation de la sortie en catastrophe : l'utilisateur a défini la matrice des données et a appelé un module d'analyse factorielle. Le programme est en train de diagonaliser la matrice des données, lorsque l'utilisateur constate que la matrice des données est erronée. Dès lors, il est inutile de continuer le traitement, et l'utilisateur provoque une interruption conformément à ce qui vient d'être décrit.

Interruption par "OUT"

Lorsque l'utilisateur introduit une réponse inappropriée à la suite d'une question de type 2, un message d'erreur apparaît et la question est reposée, et ceci jusqu'à l'obtention d'une réponse appropriée. Si l'utilisateur, même aidé par la fonction HELP, n'arrive pas à donner une réponse acceptable, il peut sortir du package en catastrophe en entrant la signe "▯ ", superposition de O, U, T.

5. LE MODE PSEUDO-BATCH

Il existe deux modes d'utilisation du dialogue : le mode interactif et le mode pseudo-batch.
Avec le mode interactif, l'utilisateur répond aux questions au moment où elles sont posées. Le mode pseudo-batch permet d'anticiper les réponses aux questions futures. L'impression des questions futures auxquelles l'utilisateur a déjà répondu est alors supprimée (suppression du dialogue) et le programme est exécuté jusqu'à l'épuisement de la liste des réponses anticipées ou jusqu'à la détection de la première erreur. Dans ce dernier cas, les réponses non encore traitées sont perdues.
L'entrée en mode pseudo-batch est permise seulement après une question de type 1.
La syntaxe d'une réponse en pseudo-batch est :

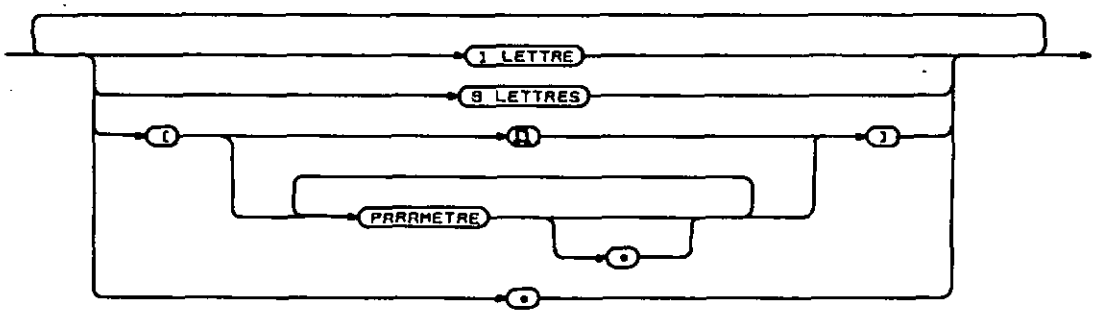
QUESTION ? /OP1/OP2/.../OPn/ : REP0 REP1 REP2 ... REPm

REP0 : est la réponse à la question posée

REP1, REP2, ..., REPm sont les réponses anticipées à des questions futures

Comme en mode interactif, la forme des réponses anticipées est une chaîne de 1 ou 3 caractères, si la réponse est relative à une question de type 1.
Si la réponse anticipée se rapporte à une question de type 2, la liste des paramètres doit être entourée de parenthèses carrées : [paramètre(s)].
A l'intérieur d'un module, l'utilisateur peut reprendre les anciens parmètres qui ont été définis précédemment à la suite de la même question de type 2, en entrant le signe "quad", noté "⍞", entre parenthèses carrées [⍞].
Une réponse en mode pseudo-batch peut s'étendre sur une ou plusieurs lignes en utilisant à la fin de chaque ligne le signe "*".
Le diagramme syntaxique suivant montre l'ensemble de toutes les réponses anticipées possibles :

PSEUDO-BATCH



Un exemple de dialogue

Cette exemple montre l'appel au module de définition de données /E/ et le début de la création de la matrice des données par entrée au clavier. Le même dialogue est d'abord reproduit en mode interactif, puis en pseudo-batch.

En mode interactif

```
QUEL MODULE ? /E/F/.../S/?/  :   E
- QUEL EST VOTRE NOM ? (5 LETTRES)  :   VOIDE
QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/  :   CRE
DE QUELLE FACON ? /ENT/DIS/WSP/  :   ?
POSSIBILITES:
/ENT/ : ENTRER LA MATRICE DES DONNEES AU CLAVIER
/DIS/ : LECTURE DE LA MATRICE DES DONNEES SUR DISQUE
/WSP/ : AFFECTATION DE VARIABLES DEFINIES DANS LA WORKSPACE
...
DE QUELLE FACON ? /ENT/WSP/DIS/ :    ENT
- NOMBRE D'OBSERVATIONS ?  :   5
- ENTREZ LES NOMS DES OBSERVATIONS ?  :   A1 A2 A3 A4 A5
- NOMBRE DE VARIABLES ? :
```

Remarque:
L'utilisateur a appelé la fonction HELP en répondant par un point d'interrogation. La possibilité WSP est décrite au paragraphe 8.

En mode pseudo-batch

```
QUEL MODULE ? /E/F/.../S/?/ :   E  [VIDE]  CRE ENT [5] *
[A1 A2 A3 A4 A5]
-NOMBRE DE VARIABLES ?  :
```

Remarque:
Les réponses anticipées étant toutes justes, le programme s'est arrêté après en avoir épuisé la liste.

6. LE CHEMINEMENT DANS LE PACKAGE

6.1. Introduction

Grâce à la structure modulaire du package CLASFAC, l'utilisateur peut passer sans autre d'un module à un autre.
Il incombe à l'utilisateur d'enchaîner les appels aux modules dans un ordre qui n'est pas dénué de sens. En particulier, une analyse statistique commence toujours par la définition de la matrice des données.
La figure ci-dessous montre les cheminements possibles à travers le package.

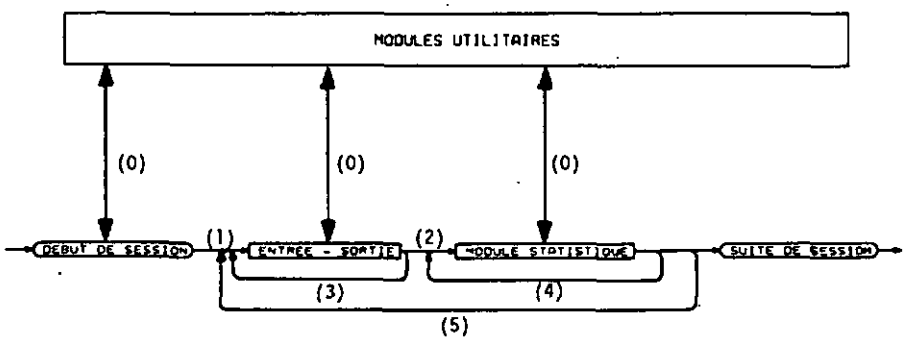


Figure 1.3

A tout moment l'utilisateur peut appeler un module utilitaire (arcs (0)).

6.2. Début d'une analyse statistique

Au début de la session, l'utilisateur passe obligatoirement dans le module /E/ pour définir la matrice des données (arc (1)). Une fois cette matrice définie, on peut passer directement à une analyse statistique (arc (2)) ou bien appeler un autre module d'entrée-sortie (arc (3)), par exemple pour imprimer la matrice des données.
La figure 1.4 explique en détail la façon dont on peut préparer une matrice de données par des appels multiples à des modules d'entrée-sortie. Il illustre comment on peut par une séquence d'appel aux modules /E/, /I/ et /M/ définir une matrice de données, l'imprimer, la corriger et en extraire une sous-matrice. Dans le package, cette dernière opération s'appelle "définition d'un masque".

6.3. Suite d'une analyse statistique

Après avoir fait une première analyse statistique sur la matrice des données qu'il vient de définir (arc (2)), l'utilisateur peut faire traiter la même matrice par une autre méthode statistique (arc (4)), ou bien revenir aux modules d'entrée-sortie (arc (5)), pour définir une sous-matrice de la matrice courante ou bien une autre matrice.
Certains modules statistiques créent des résultats qui, à leur tour, ont la forme d'une matrice de données. Une telle matrice peut être analysée statistiquement (arc (4)) ou traitée par un module d'entrée-sortie (arc (5)).

Exemple d'un dessin du plan factoriel

L'utilisateur désire exécuter une analyse factorielle sur la matrice des données et représenter graphiquement les résultats de l'analyse factorielle. Il doit séquentiellement:

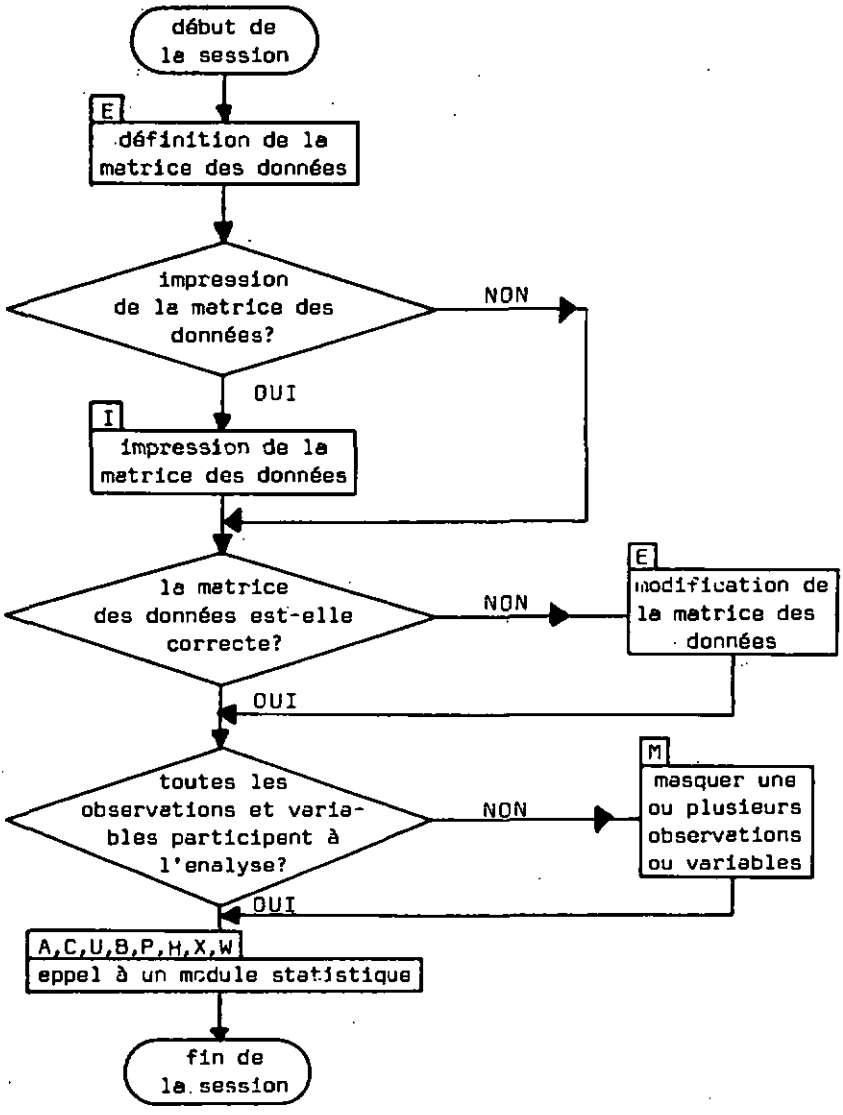


Figure 1.4 Le préperetion d'une metrice de donnéea

(1) définir la matrice des données /E/
(2) effectuer l'analyse factorielle /A/ ou /C/
(3) représenter graphiquement les réaultats de l'analyae factorielle /O/.

Lea méthodea d'analyse factorielle calculent lea projectiona dea obaervationa et des variables sur les facteura. Par la auite, l'utilisateur peut donc repréaenter graphiquement les résultata en appelant le module de deaain /O/. Il arrive souvent que le premier dessin ne soit paa satisfaisant (beaucoup de points au centre, quelquea pointa excentriques). Pour obtenir un "sgrandiaaement" du centre on procède de la manière auivante:

(i) on élimine dans la matrice dea projectiona les points excentriques en appelant le module masque /M/,
(ii) on redessine la "sous-matrice" ainal définie.

En réaumé on aura donc appelé en aéquance les modulea "deasin O", "masque M" et de nouveau "deaain O".

6.4. Le cheminement à l'intérieur d'un module

Lea propoa tenua dans ce paragraphe ont expliqué le cheminement du package par l'appel succesif de modulea.
Afin d'être complet, dea informationa concernant le cheminement à l'intérieur des modules d'entrée-aortie et de méthodea atatiatiques vont suivre. Toua cea moduies ont la même structure.

6.4.1. Choix de la matrice à traiter

Lsa modulaa d'entrée-sortie et de méthodea atatiatiquea commancent par aélectionner la matrica dea données qui va être traitée. L'utiliaateur doit répondre à une question dont l'intitulé est :

"QUELLE DONNEE ? /OP1/OP2/.../OPn/ :"
OPi, OP2, ..., OPn eat la liata de toutea les matrices qui

peuvent être traitées par le module. Cette liste dépend du module où la question est posée. Pour les modules qui traitent exclusivement la matrice des données, à savoir /E/, /A/, /C/ et /W/, la question est omise.
L'exemple suivant est relatif à un module de partitionnement:

```
QUELLE DONNEE ? /OBS/VAR/POB/PVA/ : ?

LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE
UTILISEES POUR LE PARTITIONNEMENT :
/OBS/ : OBSERVATIONS DE LA MATRICE DES DONNEES
/VAR/ : VARIABLES DE LA MATRICE DES DONNEES
/POB/ : PROJECTIONS DES OBSERVATIONS
/PVA/ : PROJECTIONS DES VARIABLES

SEULES LES MATRICES CI-DESSOUS SONT DEJA DEFINIES:
- LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
- LES VARIABLES DE LA MATRICE DES DONNEES /VAR/

QUELLE DONNEE ? /OBS/VAR/POB/PVA/ :
```

L'abréviation retenue pour une matrice particulière reste la même pour tous les modules. Donc, OBS représente toujours les observations de la matrice des données.

L'existence des matrices à traiter est contrôlée automatiquement par le programme. Si l'utilisateur sélectionne une matrice qui n'est pas encore définie, un message d'erreur apparaît et la question est reposée. Si nous gardons l'exemple précédent, l'utilisateur recevra un message d'erreur s'il choisit POB.

```
QUELLE DONNEE ? /OBS/VAR/POB/PVA/ : POB
LE JEU DE DONNEES CHOISI POUR LE PARTITIONNEMENT N'EST PAS
DEFINI
QUELLE DONNEE ? /OBS/VAR/POB/PVA/ :
```

6.4.2. Aiguillage à l'intérieur d'un module

Dans tous les modules, il existe plusieurs possibilités d'aiguillage.
Dans un module d'entrée-sortie, un aiguillage revient à se référer à l'une des fonctions constituant le module, soit par exemple pour le module /E/: créer, modifier, étendre, ... la matrice des données.
Dans les modules statistiques, les aiguillages permettent de sélectionner les options pour l'exécution de l'algorithme.

6.4.3. Possibilités de se rebrancher en divers points du module

A la fin de chaque module, l'utilisateur peut se rebrancher en différents points du module. Cette possibilité sera offerte par la question :
"QUEL BRANCHEMENT ? /OP1/OP2/.../OPn/".
Après le branchement, le module est exécuté séquentiellement. De cette façon l'utilisateur peut exécuter à nouveau un module statistique avec un nouveau jeu d'options ou utiliser une autre fonction d'un module d'entrée-sortie.

6.5. Les modules utilitaires

Les modules utilitaires facilitent l'emploi du package. Ils visent trois objectifs:

- faciliter les appels aux différents modules /T/ et /Q/
- personnaliser les sorties du programme en imprimant des commentaires /J/
- gérer la place disponible en mémoire centrale /R/

Le dernier point sera développé dans la deuxième partie au chapitre 5, car il nécessite de plus amples connaissances sur le processus de définition de la matrice des données.
L'analyse des données est un processus exploratoire itératif. Au début de l'analyse, l'utilisateur a une idée sur les méthodes qu'il veut employer, mais des

réaultata intermédiaires inattendua peuvent modifier aon plan. Deux modules, appelés IRACE et MENU dynamique, facilitent le cheminement dans le package.
TRACE résume la sesslon de travail, en donnant la liste des fonctions et méthodes déjà utiliséea. Le MENU dynamique donne la liste des modules que l'utiliaateur peut appeler, compte tenu des opérations déjà effectuéea.

7. LA STRUCTURE DE LA DOCUMENTATION

La documentation du package commence par une table des matières. L'utiliaateur peut alors sélectionner les descriptions qui l'intéressent. Cea descriptions peuvent être classées en trois catégories, selon leur contenu :

a) recueil de toutes lea informationa néceaaairea pour exécuter le package. Cette deacription répond aux queationa auivantes :
Comment exécuter le package ?
Comment répondre aux différentea queationa ?

b) informationa sur le principe des méthodea atatiatiquea implantées dans le package et la bibliographie y relative

c) le contenu et lea poaaibilitéa offertea par chaque module

Lora d'un premier eaaai du package, aeulea lea deacriptiona de type a) doivent être examinéea. La fonction MELP, et lea modulea utilitaires tels que le MENU guideront l'utiliaateur dana aaa choix. Lea deacriptions b) aont utiles à l'utiliaateur débutant en atatiatique.
Le but dea deacriptiona c) eat d'aider l'utiliaateur à employer efficacement le package.
La figura 1.5 réaume le contenu da chaque deacription. La table dea matièrea ae trouve dana la fonction DESCRIBE. L'appel à une deacription ae fait en entrant au clavier le

nom de la description. Pour obtenir, par exemple, la table des matières l'utilisateur entrera le mot-clef DESCRIBE.

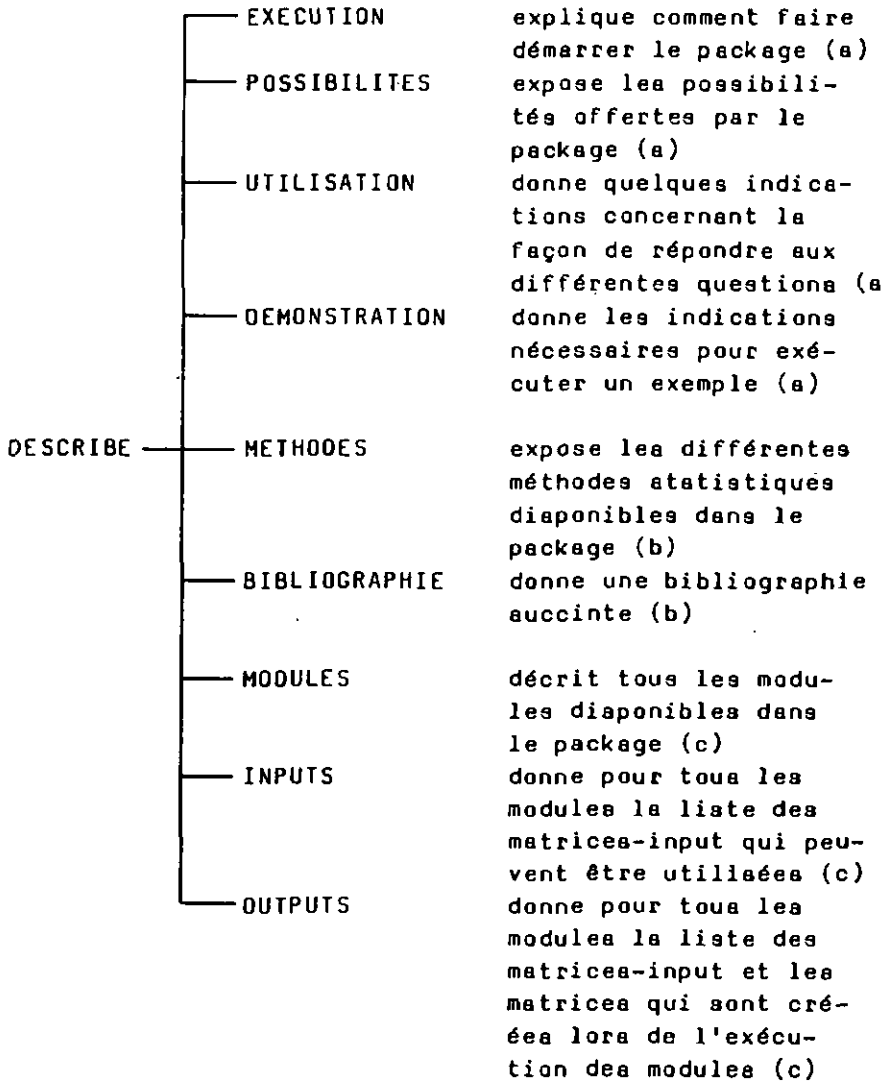


Figure 1.5

8. L'EXECUTION DU PACKAGE

Parallèlement aux explications qui permettent de faire démarrer le package et d'accéder aux fichiers de données sauvés sur disque, nous situerons l'environnement informatique dans lequel CLASFAC est réalisé.

L'exécution du package se fait en deux étapes :
- le chargement du package,
- l'appel de la fonction qui fait démarrer le package.

Le tableau suivant résume les instructions que l'utilisateur doit effectuer pour accéder aux différentes versions ou parties de CLASFAC.

	INSTRUCTIONS POUR LA VERSION 1 DE CLASFAC	INSTRUCTIONS POUR LA VERSION 2 DE CLASFAC	INSTRUCTIONS POUR LA DOCUMENTATION
CHARGEMENT	)LOAD CLASFAC1 "RET"	)LOAD CLASFAC2 "RET"	)LOAD DESCRIBE "RET"
EXECUTION	GO "RET"	GO "RET"	DESCRIBE "RET"

"RET" signifie appuyer sur la touche "retour de chariot". [Uniquement pour des raisons typographiques, "RET" est écrit sur une nouvelle ligne.]

Avant d'expliquer le sens des deux commandes qui permettent l'exécution du package, quelques notions concernant l'envi-

ronnement du langage APL vont être introduites.
Lorsqu'un utilisateur travaille en APL sur un ordinateur, il dispose d'une zone de mémoire, dont la taille est fixée par le système, et qu'on appelle espace de travail actif. Parfois nous utiliserons le terme anglais workspace active. Cette zone contient durant la session de travail les fonctions du package et les données.
L'utilisateur a en plus accès à un espace disque qui contient l'ensemble des workspaces sauvées sur disque, appelée bibliothèque privée de l'utilisateur, ainsi que des fichiers de données.

Illustration:
Souvent dans la littérature, la workspace active est assimilée à un bloc-note et la workspace sauvée sur disque à un livre d'une bibliothèque. Lorsque le bloc-note contient de l'information que l'utilisateur désire conserver, celui-ci peut, par une commande système, copier le bloc-note et le transformer en livre. Inversément, lorsque l'utilisateur veut extraire l'information contenue dans le livre, une commande système permet de copier le contenu du livre dans le bloc-note.

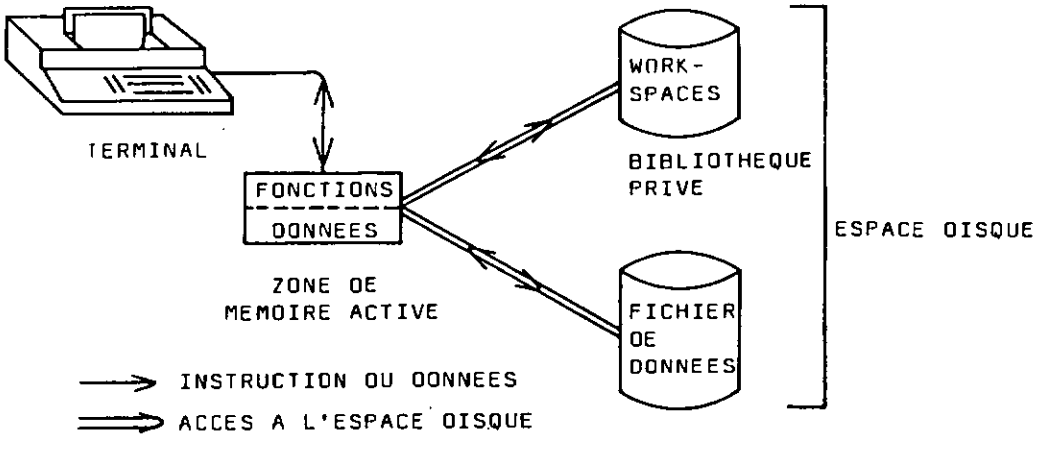


Figure 1.6

L'utilisateur du package CLASFAC-APL dispose d'une bibliothèque privée formée du contenu de trois workspaces sauvées sur disque : CLASFAC1, CLASFAC2 et DESCRIBE. Lors de la connexion au système, la zone de travail active est vide c'est-à-dire qu'aucune fonction ou variable n'est définie. L'utilisateur peut charger le contenu d'une workspace sauvée sur disque par la commande :

)LOAD NOM_DE_LA_WORKSPACE_SAUVEE_SUR_DISQUE

Après le chargement, l'utilisateur possède dans la workspace active le texte de toutes les fonctions de la workspace. En APL, le terme de fonction recouvre les termes de programme et de sous-programme des autres langages de programmation. Si l'utilisateur charge, par exemple, la zone de travail CLASFAC1, par la commande)LOAD CLASFAC1, toutes les fonctions nécessaires à l'exécution du package sont transférées dans la workspace active. L'utilisateur peut alors exécuter CLASFAC en appelant la fonction de guidage GO. Pour accéder à la partie documentation du package, l'utilisateur doit d'abord charger la workspace DESCRIBE dans la zone de travail active et exécuter ensuite la fonction DESCRIBE, qui est donc une fonction particulière de la workspace de même nom.

9. JEUX DE DONNEES POUR ESSAI ET DEMONSTRATION

Pour essayer le package CLASFAC, l'utilisateur a à sa disposition deux jeux de données mémorisés sur disque et fournis avec le package :

1) le premier jeu concerne des données agricoles de la Suisse occidentale,

2) le deuxième concerne des données relatives à la qualité de marques de montres.

L'utilisateur y accédera en répondant aux questions du dialogue de la façon suivante :

```
 MODULE CHOISI ? /E/F/.../Q/S/?/  :  E
- QUEL EST VOTRE NOM ? (5 CARACTERES AU MAXIMUM) : AGRIC
                                    si on choisit  1)
                                                MONTR
                                    si on choisit  2)
QUELLE OPERATION /CRE/ETE/MOD/DEL/SAV/?/  :   CRE
DE QUELLE FACON ? /ENT/DIS/WSP/  :   DIS
```

En utilisant le mode pseudo-batch, le dialogue se résume à :

```
MODULE CHOISI ? /E/F/.../Q/S/?/  :  REP
  Si l'utilisateur choisit  1)  REP sera :
     E  [AGRIC]  CRE DIS
  Si l'utilisateur choisit  2)  REP sera :
     E  [MONTR]  CRE DIS
```

Après ces opérations, l'utilisateur peut appeler n'importe quel module pour travailler sur la matrice choisie.

10. LA CREATION DE FICHIERS DE DONNEES APPARTENANT A L'UTILISATEUR

En cours d'analyse, la workspace active contient des données et des fonctions.
Lorsque l'utilisateur cesse de travailler, ou lorsqu'il charge une autre workspace, le contenu de l'espace de travail actif est détruit.
CLASFAC offre la possibilité à l'utilisateur de sauver sur disque la matrice des données et/ou des matrices subsidiaires. Cette opération s'effectue par un appel au module /E/.
Le fichier ainsi créé portera le nom de l'utilisateur. Par contre il n'est pas possible de sauver sur disque les résultats calculés par un module statistique. Ceux-ci doivent donc être recréés lors d'une session ultérieure.
Cette option a été prise afin de limiter le nombre de fichiers à gérer.

Le figure auivante illuatre CLASFAC et aon environnement informatique :

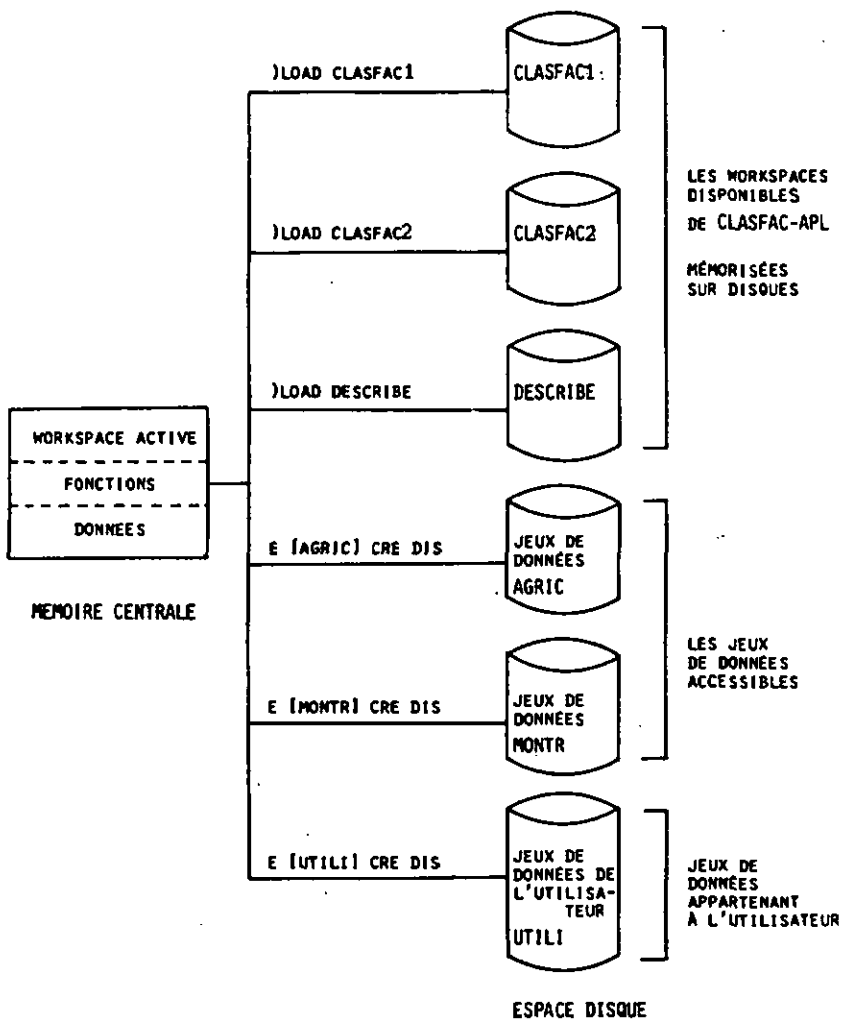


Figure 1.7

11. L'ARRET DE L'EXECUTION DE CLASFAC

Fin normale

L'exécution du package est arrêtée par la réponse d'une lettre "S" (STOP) à une question de type 1.

Arrêt en catastrophe

Si l'utilisateur répond à une question de type 2, par la réponse "0", superposition des caractères OUT, l'exécution du package sera arrêtée définitivement, si l'on appuie sur la touche "→".
Si le package est en train d'exécuter des opérations, l'utilisateur doit appuyer deux fois sur la touche "ATTN", puis sur la touche "→".
Ces deux possibilités d'arrêt en catastrophe ont déjà été décrites au paragraphe 4.

12. LE LANGAGE DE PROGRAMMATION APL

A titre d'information (puisque l'utilisateur n'a pas besoin de connaître APL pour tirer profit du package), nous conclurons ce chapitre par un survol des caractéristiques du langage de programmation APL.

Historiquement, APL est né des travaux de Kenneth E. Iverson sur la formalisation des algorithmes. Le but d'Iverson était le développement d'un langage formel concis et consistant de description d'algorithmes. Les concepts fondamentaux en sont exposés dans son ouvrage intitulé : A Programming Language publié en 1962. Pendant plus de six ans, APL a été utilisé comme langage formel. La première réalisation expérimentale a été implantée par IBM sur un système 360.

APL est un langage interactif utilisant une notation non ambiguë. Dans ce langage, aucun mot réservé n'existe. Tous les opérateurs sont définis par des symboles d'un seul caractère +, -, /, !, ?, *, \,....

APL est conçu pour travailler sur des tableaux. De nombreux opérateurs très puissants facilitent leur traitement. APL permet l'allocation dynamique en mémoire. Ainsi tout nombre, nommé scalaire en APL, peut être immédiatement converti en un tableau de n'importe quelle dimension.
A l'heure actuelle de nombreux constructeurs possèdent APL sur leur machine.

LES METHODES STATISTIQUES ET LEUR IMPLANTATION

CHAPITRE 1 : LA MATRICE DES DONNEES ET LES MODULES D'ENTREE-SORTIE

1. LA MATRICE DES DONNEES

1.1. Introduction

Le point de départ de toutes les méthodes statistiques utilisées dans le package est une matrice de données. Cette dernière contient l'information qui doit être analysée.
La matrice des données est un tableau de chiffres dont les lignes forment les observations et les colonnes les variables. Les termes objet, entité, sujet et individu sont souvent utilisés dans la littérature comme synonymes d' "observation", alors que les notions d'attribut, caractéristique et propriété coïncident avec le mot "variable".
Dans les modules de classification, nous utiliserons parfois le terme "objet" pour désigner les entités à classer, qui peuvent être les observations ou les variables.
Comme le champ d'application de l'analyse des données embrasse tous les domaines de l'investigation scientifique, les observations peuvent représenter les entités les plus diverses, soit par exemple, en médecine, des maladies ou des patients, en marketing, des clients ou des produits, en gestion financière, des entreprises, en économie, des districts, en biologie, des animaux .... La matrice des données est complètement définie, si pour chaque observation, on connaît les valeurs des variables qui la caractérisent.
Avant de traiter la matrice des données par le package, les observations et les variables doivent être choisies avec soin. Seul le chercheur, en fonction de sa connaissance a priori du phénomène étudié, peut juger de l'opportunité de la participation de certaines observations ou variables.

Par exemple, dans une étude de la criminalité dans différentes villes des Etats-Unis, les observations retenues pour l'analyse sont les plus grandes villes des Etats-Unis. Chaque observation est décrite par 5 variables : les taux de meurtre, de viol, de cambriolage, d'effraction et de vol de voitures. Intuitivement, sans être un criminologue, la sélection des variables illustre d'une façon appropriée le but poursuivi par l'étude. Cet exemple est tiré de l'ouvrage d'Hartigan /16/.
La distinction entre variables et observations est souvent arbitraire, mais en principe, l'observation diffère de la variable par son caractère répétitif.
Chaque variable peut prendre différentes valeurs, nommées ses états. Ces états peuvent être des valeurs réelles ou entières. Dans l'exemple de l'étude de la criminalité, toutes les variables prennent des valeurs réelles, puisqu'elles représentent des taux. Parfois, des variables, par la nature même de l'étude, ne représentent pas a priori des états numériques. Citons comme exemple la variable état civil qu'on rencontre fréquemment dans les questionnaires ; ses états possibles sont: célibataire, marié, veuf et divorcé. Lors de la saisie du questionnaire, les états de cette variable sont souvent codés par les nombres 1,2,3,4.
Dans un premier temps, nous allons introduire les notations relatives à la matrice des données. Ces dernières seront utilisées lors de la présentation des différentes méthodes statistiques.
Dans un deuxième temps, nous présenterons les différents types de variables et les techniques de recodification de variables qui permettent d'obtenir une matrice de données homogène.

1.2. Notations relatives à la matrice des données

Nous noterons $I=\{1,2,\ldots,m\}$ l'ensemble des observations, $J=\{1,2,\ldots,n\}$ l'ensemble des variables et X la matrice des données de dimension (m,n) indexée par les ensembles I et J.

J /	1	2	3	...	j	...	n
1	x_1^1	x_1^2	x_1^3	...	x_1^j	...	x_1^n
2	x_2^1	x_2^2	x_2^3	...	x_2^j	...	x_2^n
...	...	...	...	...	...	...	...
i	x_i^1	x_i^2	x_i^3	...	x_i^j	...	x_i^n
...	...	...	...	...	...	...	...
m	x_m^1	x_m^2	x_m^3	...	x_m^j	...	x_m^n

L'élément x_i^j désigne la valeur prise par la i-ième observation pour la j-ième variable.
Nous noterons X_i, le i-ième vecteur-ligne de la matrice X:

$$X_i = (x_i^1 \quad x_i^2 \quad x_i^3 \quad \ldots \quad x_i^j \quad \ldots \quad x_i^n)$$

Les éléments de ce vecteur déterminent les valeurs prises par la i-ième observation pour les différentes variables.
Le vecteur X_i peut être considéré comme une représentation de la i-ième observation dans l'espace vectoriel à n dimensions, R^n.
Le sous-ensemble de R^n suivant:

$$N(1) = \{ X_i \mid 1 \leq i \leq m \}$$

est appelé le nuage des points-observation.

34 LA MATRICE DES DONNEES ET LES MODULES D'ENTREE-SORTIE

Nous noterons par X^j le j-lème vecteur-colonne de la matrice X:

$$X^j = \begin{bmatrix} x_1^j \\ x_2^j \\ \cdot \\ \cdot \\ \cdot \\ x_i^j \\ \cdot \\ \cdot \\ \cdot \\ x_m^j \end{bmatrix}$$

Les éléments de ce vecteur sont les valeurs prises par les différentes observations pour la j-ième variable.
Le vecteur X^j peut être considéré comme une représentation de la variable dans l'espace vectoriel à m dimensions, R^m.
La sous-ensemble de R^m suivant :

$$N(J) = \{ X^j \mid 1 \leq j \leq n \}$$

est appelé le nuage des points-variable.
Remarquons que la matrice des données peut être représentée soit par les m observations soit par les n variables :

$$X = [X^1 \quad X^2 \quad \ldots \quad X^j \quad \ldots \quad X^n] = \begin{bmatrix} X_1 \\ X_2 \\ \ldots \\ X_m \end{bmatrix}$$

Les modules de classification permettent de classer les observations ou les variables de la matrice des données. Pour conclure, remarquons que si l'on transpose la matrice des données, les notions d'observations et de variables sont échangées. Cette constatation nous permet d'utiliser les mêmes algorithmes de classification pour traiter les observations et les variables.

1.3. Les types de variables

Les propriétés des états possibles d'une variable déterminent son type. Ce paragraphe donne les caractéristiques et des exemples pour tous les types de variables.
Il existe deux catégories de variables : les variables quantitatives et les variables qualitatives.

1.3.1. Les variables quantitatives

Par la nature même du phénomène étudié, les états d'une variable quantitative sont mesurés par des nombres réels ou entiers. Si les états résultent d'un comptage, il s'agit de la sous-classe importante des fréquences.
Les opérations statistiques suivantes peuvent être effectuées : le calcul de la moyenne, de la variance, du moment d'ordre h et du coefficient de corrélation.
Toutes les variables obtenues par des dispositifs de mesure sont des variables quantitatives. Dans un tel contexte, la variable mesurée peut être un poids, une densité, ou une longueur d'onde. En économie, toutes les mesures monétaires sont des variables quantitatives : PNB, chiffre d'affaires, taux de change.

1.3.2. Les variables qualitatives

Les états d'une variable qualitative ne sont qu'une représentation possible adoptée par le chercheur. Si les états sont structurés par une relation d'ordre total, la variable est ordinale, dans le cas contraire, elle est nominale.

Les variables ordinales

Les opérations statistiques suivantes peuvent être effectuées : le calcul de la médiane, des quartiles, des déciles, du coefficient de rang de Spearman. Les variables ordinales englobent toutes les variables qui réalument des préférences (entre différents produits) et des notes (lors d'une épreuve ou lors de la mesure de la qualité d'un produit).

Les variables nominales

Les opérstions suivantes peuvent être exécutées : le calcul du mode, des comptages, la créstion de tsbleau de fréquences.
Une variable nominale qui ne possède que deux états possibles est une variable dichotomique. On utilise également les synonymes de variable booléenne, logique ou binaire. Les exemples les plus courants sont le sexe : homme ou femme, codé 0 et 1, et les questions qui n'admettent que les réponses oui et non.
Une variable nominale est dite multinominale si elle possède plus de deux étets possibles. Les variables multinominales recouvrent par exemple, les codifications de catégories socio-professionnelles ou socio-culturelles, de questions auxquelles on répond par oui, non, sans réponse ou indifférent.

1.4. La recodification d'une variable

Quoique CLASFAC ne possède aucun module pour transformer une variable d'un type en une variable d'un autre type, quelques indications concernant les méthodes possibles vont être données.
Dans la présentation des différents types de variables, aucune notion d'échelle de mesure n'a été utilisée, car le but poursuivi était de les introduire d'une façon intuitive. Mathématiquement, la distinction entre deux types de variables est basée sur les propriétés de ses états par rapport à l'échelle sur laquelle ils ont été mesurés.

Toua lea principes de recodification aont baaés sur le paaaage d'un type d'échelle de meaure à un autre type d'échelle.
La figure 2.1 indique, pour chaque type de variable, l'échelle de mesure qui eat utiliaée avec ses caractéristiques et les tranaformationa qui laissent invariant la variable. Elle donne en plua les changements d'échelles possiblea qui modifient le type de la variable.
Le changement d'un type de variable en un autre type peut se faire soit par le passage à une échelle de meaure plua riche (transformation d'une variable nominale en une variable ordinale), soit par le passage à une échelle de mesure plus pauvre (transformation d'une variable quantitative en variable qualitative).
La première transformation implique un enrichiaaement dea variablea par l'adjonction d'hypothèses supplémentaires baaéea aur les propriétés dea variables et la deuxième un appauvriaaement de l'information contenue dana la variable.
La fin de ce paragraphe eat conaacrée aux différentea transformations possibles de variablea.

1) Les transformationa par appauvriaaement de l'information

a) Iranaformation d'une variable quantitative en une variable ordinale

Opération:
Il existe deux poaaibilitéa :
1) on définit dea classea contiguëa sur l'échelle d'intervalle. Lea objeta d'une même claaae ont le même rang et la relation d'ordre entra les claaaes eat conservée.
2) on affectue dea opérationa atatiatiquea (cf. Anderberg /9/). L'idée eat baaée aur la répartition égale de l'effectif dans les diveraes claaaes.

Types de variables	Echelles et relations mathématiques	Transformation qui laissent les variables invariantes
variable quantitative	Echelle de rapport Equivalence. Ordre. Rapport entre 2 valeurs. Rapport entre 2 intervalles.	Transformation de similarité ou homothétie $x' = cx$ $c>0$
variable ordinale	Echelle d'intervalle Equivalence entre les objets ayant même valeur. Ordre large. Rapport entre deux intervalles.	Transformation linéaire ou affine $x' = ax + b$ $a>0$
variable ordinale	Echelle ordinale Equivalence entre les objets ayant le même ordre (strict, large, total, partiel) préordre quasi-ordre	Transformation monotone $x' = f(x)$ avec $f(x)$ fonction monotone croissante
variable nominale	Echelle nominale Equivalence entre les membres d'une même classe	Substitution $x' = f(x)$ avec $f(x)$ fonction discrète et bijective

Légende:

→ : transformations par appauvrissement des variables

⟹ : transformations par enrichissement des variables

Figure 2.1

Remarque: Ce tableau est inspiré de Chandon et Pinson /12/.

Perte d'information:
La distinction entre les objets d'une même classe est perdue et les classes ne sont structurées que par une relation d'ordre.

Remarque:
L'opération est identique pour transformer une variable quantitative mesurée sur une échelle d'intervalle ou de rapport. La seule distinction est la signification du zéro.

b) Transformation d'une variable ordinale en une variable nominale

Opération:
Cette transformation est purement "philosophique" : il suffit de considérer les rangs comme des états différents sans aucune relation d'ordre.

Perte d'information:
Toute relation d'ordre est perdue.

c) Transformation d'une variable multinominale en plusieurs variables binaires

Opération:
Chaque état de la variable multinominale définit une variable binaire.

Exemple: Etat civil

états	code originel	code binaire
célibataire	1	1000
marié	2	0100
veuf	3	0010
divorcé	4	0001

Perte d'information: aucune

2) Les transformations par enrichissement de l'information

a) Transformation d'une variable binaire en une variable multinominale

Une variable binaire peut toujours être condidérée comme un cas particulier d'une variable multinominale.

b) Transformation d'une variable nominale en une variable ordinale

Opération:
Cette transformation est purement "philosophique". Il suffit de considérer les états de la variable nominale comme ordonnés.

Hypothèse:
Il faut supposer l'existence d'une relation d'ordre.

c) Transformation d'une variable ordinale en une variable quantitative

Opération:
Cette transformation est difficile, il faut se référer à Anderberg /9/.

Hypothèse:
Il faut affecter une valeur à chaque classe ordonnée qui préserve l'ordre et qui de plus donne un sens à la différence entre classes.

1.5. Recodification de la matrice des données

Une matrice de données peut être hétérogène, c'est-à-dire, contenir à la fois des variables quantitatives et qualitatives. Selon la méthode statistique utilisée, la matrice des

données doit être homogénéisée, ce qui signifie, transformée en une matrice dont toutes les variables sont de même type. Les différentes méthodes statistiques de CLASFAC utilisent uniquement des matrices de données homogènes, qualitatives ou quantitatives.
Les passages possibles d'un type de matrice de données à un autre dépendent toujours de l'état initial de la matrice des données et de la méthode statistique qui va être utilisée. Si, par exemple, la matrice des données contient à la fois des variables quantitatives, qualitatives nominales, ordinales et dichotomiques et si l'utilisateur veut exécuter une classification descendante par la méthode de segmentation, il doit transformer les variables quantitatives en n'importe quel type de variable qualitative. Pour d'autres méthodes statistiques, l'utilisateur devra utiliser d'autres transformations.

2. LES MODULES D'ENTREE-SORTIE

2.1. La notion de jeux de données ou de matrice de données dans CLASFAC

Dans le package CLASFAC, un jeu de données est formé:

- de la matrice des données proprement dite (tableau de chiffres)
- des noms des observations
- des noms des variables

Si ces trois objets sont compatibles, on dit que le jeu de données est défini. La règle de compatibilité est la suivante : la matrice des données doit contenir autant de lignes qu'il y a de noms d'observations et autant de colonnes qu'il y a de noms de variables. La figure ci-dessous illustre cette règle.

42 LA MATRICE DES DONNEES ET LES MODULES D'ENTREE-SORTIE

	1VAR	2VAR	...	JVAR	...	NVAR
1OBS	x_1^1	x_1^2	...	x_1^j	...	x_1^n
2OBS	x_2^1	x_2^2	...	x_2^j	...	x_2^n
...	...	...	...	...	...	...
MOBS	x_m^1	x_m^2	...	x_m^n	...	x_m^n

Parfois, par abus de langage, nous utiliserons l'expression "matrice des données" pour désigner le jeu de données.
Comme on peut exécuter les modules avec différents jeux de données, chaque jeu porte un nom. La matrice des données s'appelle DON. Dans les modules de classification, l'utilisateur peut classer les observations ou les variables de la matrice des données. Les observations de la matrice des données sont notées OBS et les variables VAR. Il est important de souligner que si DON est définie, OBS et VAR le sont également.
Seulement trois jeux de données sont définis par l'utilisateur :

- la matrice des données, abrégée DON
- la matrice des observations supplémentaires, abrégée OSU
- la matrice des variables supplémentaires, abrégée VSU

Ces deux dernières sont appelées des <u>matrices subsidiaires</u>. Leur utilisation est expliquée au chapitre 3.
Les matrices subsidiaires doivent vérifier les conditions suivantes : la matrice des observations supplémentaires doit porter sur les mêmes variables que la matrice des données et la matrice des variables supplémentaires sur les mêmes observations que la matrice des données. Ces conditions peuvent être résumées par la figure suivante:

	1VAR	2VAR	...	NVAR
1OBS	x_1^1	x_1^2	...	x_1^n
...	...	...	...	...
MOBS	x_m^1	x_m^2	...	x_m^n

1VSU	2VSU
v_1^1	v_1^2
...	...
v_m^1	v_m^2

1OSU	o_1^1	o_1^2	...	o_1^n
2OSU	o_2^1	o_2^2	...	o_2^n

Toutes les autres matrices qui peuvent être utilisées dans les modulea d'entrée-sortie ou dans les modulea statistiques aont des matricea-résultats, c'est-à-dire dea matrices qui ont été créées par des modules statistiques.

2.2. Typologie dea modulea d'entrée-sortie

Les modulea réunis aous le nom de module d'entrée-sortie peuvent être regroupés en 4 catégories:

1) lea modules de définition de la matrice dea donnéea et dea matrices subsidiairea /E/F/
2) lea modules de sortie /I/D/
3) le module de aélection d'une aous-matrice, appelé module de maaque /M/
4) le module d'initialiaation de jeux de donnéea /V/

Il aat important de diatinguer parmi les modulea ceux qui effectuent des opérations physiques de ceux qui se basent aur dea opérations logiques.
Une opération physique change le jeu de données et l'ancien

jeu est donc "perdu". Tel est le cas pour les catégories 1) et 4).
Par contre une opération logique modifie le contenu d'un jeu de données qui va être traité sans qu'on perde pour autant le jeu de données initial. La catégorie 3) permet de sélectionner une sous-matrice sans détruire la matrice initiale.
Remarquons qu'une opération logique nécessite parfois des initialisations. Celles-ci se font automatiquement par le package. Ainsi, si l'utilisateur masque la matrice des données, toutes les matrices-résultats obtenues précédemment ne sont plus compatibles et doivent donc être détruites.

2.3. Les fonctions des modules et quelques exemples

Nous allons maintenant décrire les possibilités offertes par chaque module. La plupart du temps nous montrerons par un exemple une ou plusieurs possibilités du module décrit. Parfois un renvoi indique le paragraphe, où un exemple est donné conjointement avec un autre module.
Les numéros entre parenthèses renvoient à des explications complémentaires données à la fin de la description générale du module.

2.3.1. /E/ Module de définition, de correction et de mémorisation de la matrice des données

Ce module permet :
- de définir la matrice des données, les noms des observations et les noms des variables (1)
  - par entrée au clavier
  - par affectation de variables définies dans la workspace (zone de travail) (2)
  - par lecture sur disque (3)
- d'étendre la matrice des données (augmenter le nombre d'observations ou de variables)
- de modifier :
  - les valeurs d'une observation ou d'une variable
  - la valeur d'un élément de la matrice des données

- le nom d'une observation ou d'une variable
- de supprimer une observation ou une variable
- de sauver la matrice des données sur disque (création d'un fichier sur disque) avec la possibilité de sauver également la matrice des observations supplémentaires et des variables supplémentaires si elles sont définies (4).
- d'effacer sur disque un fichier de données créé par l'utilisateur. Cette opération ne détruit pas pour autant la matrice des données définie dans l'espace de travail actif.

Compléments:

(1) : Lors de la définition des noms des observations et des variables, les deux premiers caractères doivent à eux seuls permettre de distinguer les observations ou les variables.

Exemple:
Il ne faut pas définir les noms :
65FEMMES 65HOMMES 6510 6511
mais,
FEMME65 HOMME65 1065 1165

(2) : Cette possibilité est offerte à l'utilisateur "APListe" qui possède une workspace privée sauvée sur disque. Celle-ci doit contenir 3 variables qui définissent la matrice des données, les noms des observations et des variables. Ces variables peuvent alors être transmises au programme par l'intermédiaire de cette possibilité. Les 3 variables APL doivent respecter la condition de compatibilité donnée au paragraphe 2.1. Si ces trois variables s'appellent dans l'ordre X1, X2, X3, l'utilisateur doit exécuter les commandes suivantes :

```
)LOAD CLASFAC1 ou CLASFAC2
)COPY BIBLIOPRIVEE X1 X2 X3
GO
```

(3) : Si l'utilisateur ne possède pas de matrice de données sauvée sur disque, il peut accéder à un des deux jeux de données intégrés au package.
(voir: première partie paragraphe 9 : Jeux de données pour essai et démonstration).

(4) : L'environnement de CLASFAC est décrit au paragraphe 10 de la première partie.

Exemples:

L'exemple illustre la création, l'extension et la modification d'une petite matrice de données. Ne connaissant pas encore le package, l'utilisateur appelle à la suite de chaque question la fonction HELP dans le but d'obtenir des explications complémentaires.
A la suite de chaque opération sur la matrice des données (module /E/) on imprime cette dernière en utilisant le mode pseudo-batch. Les autres options du modules d'impression /I/ seront illustrées en 2.3.3.

GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ ?

L'UTILISATEUR DOIT DANS PRESQUE TOUS LES CAS REPONDRE A CETTE QUESTION PAR O(UI).
LA REPONSE N(ON) AURA DE GRAVES CONSEQUENCES SI L'UTILISATEUR L'UTILISE A MAUVAIS ESCIENT. ELLE COMPROMETTRA TOTALEMENT L'EXECUTION DU PACKAGE.

/OUI/ : LE PACKAGE CLASFAC DOIT TOUJOURS COMMENCER PAR L'INITIALISATION DE TOUTES LES VARIABLES QUI SONT UTILISEES DANS LE PACKAGE.
ATTENTION:
SI L'UTILISATEUR OMET CETTE INITIALISATION, DES MESSAGES D'ERREUR '' VALUE ERROR '' APPARAITRONT LORS DE L'EXECUTION.

/NON/ : CETTE OPTION DOIT ETRE PRISE AVEC PRECAUTION ET APRES REFLEXION. ELLE EST IMPLANTEE DANS LE SEUL BUT DE PERMETTRE A UN UTILISATEUR QUI A TERMINE NORMALEMENT L'EXECUTION DU PACKAGE ET QUI N'EST PAS SORTI DE LA WORKSPACE CLASFAC1 OU CLASFAC2 (SELON LA VERSION QUI EST EXECUTEE), D'EXECUTER A NOUVEAU LE PACKAGE SANS INITIALISER LES VARIABLES QUI ONT DEJA ETE DEFINIES.

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ O
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ ?

LISTE DES MODULES DISPONIBLES
****

MODULES D'ENTREE-SORTIE
****
/E/ MODULE DE DEFINITION, DE CORRECTION ET DE MEMORISATION DE LA MATRICE DES DONNEES
/F/ MODULE DE DEFINITION DE MATRICES SUBSIDIAIRES
/M/ MODULE DE MASQUE
/D/ MODULE DE DESSIN
/I/ MODULE D'IMPRESSION
/V/ MODULE D'INITIALISATION

MODULES D'ANALYSE
****
/A/ MODULE D'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES
/C/ MODULE D'ANALYSE FACTORIELLE DES CORRESPONDANCES
/U/ MODULE DE PROJECTION DE POINTS
/P/ MODULE DE PARTITIONNEMENT PAR LA METHODE DE KMEAN
/B/ MODULE DE PARTITIONNEMENT PAR LA METHODE SEQUENTIELLE ADAPTATIVE (MSA)
/H/ MODULE DE CLASSIFICATION HIERARCHIQUE ASCENDANTE
/X/ MODULE DE CLASSIFICATION HIERARCHIQUE DESCENDANTE POLYTHETIQUE
/W/ MODULE DE CLASSIFICATION HIERARCHIQUE DESCENDANTE MONOTHETIQUE PAR LA METHODE AID, APPELEE AUSSI LA SEGMENTATION.
LES VARIABLES EXPLICATIVES DOIVENT ETRE QUALITATIVES ET LA VARIABLE EXPLIQUEE BOOLEENNE.

MODULES UTILITAIRES
****
/T/ MODULE DE TRACE AUTOMATIQUE
/Q/ MENU
/R/ PLACE DISPONIBLE
/J/ COMMENTAIRES
/S/ MODULE D'ARRET D'EXECUTION DU PACKAGE (STOP)
/?/ RENSEIGNEMENTS SUPPLEMENTAIRES

CONSEILS POUR LE DEBUTANT
****
POUR CHOISIR UN MODULE REPONDEZ Q.
POUR CHOISIR UNE REPONSE A L'INTERIEUR D'UN MODULE REPONDEZ D'ABORD PAR ?.

```
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ Q

*****************
*MODULE MENU /Q/*
*****************

L'UTILISATEUR DOIT APPELER LE MODULE /E/ POUR DEFINIR LA MATRICE
DES DONNEES

MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ E

*******************************************************************************
*MODULE DE DEFINITION, CORRECTION, MEMORISATION DE LA MATRICE DES DONNEES /E/*
*******************************************************************************

→ QUEL EST VOTRE NOM ? (5 LETTRES) ←⊞→ ?

LE NOM QUE VOUS ENTREREZ, SERA UTILISE POUR DEFINIR LES NOMS DES FICHIERS QUE
VOUS CREEREZ QUE VOUS LIREZ OU QUE VOUS EFFACEREZ.
LE PREMIER CARACTERE DU NOM DOIT ETRE UNE LETTRE, SEULS LES 5 PREMIERS
CARACTERES SERONT PRIS EN CONSIDERATION.

→ QUEL EST VOTRE NOM ? (5 LETTRES) ←⊞→ VOIDE
QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ ?

POSSIBILITES:
  /CRE/ : CREER UNE NOUVELLE MATRICE DES DONNEES
  /ETE/ : ETENDRE LA MATRICE DES DONNEES
  /MOD/ : MODIFIER DES ELEMENTS DE LA MATRICE DES DONNEES
  /DEL/ : DETRUIRE UNE MATRICE DE DONNEES SAUVEE SUR DISQUE
  /SAV/ : SAUVER UN JEU DE DONNEES SUR DISQUE

QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ CRE
DE QUELLE FACON ? /ENT/DIS/WSP/?/ ←⊞→ ?

POSSIBILITES:
  /ENT/ : ENTRER LA MATRICE DES DONNEES AU CLAVIER
  /DIS/ : LECTURE DE LA MATRICE DES DONNEES SUR DISQUE
  /WSP/ : AFFECTATION DE VARIABLES DEFINIES DANS LA WORKSPACE
          IL FAUT DEFINIR 3 VARIABLES DANS LA WORKSPACE :
            - LA MATRICE DES DONNEES (MATRICE NUMERIQUE DE DIMENSION
              (NOBS,NVAR))
            - LES NOMS DES OBSERVATIONS DE LA MATRICE DES DONNEES
              (LIGNES), (MATRICE ALPHANUMERIQUE DE DIMENSION (NOBS,X))
            - LES NOMS DES VARIABLES DE LA MATRICE DES DONNEES
              (COLONNES), (MATRICE ALPHANUMERIQUE DE DIMENSION (NVAR,X))
              AVEC:
               NOBS : LE NOMBRE D'OBSERVATIONS
               NVAR : LE NOMBRE DE VARIABLES
                  X : UN NOMBRE QUELCONQUE
```

```
DE QUELLE FACON ? /ENT/DIS/WSP/?/ ←┤├→ ENT
→ NOMBRE D'OBSERVATIONS ? ←┤├→ 4
→ ENTREZ LES NOMS DES OBSERVATIONS DANS L'ORDRE. UTILISEZ PLUSIEURS LIGNES
  SI NECESSAIRE.
A  B  C
ENCORE 1
D
→ NOMBRE DE VARIABLES ? ←┤├→ 3
→ ENTREZ LES NOMS DES VARIABLES DANS L'ORDRE. UTILISEZ PLUSIEURS LIGNES
  SI NECESSAIRE.
S  T  U
→ ENTREZ LES VALEURS DE LA MATRICE PAR LIGNE.
  UTILISEZ PLUSIEURS LIGNES SI CELA EST NECESSAIRE.
  EN CAS D'ERREUR VOUS POUVEZ REINTRODUIRE UNE LIGNE DE LA MATRICE.
  POUR INDIQUER UNE CORRECTION ENTREZ LE SIGNE Δ.
LIGNE=1 NOM: A
12  2.34  ¯12.3
LIGNE=2 NOM: B
23.1  1.21€¯1  ¯2.2
ERREUR D'ENTREE REINTRODUISEZ VOTRE DERNIERE ENTREE
LIGNE=2 NOM: B
23.1  1.21E¯1  2.2
LIGNE=3 NOM: C
12  2.35  ¯12.3
LIGNE=4 NOM: D
Δ
ENTREZ LE NUMERO DE LA LIGNE A CORRIGER ? ←┤├→ ?

ENTREZ LE NUMERO DE LA LIGNE QUE VOUS VOULEZ MODIFIER.
APRES LA CORRECTION, L'UTILISATEUR SERA AUTOMATIQUEMENT GUIDE A LA LIGNE
OU IL S'EST ARRETE.

ENTREZ LE NUMERO DE LA LIGNE A CORRIGER ? ←┤├→ 2
LIGNE=2 NOM: B
23.1  1.21E¯1  ¯2.2
LIGNE=4 NOM: D
3  2  33
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←┤├→ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE ENTREE:
  /CRE/ : CREER UNE NOUVELLE MATRICE DES DONNEES
  /ETE/ : ETENDRE LA MATRICE DES DONNEES
  /MOD/ : MODIFIER DES ELEMENTS DE LA MATRICE DES DONNEES
  /DEL/ : DETRUIRE UNE MATRICE DE DONNEES SAUVEE SUR DISQUE
  /SAV/ : SAUVER UN JEU DE DONNEES SUR DISQUE

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←┤├→ N
```

5D LA MATRICE DES DONNEES ET LES MODULES D'ENTREE-SORTIE

```
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ I DON O [2]

**************************
*MODULE D'IMPRESSION /I/*
**************************

IMPRESSION DE LA MATRICE DES DONNEES

              S          T          U
A         12.00       2.34     ¯12.30
B         23.10       0.12      ¯2.20
C         12.00       2.35     ¯12.30
D          3.00       2.00      33.00

NOMS DES VARIABLES /O/N/ ←⊞→ E [□]

*****************************************************************************
*MODULE DE DEFINITION, CORRECTION, MEMORISATION DE LA MATRICE DES DONNEES /E/*
*****************************************************************************

QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ ?

POSSIBILITES:
  /CRE/ : CREER UNE NOUVELLE MATRICE DES DONNEES
  /ETE/ : ETENDRE LA MATRICE DES DONNEES
  /MOD/ : MODIFIER DES ELEMENTS DE LA MATRICE DES DONNEES
  /DEL/ : DETRUIRE UNE MATRICE DE DONNEES SAUVEE SUR DISQUE
  /SAV/ : SAUVER UN JEU DE DONNEES SUR DISQUE

QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ ETE
ETENDRE LA MATRICE DES DONNEES EN AUGMENTANT LE NOMBRE D'OBSERVATIONS ? /O/N/
←⊞→ O
→ NOMBRE DE NOUVELLES OBSERVATIONS ? ←⊞→ 2
→ ENTREZ LES NOMS DES NOUVELLES OBSERVATIONS. UTILISEZ PLUSIEURS LIGNES
  SI NECESSAIRE.
E  F
→ ENTREZ LES VALEURS DES OBSERVATIONS SUR UNE OU PLUSIEURS LIGNES.
  EN CAS D'ERREUR VOUS POUVEZ REINTRODUIRE UNE LIGNE DE LA MATRICE.
  POUR INDIQUER UNE CORRECTION ENTREZ LE SIGNE Δ .
LIGNE=1 NOM: E
2X  3  3
ERREUR D'ENTREE REINTRODUISEZ VOTRE DERNIERE ENTREE
LIGNE=1 NOM: E
2  3  3
LIGNE=2 NOM: F
2  33  44
```

```
ETENDRE LA MATRICE DES DONNEES EN AUGMENTANT LE NOMBRE DE VARIABLES ? /O/N/
←⊞→ O
→ NOMBRE DE NOUVELLES VARIABLES ? ←⊞→ 3
→ ENTREZ LES NOMS DES NOUVELLES VARIABLES. UTILISEZ PLUSIEURS LIGNES
  SI NECESSAIRE.
V  W  X
→ ENTREZ LES VALEURS DES VARIABLES SUR UNE OU PLUSIEURS LIGNES.
  EN CAS D'ERREUR VOUS POUVEZ REINTRODUIRE UNE LIGNE DE LA MATRICE.
  POUR INDIQUER UNE CORRECTION ENTREZ LE SIGNE Δ .
LIGNE=1 NOM: V
1  2  2  33  3
ENCORE 1
3
LIGNE=2 NOM: W
3  2.2  4.12  5  3.2  2
LIGNE=3 NOM: X
2  3.2  1.  2.  2.23  3
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ I DON O [□]

*************************
*MODULE D'IMPRESSION /I/*
*************************


IMPRESSION DE LA MATRICE DES DONNEES

          S        T        U        V        W        X
A     12.00     2.34   -12.30     1.00     3.00     2.00
B     23.10     0.12    -2.20     2.00     2.20     3.20
C     12.00     2.35   -12.30     2.00     4.12     1.00
D      3.00     2.00    33.00    33.00     5.00     2.00
E      2.00     3.00     3.00     3.00     3.20     2.23
P      2.00    33.00    44.00     3.00     2.00     3.00


NOMS DES VARIABLES /O/N/ ←⊞→ E [□]

*****************************************************************************
*MODULE DE DEFINITION, CORRECTION, MEMORISATION DE LA MATRICE DES DONNEES /E/*
*****************************************************************************

QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ ?

POSSIBILITES:
  /CRE/ : CREER UNE NOUVELLE MATRICE DES DONNEES
  /ETE/ : ETENDRE LA MATRICE DES DONNEES
  /MOD/ : MODIFIER DES ELEMENTS DE LA MATRICE DES DONNEES
  /DEL/ : DETRUIRE UNE MATRICE DE DONNEES SAUVEE SUR DISQUE
  /SAV/ : SAUVER UN JEU DE DONNEES SUR DISQUE
```

```
QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←┘ MOD
QUELLE OPERATION ? /DOB/DVA/COB/CVA/COV/CNO/CNV/?/ ←┘ ?

POSSIBILITES DE MODIFICATION:
  /DOB/ : EFFACER UNE OU PLUSIEURS OBSERVATIONS
  /DVA/ : EFFACER UNE OU PLUSIEURS VARIABLES
  /COB/ : CHANGER TOUTES LES VALEURS D'UNE OBSERVATION
  /CVA/ : CHANGER TOUTES LES VALEURS D'UNE VARIABLE
  /COV/ : CHANGER LA VALEUR D'UN ELEMENT DEFINI PAR LE NOM D'UNE OBSERVATION
          ET D'UNE VARIABLE
  /CNO/ : CHANGER LE NOM D'UNE OBSERVATION
  /CNV/ : CHANGER LE NOM D'UNE VARIABLE

QUELLE OPERATION ? /DOB/DVA/COB/CVA/COV/CNO/CNV/?/ ←┘ COV
→ NOM DE L'OBSERVATION ET DE LA VARIABLE DE L'ELEMENT A MODIFIER ? ←┘ A  U
→ NOUVELLE VALEUR DE L'ELEMENT  ? ←┘ 12.30
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←┘ MOD
QUELLE OPERATION ? /DOB/DVA/COB/CVA/COV/CNO/CNV/?/ ←┘ DOB
→ NOM DES OBSERVATIONS A SUPPRIMER ? ←┘ F
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←┘ MOD
QUELLE OPERATION ? /DOB/DVA/COB/CVA/COV/CNO/CNV/?/ ←┘ CVA
→ NOM DE LA VARIABLE A MODIFIER ? ←┘ X
→ ENTREZ LES VALEURS DE LA VARIABLE A MODIFIER ? ←┘
  EN CAS D'ERREUR VOUS POUVEZ REINTRODUIRE UNE LIGNE DE LA MATRICE.
  POUR INDIQUER UNE CORRECTION ENTREZ LE SIGNE Δ .
LIGNE=1 NOM: X
20 32  10  20  22.3
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←┘ I DON O [□]

*************************
*MODULE D'IMPRESSION /I/*
*************************


IMPRESSION DE LA MATRICE DES DONNEES

```

	S	T	U	V	W	X
A	12.00	2.34	12.30	1.00	3.00	20.00
B	23.10	0.12	2.20	2.00	2.20	32.00
C	12.00	2.35	12.30	2.00	4.12	10.00
D	3.00	2.00	33.00	33.00	5.00	20.00
E	2.00	3.00	3.00	3.00	3.20	22.30

```
NOMS DES VARIABLES /O/N/ ←┘ S
FIN DE L'EXECUTION DU PACKAGE ? /O/N/ (N'OUBLIEZ PAS DE SAUVER VOTRE JEU DE
DONNEES SUR DISQUE) ←┘ ?
```

/O/ : FIN DE L'EXECUTION DU PACKAGE
/N/ : CHOIX D'UN AUTRE MODULE

REMARQUE:
SI L'UTILISATEUR VEUT SAUVER SES DONNEES SUR DISQUE, IL DEVRA APPELER LE MODULE /E/ ET IL REPONDRA A CETTE QUESTION PAR:
... ? ←□□→ E [SONNOM] SAV

FIN DE L'EXECUTION DU PACKAGE ? /O/N/ (N'OUBLIEZ PAS DE SAUVER VOTRE JEU DE DONNEES SUR DISQUE) ←□□→ O
FIN DE L'EXECUTION DU PACKAGE

2.3.2. /F/ Module de définition de matrices subsidiaires

Il permet :
- de définir :
  - la matrice des observations supplémentaires /OSU/
  - la matrice des variables supplémentaires /VSU/
    - par entrée au clavier
    - par affectation de variables définies dans la workspace active (5)
    - par lecture sur disque
    - à partir de la matrice des données
  - d'étendre /OSU/ (d'ajouter des observations supplémentaires)
  - d'étendre /VSU/ (d'ajouter des variables supplémentaires)
  - de supprimer une ou plusieurs observations ou variables supplémentaires.

Complément:

(5) : Cette possibilité est offerte à l'utilisateur "APListe". Dans la workspace sauvée sur disque 3 variables doivent être définies : la matrice subsidiaire (tableau de chiffres), les noms des lignes de la matrice subsidiaire et les poids associés à chaque objet.

Un exemple est donné au chapitre 3, paragraphe 7.

2.3.3. /I/ Module d'impreasion

Il permet d'imprimer un jeu de données.

Exemple:

Avant de montrer les possibilités offertes par le module /I/, un jeu de données mis à la disposition de l'utilisateur pour des essais et des démonstrations est lu sur disque. Il s'agit des données relatives à la qualité de marques de montres (cf. partie 1 paragraphe 9).
Les différentes méthodes statistiques implantées dans CLASFAC seront exécutées sur ce jeu de données à l'exception de la méthode de segmentation, module /W/ pour laquelle un exemple particulier sera introduit.
Le jeu de données relatif aux marques de montres est un tableau de contingence. Les observations représentent les différentes qualités attribuées à une montre:

- bonne qualité (QUALITE)
- technique avancée (TECHNIQUE)
- très précis (PRECISION)
- esthétique (ESTHETIQUE)
- fiable (FIABILITE)
- vaut son prix (VALEUR)
- solide (SOLIDITE)
- élégant (ELEGANT)
- fabrique des montres électroniques (ELECTRON)

et les variables représentent onze marques de montres désignées par les noms : 1MARQUE, 2MARQUE, ..., 11MARQUE.
La valeur d'un élément de la matrice des données correspond au nombre de personnes qui ont attribué la qualité X (par exemple FIABILITE) à la marque Y (par exemple 6MARQUE).

```
      GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←↵→ O
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←↵→ E

*****************************************************************************
*MODULE DE DEFINITION, CORRECTION, MEMORISATION DE LA MATRICE DES DONNEES /E/*
*****************************************************************************

→ QUEL EST VOTRE NOM ? (5 LETTRES) ←↵→ MONTR
QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←↵→ CRE
DE QUELLE FACON ? /ENT/DIS/WSP/?/ ←↵→ DIS
LECTURE CORRECTE
QUEL BRANCHEMENT ? /CRE/ETE/MOD/DEL/SAV/?/ ←↵→ I

************************
*MODULE D'IMPRESSION /I/*
************************

QUELLE DONNEE ? /DON/POB/PVA/SOB/SVA/PCO/PCV/IVO/IVV/IFO/IFV/OSU/VSU/?/ ←↵→ ?

LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE IMPRIMEES:
  /DON/ : MATRICE DES DONNEES
  /POB/ : PROJECTIONS DES OBSERVATIONS
  /PVA/ : PROJECTIONS DES VARIABLES
  /OSU/ : OBSERVATIONS SUPPLEMENTAIRES
  /VSU/ : VARIABLES SUPPLEMENTAIRES
  /IVO/ : CENTRES D'INERTIE DES OBSERVATIONS
  /IVV/ : CENTRES D'INERTIE DES VARIABLES
  /IFO/ : CENTRES D'INERTIE DES COMPOSANTES DES OBSERVATIONS
  /IFV/ : CENTRES D'INERTIE DES COMPOSANTES DES VARIABLES
  /PCO/ : PROJECTIONS DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /PCV/ : PROJECTIONS DES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES
  /SOB/ : PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
  /SVA/ : PROJECTIONS DES VARIABLES SUPPLEMENTAIRES

SEULES LES MATRICES CI-DESSOUS SONT DEFINIES:
 - LA MATRICE DES DONNEES /DON/


QUELLE DONNEE ? /DON/POB/PVA/SOB/SVA/PCO/PCV/IVO/IVV/IFO/IFV/OSU/VSU/?/ ←↵→ DON
MATRICE DES DONNEES ? /O/N/ ←↵→ ?

/O/ : IMPRESSION DE LA MATRICE CHOISIE ENCADREE DES NOMS DE SES VARIABLES
      ET DE SES OBSERVATIONS
/N/ : POURSUITE SEQUENTIELLE DU MODULE

MATRICE DES DONNEES ? /O/N/ ←↵→ O
→ AVEC QUELLE PRECISION DECIMALE ? /0/1/2/3/4/5/ ←↵→ ?
```

LA PRECISION, POUR L'ECRITURE DE LA MATRICE, PEUT ETRE 0,1,2,3,4 *OU* 5 *POSITIONS APRES LA VIRGULE.*

→ *AVEC QUELLE PRECISION DECIMALE ? /0/1/2/3/4/5/* ←⊞→ 0

IMPRESSION DE LA MATRICE DES DONNEES

	1*MARQUE*	2*MARQUE*	3*MARQUE*	4*MARQUE*	5*MARQUE*	6*MARQUE*
QUAL	345	828	46	132	1348	492
TECH	100	387	29	172	1065	571
PREC	173	586	11	110	1156	410
ESTH	449	463	28	98	861	311
FIAB	194	568	20	87	1133	358
VALE	181	396	20	88	779	318
SOLI	183	611	17	72	1105	290
ELEG	481	475	17	71	879	209
ELEC	34	201	33	230	892	701

	7*MARQUE*	8*MARQUE*	9*MARQUE*	10*MARQUE*	11*MARQUE*
QUAL	365	167	111	580	23
TECH	200	156	42	356	68
PREC	211	107	69	402	50
ESTH	218	110	72	598	53
FIAB	219	90	58	421	116
VALE	155	157	56	353	320
SOLI	197	115	79	419	103
ELEG	194	76	56	542	62
ELEC	166	201	43	177	322

NOMS DES VARIABLES /O/N/ ←⊞→ *?*

/O/ : IMPRESSION DES NOMS DES VARIABLES DE LA MATRICE CHOISIE
/N/ : POURSUITE SEQUENTIELLE DU MODULE

NOMS DES VARIABLES /O/N/ ←⊞→ *O*

IMPRESSION DES NOMS DES VARIABLES DE LA MATRICE DES DONNEES

1*MARQUE*
2*MARQUE*
3*MARQUE*
4*MARQUE*
5*MARQUE*
6*MARQUE*
7*MARQUE*
8*MARQUE*
9*MARQUE*
10*MARQUE*
11*MARQUE*

NOMS DES OBSERVATIONS /O/N/ ←⊞→ ?

/O/ : IMPRESSION DES NOMS DES OBSERVATIONS DE LA MATRICE CHOISIE
/N/ : POURSUITE SEQUENTIELLE DU MODULE

NOMS DES OBSERVATIONS /O/N/ ←⊞→ O

IMPRESSION DES NOMS DES OBSERVATIONS DE LA MATRICE DES DONNEES

QUALITE
TECHNIQUE
PRECISION
ESTHETIQUE
FIABILITE
VALEUR
SOLIDITE
ELEGANT
ELECTRON

QUEL BRANCHEMENT ? /OUT/OVA/OOB/?/ ←⊞→ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE D'IMPRESSION :
/OUT/ : IMPRESSION DE LA MATRICE (MEME JEU DE DONNEES)
/OVA/ : IMPRESSION DES NOMS DES VARIABLES (MEME JEU DE DONNEES)
/OOB/ : IMPRESSION DES NOMS DES OBSERVATIONS (MEME JEU DE DONNEES)

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←⊞→ N

QUEL BRANCHEMENT ? /OUT/OVA/OOB/?/ ←⊞→ S O
FIN DE L'EXECUTION DU PACKAGE

2.3.4. /M/ Module de masque

Il permet :

- de masquer (supprimer) des lignes ou des colonnes d'une matrice afin de définir une sous-matrice. Cette opération correspond à une opération logique.

Exemple:

Par l'appel au module /M/, nous allons éliminer plusieurs observations et variables dans la matrice des marques de montres. Après avoir imprimé la sous-matrice ainsi définie, nous allons la "démasquer".

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ N
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ I DON N O O

*************************
*MODULE D'IMPRESSION /I/*
*************************

IMPRESSION DES NOMS DES VARIABLES DE LA MATRICE DES DONNEES

1MARQUE
2MARQUE
3MARQUE
4MARQUE
5MARQUE
6MARQUE
7MARQUE
8MARQUE
9MARQUE
10MARQUE
11MARQUE


IMPRESSION DES NOMS DES OBSERVATIONS DE LA MATRICE DES DONNEES

QUALITE
TECHNIQUE
PRECISION
ESTHETIQUE
FIABILITE
VALEUR
SOLIDITE
ELEGANT
ELECTRON
```

```
QUEL BRANCHEMENT ? /OUT/OVA/OOB/?/ ←⊞→ M

**********************
*MODULE DE MASQUE /M/*
**********************

QUELLE DONNEE ? /DON/IVO/IVV/IFO/IFV/POB/PVA/SOB/SVA/?/ ←⊞→ ?

LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE MASQUEES:
  /DON/ : LA MATRICE DES DONNEES
  /IVO/ : LES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /IVV/ : LES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES
  /IFO/ : LES CENTRES D'INERTIE DES PROJECTIONS DES OBSERVATIONS
  /IFV/ : LES CENTRES D'INERTIE DES PROJECTIONS DES VARIABLES
  /POB/ : LES PROJECTIONS DES OBSERVATIONS
  /PVA/ : LES PROJECTIONS DES VARIABLES
  /SOB/ : LES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
  /SVA/ : LES PROJECTIONS DES VARIABLES SUPPLEMENTAIRES

CONSEQUENCE DE L'APPLICATION DU MASQUE DANS TOUS LES MODULES

 DON : A, C, P, B, H, X, W, D, I
 POB :       P, B, H, X,    D, I
 PVA :       P, B, H, X,    D, I
 SOB,SVA :                  D, I
 IVO,IVV :         H, X,    D, I, U
 IFO,IFV :         D, I

SEULES LES MATRICES CI-DESSOUS QUI SONT DEJA DEFINIES PEUVENT ETRE MASQUEES:
 - LA MATRICE DES DONNEES /DON/

QUELLE DONNEE ? /DON/IVO/IVV/IFO/IFV/POB/PVA/SOB/SVA/?/ ←⊞→ DON
MASQUER OU DEMASQUER ? /MAS/DEM/ ←⊞→ ?

POSSIBILITES OFFERTES PAR LE MODULE MASQUE:
  /MAS/ : MASQUER UNE MATRICE
  /DEM/ : DEMASQUER UNE MATRICE

MASQUER OU DEMASQUER ? /MAS/DEM/ ←⊞→ MAS
MASQUER LES OBSERVATIONS ? /O/N/ ←⊞→ ?

L'INTERROGATION : 'MASQUER OU DEMASQUER LES OBSERVATIONS ', DOIT ETRE PRISE
AU SENS LARGE DU TERME : MASQUER LES LIGNES DE LA MATRICE CROISIE.

MASQUER LES OBSERVATIONS ? /O/N/ ←⊞→ O
DE QUELLE MANIERE ? /POS/NEG/ENU/NUM/?/ ←⊞→ ?
```

```
POSSIBILITES POUR DEFINIR UN MASQUE:
  /POS/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A MASQUER
  /NEG/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A NE PAS MASQUER
  /ENU/ : L'UTILISATEUR PEUT SELECTIONNER, A L'AIDE DES CHIFFRES 0 ET 1
          LES NOMS QU'IL VEUT MASQUER. LES NOMS APPARAISSENT SUR LE TERMINAL.
  /NUM/ : MASQUER TOUS LES NOMS COMPRIS ENTRE DEUX NOMS

CONSEIL:
SI L'UTILISATEUR NE CONNAIT PAS EXACTEMENT LES NOMS DES OBSERVATIONS
OU DES VARIABLES A MASQUER, IL EST RECOMMANDE D'APPELER LE MODULE
/I/ POUR LES IMPRIMER.

DE QUELLE MANIERE ? /POS/NEG/ENU/NUM/?/ ←⊞→ ENU
→ ENTREZ AU-DESSOUS DES NOMS DES OBSERVATIONS LA VALEUR 0 SI LE NOM DOIT ETRE
MASQUE ET 1 SINON
QUALITE      TECHNIQUE    PRECISION    ESTHETIQUE    FIABILITE    VALEUR
   1              0            1             0             0           1
SOLIDITE     ELEGANT      ELECTRON
1               1            0
MASQUER LES VARIABLES ? /O/N/ ←⊞→ ?

L'INTERROGATION : ' MASQUER LES VARIABLES ', DOIT ETRE PRISE AU SENS
LARGE DU TERME: MASQUER LES COLONNES DE LA MATRICE CHOISIE.

MASQUER LES VARIABLES ? /O/N/ ←⊞→ O
DE QUELLE MANIERE ? /POS/NEG/ENU/NUM/?/ ←⊞→ ?

POSSIBILITES POUR DEFINIR UN MASQUE:
  /POS/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A MASQUER
  /NEC/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A NE PAS MASQUER
  /ENU/ : L'UTILISATEUR PEUT SELECTIONNER, A L'AIDE DES CHIFFRES 0 ET 1
          LES NOMS QU'IL VEUT MASQUER. LES NOMS APPARAISSENT SUR LE TERMINAL.
  /NUM/ : MASQUER TOUS LES NOMS COMPRIS ENTRE DEUX NOMS

CONSEIL:
SI L'UTILISATEUR NE CONNAIT PAS EXACTEMENT LES NOMS DES OBSERVATIONS
OU DES VARIABLES A MASQUER, IL EST RECOMMANDE D'APPELER LE MODULE
/I/ POUR LES IMPRIMER.

DE QUELLE MANIERE ? /POS/NEG/ENU/NUM/?/ ←⊞→ NUM
→ DONNER LES DEUX NOMS DES BORNES DEFINISSANT L'INTERVALLE A MASQUER (BORNES Y
COMPRIS) ←⊞→ 7MARQUE  11MARQUE
QUEL BRANCHEMENT ? /MOB/MVA/DOB/DVA/?/ ←⊞→ ?

LISTE DES BRANCHEMENTS DANS LE MODULE MASQUE:
  /MOB/ : MASQUER LES OBSERVATIONS SUR LE MEME JEU DE DONNEES
  /MVA/ : MASQUER LES VARIABLES SUR LE MEME JEU DE DONNEES
  /DOB/ : DEMASQUER LES OBSERVATIONS SUR LE MEME JEU DE DONNEES
  /DVA/ : DEMASQUER LES VARIABLES SUR LE MEME JEU DE DONNEES
```

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←⊞→ N

QUEL BRANCHEMENT ? /MOB/MVA/DOB/DVA/?/ ←⊞→ I DON O [O]

****
MODULE D'IMPRESSION /I/
****

IMPRESSION DE LA MATRICE DES DONNEES

	1MARQUE	2MARQUE	3MARQUE	4MARQUE	5MARQUE	6MARQUE
QUAL	345	828	46	132	1348	492
PREC	173	586	11	110	1156	410
VALE	181	396	20	88	779	318
SOLI	183	611	17	72	1105	290
ELEC	481	475	17	71	879	209

NOMS DES VARIABLES /O/N/ ←⊞→ M

****
MODULE DE MASQUE /M/
****

QUELLE DONNEE ? /DON/IVO/IVV/IFO/IFV/POB/PVA/SOB/SVA/?/ ←⊞→ DON
MASQUER OU DEMASQUER ? /MAS/DEM/ ←⊞→ ?

POSSIBILITES OFFERTES PAR LE MODULE MASQUE:
/MAS/ : MASQUER UNE MATRICE
/DEM/ : DEMASQUER UNE MATRICE

MASQUER OU DEMASQUER ? /MAS/DEM/ ←⊞→ DEM
DEMASQUER LES OBSERVATIONS ? /O/N/ ←⊞→ O
DE QUELLE MANIERE ? /TOU/POS/NEG/ENU/NUM/?/ ←⊞→ ?

POSSIBILITES POUR DEMASQUER:
/TOU/ : DEMASQUER TOUTES LES OBSERVATIONS OU TOUTES LES VARIABLES
/POS/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A DEMASQUER
/NEG/ : ENTRER LA LISTE EXHAUSTIVE DES NOMS A NE PAS DEMASQUER
/ENU/ : L'UTILISATEUR PEUT SELECTIONNER A L'AIDE DES DEUX CHIFFRES 0 ET 1 LES NOMS QU'IL VEUT DEMASQUER
LES NOMS APPARAISSENT AU CLAVIER
/NUM/ : DEMASQUER TOUS LES NOMS COMPRIS ENTRE DEUX NOMS

CONSEIL:
SI L'UTILISATEUR NE CONNAIT PAS EXACTEMENT LES NOMS DES OBSERVATIONS OU DES VARIABLES A MASQUER, IL EST RECOMMANDE D'APPELER LE MODULE /I/ POUR LES IMPRIMER.

DE QUELLE MANIERE ? /TOU/POS/NEG/ENU/NUM/?/ ↵ TOU
DEMASQUER LES VARIABLES ? /O/N/ ↵ O
DE QUELLE MANIERE ? /TOU/POS/NEG/ENU/NUM/?/ ↵ TOU
QUEL BRANCHEMENT ? /MOB/MVA/DOB/DVA/?/ ↵ I DON N O O S O

****
MODULE D'IMPRESSION /I/
****

IMPRESSION DES NOMS DES VARIABLES DE LA MATRICE DES DONNEES

1MARQUE
2MARQUE
3MARQUE
4MARQUE
5MARQUE
6MARQUE
7MARQUE
8MARQUE
9MARQUE
10MARQUE
11MARQUE

IMPRESSION DES NOMS DES OBSERVATIONS DE LA MATRICE DES DONNEES

QUALITE
TECHNIQUE
PRECISION
ESTHETIQUE
FIABILITE
VALEUR
SOLIDITE
ELEGANT
ELECTRON

FIN DE L'EXECUTION DU PACKAGE

2.3.5. /O/ Module de deaain

11 permet de repréaenter graphiquement un ou plusieura nuages de pointa.

Un exemple eat donné au chapitre 3, paragraphe 7.

2.3.6. /V/ Module d'initialisation

11 permet de détruire (effacer) un jeu de donnéea. Le but est de rendre disponible de la place en mémoire centrale. Ce module ne doit être utilisé qu'en cas de nécessité, car les initialisations se font automatiquement par le package, lors de l'appel à certains modules.
Un exemple eat présenté au chapitre 5, paragraphe 3.

PRELIMINAIRES STATISTIQUES

Afin d'alléger les descriptions des méthodes statistiques de CLASFAC, nous allons introduire quelques notions statistiques communes à plusieurs méthodes.

1. PRELIMINAIRES

Les propos tenus dans ce paragraphe ne sont valables que pour des matrices de données quantitatives.

1.1. Les distances

1.1.1. Les propriétés des distances

Puisqu'une simple contemplation de la matrice des données ne permet pas de découvrir la structure des données, un premier pas serait franchi s'il était possible de définir un outil qui mesure la dissemblance ou la ressemblance entre les objets. Deux outils sont proposés dans la littérature statistique : l'indice ou la fonction de distance et l'indice de similarité. Nous ne nous intéresserons qu'à la fonction de distance qui mesure la dissemblance entre deux objets X_i et X_j et qui est utilisée dans de nombreuses méthodes statistiques de CLASFAC.
Si nous considérons C, l'ensemble des objets de la matrice des données (il peut s'agir des observations ou des variables), une fonction de distance d est une application de CxC dans R - elle associe donc à toute paire d'objets un nombre réel - qui vérifie les propriétés suivantes :

(1) $d(X_i,X_j) \geq 0$

(2) $d(X_i,X_j) = d(X_j,X_i)$

(3) $d(X_i,X_j) = 0$ si $X_i=X_j$

(4) si $X_i \neq X_j$ alors $d(X_i,X_j) > 0$

Une fonction de distance est une distance métrique si l'inégalité triangulaire est vérifiée.

$$(5)\quad d(X_i,X_j) \leq d(X_i,X_k)+d(X_k,X_j)$$

L'inégalité triangulaire signifie que trois objets forment un triangle ou sont alignés.

Remarque:
Par la suite, nous dirons distance à la place de fonction de distance, même si les "distances" étudiées n'en sont pas au sens de la topologie générale.

1.1.2. Les distances de CLASFAC

Rappelons que si X_i et X_j sont deux objets, la notation adoptée est la suivante :

$$X_i = (x_i^1 , x_i^2 , \ldots, x_i^n)$$

$$X_j = (x_j^1 , x_j^2 , \ldots, x_j^n)$$

La distance de Minkowski

On peut définir toute une famille de distances par la formule:

$$d(X_i,X_j) = \left[\sum_{k=1}^{n} w_k \, (x_i^k - x_j^k)^r \right]^{1/r}$$

Les différentes distances de Minkowski s'obtiennent en se donnant des valeurs particulières à r et à w_k, poids de la k-ième variable.

1) Les distances quadratiques

Nous allons tout d'abord nous intéresser aux distances quadratiques, pour lesquelles r=2.

En élevant au carré les écarts, les distances quadratiques privilégient les fortes différences au détriment des petites.

a) La distance euclidienne

$$d^2(X_i,X_j) = \sum_{k=1}^{n} w_k \, (x_i^k - x_j^k)^2$$

Parmi toutes les distances, la distance euclidienne est celle qui nous paraît la plus familière, car elle correspond à la longueur d'un segment entre deux points dans un espace métrique à deux dimensions.
Cette distance très souvent utilisée présente des caractéristiques qui, si elles sont ignorées de l'utilisateur, peuvent conduire à des résultats statistiques inattendus. La distance euclidienne dépend de l'unité de mesure et de la variance de chaque variable. De plus, elle est sensible à "l'effet de taille" : si nous multiplions toutes les valeurs d'un objet par s, les distances de cet objet aux autres objets sont toutes également multipliées par s. Une transformation statistique, appelée la normalisation permet de palier à ces difficultés. Elle est expliquée dans le prochain paragraphe.

b) La distance du CHI2

$$d^2(X_i,X_j) = \sum_{k=1}^{n} \left[\frac{x_i^k}{x_i} - \frac{x_j^k}{x_j} \right]^2 \frac{x}{x^k}$$

avec $x_i = \sum_k x_i^k$

$$x^k = \sum_i x_i^k$$

$$x = \sum_i \sum_k x_i^k$$

Le vecteur

$$\left(\frac{x_i^1}{x_i}, \frac{x_i^2}{x_i}, \ldots, \frac{x_i^k}{x_i}, \ldots, \frac{x_i^n}{x_i} \right)$$

s'appelle le profil de l'objet i.
Cette distance ne peut être calculée que si les variables sont mesurées sur une même échelle et avec une même unité et si $x_i^j \geq 0 \quad \forall i,j$. Une matrice de contingence (tableau de fréquences) répond, par exemple, à tous ces impératifs.
Cette distance fait disparaître "l'effet de taille" reproché à la distance euclidienne, puisqu'elle calcule la somme des écarts au carré des profils des objets pondérés par l'inverse de l'importance de chaque variable.
La distance entre deux objets ayant même profil est nulle; cette propriété est utilisée dans la méthode d'analyse factorielle des correspondances.
Remarquons que la distance du CHI2 n'est en fait pas une fonction de distances entre les objets originels, notés ici X_i et X_j, mais entre leurs profils.

2) Les distances pour lesquelles r=1

A l'encontre des distances quadratiques, elles privilégient les petites distances au détriment des grandes.

a) La distance de Minkowski d'ordre 1

$$d(X_i, X_j) = \sum_{k=1}^{n} |x_i^k - x_j^k|$$

Comme cette distance peut être comparée à la distance parcourue par un promeneur dans une grande ville des Etats-Unis, formée d'avenues longitudinales et transversa-

lea, la littérature anglo-saxonne l'a baptiséa "city-block métrique" ou "métrique de Manhattan".

b) La diatance de Canberra

$$d(X_i,X_j) = \sum_{k=1}^{n} |x_i^k - x_j^k| \quad /(|x_i^k| \quad + \quad |x_j^k|)$$

Cette diatance eat une variante de la distance de Minkowski d'ordre 1.

3) La distance de Minkowski d'ordre infini

En faisant tendre le paramètre r vers l'infini, on obtient la diatance de Minkowaki d'ordre infini. Cette distance privilégie les fortes distances -en fait la plus forte- et eat donc plus robuste.

$$d(X_i,X_j) = \underset{k=1,n}{\text{Max}} \; |x_i^k - x_j^k|$$

Diatancea utiliaées dans lea différentes méthodes et modulea

Le tableau 2.2 indique lea modules statistiques qui tiennent compte dana leur algorithme de la notion de distance.

1.1.3. La matrice des distancea

Si nous calculons les distances entre tous les objets de la matrice dea donnéea, nous obtenons une matrice de distances d de taille mxm. Elle est aymétrique $\forall i,j \; i \neq j$ $d(X_i,X_j)=d(X_j,X_i)$ et sa diagonale eat formée de 0, $d(X_i,X_i)=0$. Cea propriétéa découlent de la définition de la fonction de diatance.

	distance euclidienne	distsnce du CMI2	distance de Minkowski d'ordre 1	distsnce de Csnbarra	diatance de Minkowski d'ordre infini
analyse fsctorielle en composantes principales	X				
snslyse factorielle des correspondsnces		X			
projections d'un nusge de points dans l'espace fsctoriel	X	X			
partitionnement psr la méthode de KMEAN	X	X			
psrtitionnement psr la méthode MSA	X	X			
claaaification hiérsrchique sscendante	X	X	X	X	X
clssslfication hiérsrchiqus dsscendante polythétiqus	X				

Tableau 2.2

1.2. Normalisation d'une matrice de données quantitatives

Même si la matrice des données brutes (définie par le module /E/) est une matrice quantitative, une comparaison de ses variables n'est pas possible si les variables sont mesurées avec des unités différentes. Les unités de la matrice des données formée des variables poids, taille, tension artérielle et vitesse de sédimentation, par exemple, sont exprimés en kg, mm, ... Si nous calculons la distance (différente de celle du CHI2) entre ses objets, la variable "taille" annihile presque totalement l'effet des autres variables. Dès lors, la normalisation de la matrice des données va s'imposer. Elle permet de supprimer la pondération intrinsèque des variables et de construire une matrice de données indépendantes des unités de mesure.
CLASFAC propose de centrer ou de centrer et de réduire la matrice des données. La deuxième possibilité correspond à la normalisation de la matrice.
Si $\bar{x}^j=(1/m)\sum x_i^j$ est la moyenne de la variable j et $s_j^2=(1/m)\sum(x_i^j-\bar{x}^j)$ sa variance, centrer la matrice des données brutes revient à remplacer l'élément x_i^j par $x_i^j-\bar{x}^j$. La moyenne de toute la variable centrée est alors nulle.
Centrer et réduire la matrice des données brutes revient à remplacer l'élément x_i^j par $(x_i^j-\bar{x}^j)/s_j$. La moyenne de toute les variables est alors égale à 0 et leur écart-type vaut 1.

Dans les modules statistiques /A/U/P/B/H/X/, l'utilisateur peut analyser soit la matrice "brute", soit la matrice centrée, soit la matrice centrée et réduite. Si l'utilisateur fait plusieurs analyses sur un même jeu de données, il devra toujours conserver la même option (matrice de données brute, ou centrée, ou centrée et réduite) pour que les résultats soient compatibles.

2. VUE D'ENSEMBLE DES METHODES

Les méthodes statistiques de CLASFAC appartiennent à deux

grandes familles de l'analyse statistique multivariable : l'analyse factorielle et la classification automatique.
L'analyse factorielle permet de résumer l'information contenue dans la matrice des données en remplaçant les variables par un nombre réduit de facteurs (combinaisons linéaires de variables). La méthode aboutit à une représentation graphique visualisant les similarités entre les observations et les variables par leurs proximités géométriques.
La classification automatique permet de construire des classes d'objets (méthodes de partitionnement) ou de structurer les objets en classes principales, classes, sous-classes...
Avant de décrire les différentes méthodes, nous allons les classer selon leurs caractéristiques (Tableau 2.3) et indiquer par un numéro entre parenthèses dans quel ordre les méthodes vont être exposées. Il est conseillé de respecter cet ordre pour la lecture, car les notations introduites lors de la présentation d'une méthode ne sont pas forcément redéfinies lors de l'exposé d'une autre méthode.

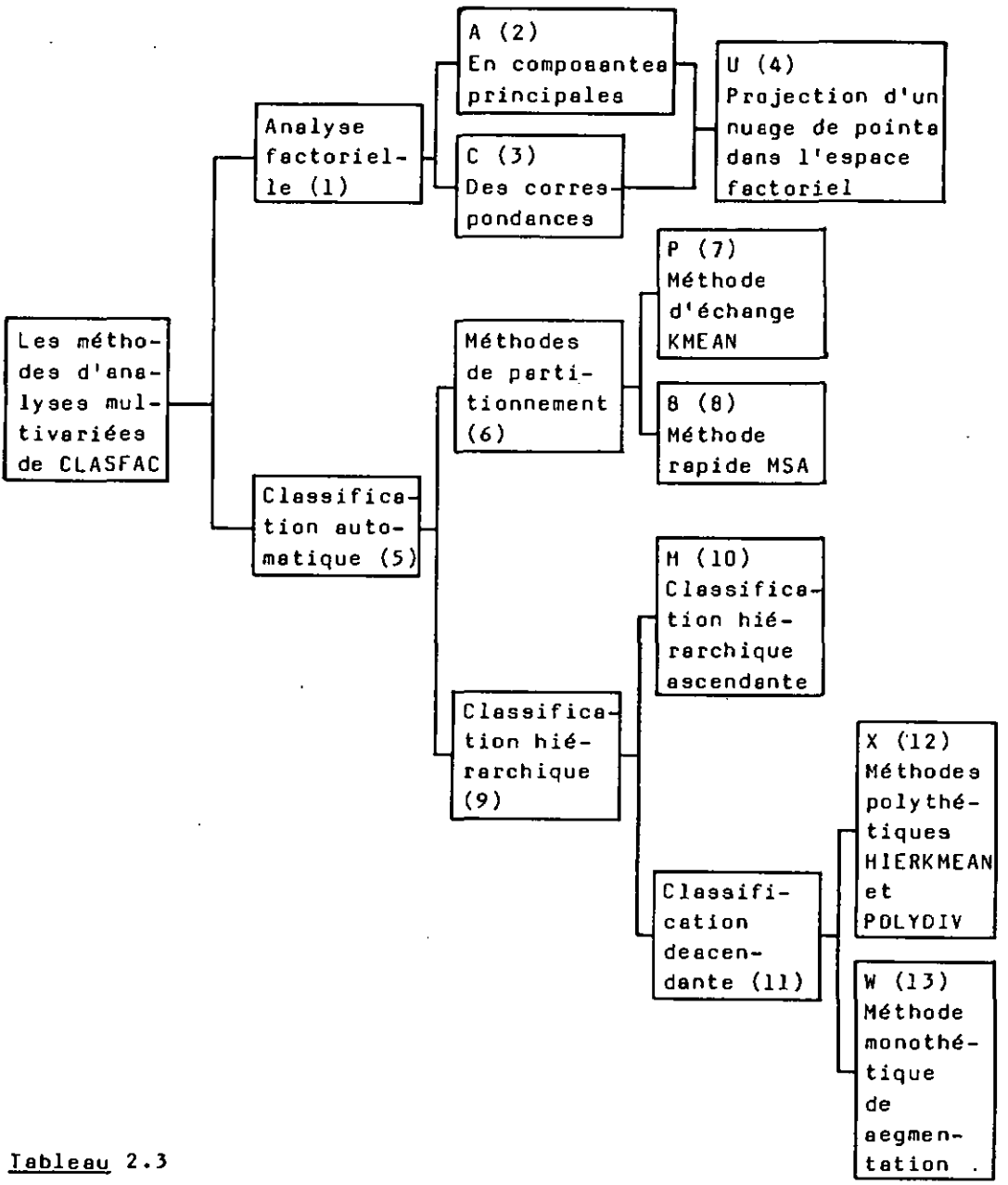

Les méthodes d'analyses multivariées de CLASFAC
Analyse factorielle (1)
A (2) En composantes principales
C (3) Des correspondances
U (4) Projection d'un nuage de points dans l'espace factoriel
Classification automatique (5)
Méthodes de partitionnement (6)
P (7) Méthode d'échange KMEAN
8 (8) Méthode rapide MSA
Classification hiérarchique (9)
M (10) Classification hiérarchique ascendante
Classification descendante (11)
X (12) Méthodes polythétiques HIERKMEAN et POLYDIV
W (13) Méthode monothétique de segmentation

Tableau 2.3

L'ANALYSE FACTORIELLE

Au lieu de présenter les principes généraux de l'analyse factorielle et de montrer les particularités de chacune des deux méthodes implantées dans CLASFAC, nous avons préféré, après un bref rappel de quelques notions de statistique, de développer en détail les principes mathématiques de l'analyse factorielle en composantes principales. L'analyse factorielle des correspondances étant basée sur les mêmes principes mathématiques, elle sera présentée comme un cas particulier.
De plus, l'emploi des différents modules de CLASFAC associés à la technique sera exposé.

1. L'ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES (Pearson, Hotelling)

1.1. Notions de base

Nous allons tout d'abord donner les formules de la moyenne, de la variance et de l'écart-type associées à une variable. Rappelons que la moyenne indique la valeur centrale des états de la variable et que la variance mesure la dispersion des états à la moyenne. Si nous notons X^j, la j-ième variable, les formules s'écrivent:

$$\bar{x}^j = \frac{1}{m}\sum_i x_i^j$$

$$VAR(X^j) = \frac{1}{m}\sum_i (x_i^j - \bar{x}^j)^2$$

$$s^j = \sqrt{VAR(X^j)}$$

La covariance et la corrélation mesurent le degré de dépendance entre deux variables. Si nous notons X^k et X^h deux variables, la covariance est définie par:

$$COV(X^k,X^h) = \sum_{i=1}^{m} (x_i^k - \bar{x}^k)\ (x_i^h - \bar{x}^h)$$

La covariance peut être interprétée comme le produit scalaire entre deux variables centrées. De ce fait, si les variables X^k et X^h sont indépendantes, la covariance est nulle. La réciproque est fausse et il faut donc être prudent lors de l'interprétation de ce coefficient.
Si nous calculons les indices de covariance entre toutes les variables de la matrice des données, nous obtenons la matrice de variances-covariances de taille nxn. Cette matrice est symétrique et sa diagonale est formée de la variance, car pour i=j on a $COV(X^i,X^j)=VAR(X^i)$, propriété qui découle de la définition même de la covariance.

Le coefficient de corrélation est défini par:

$$\rho(X^k,X^h) = \frac{COV(X^k,X^h)}{\sqrt{VAR(X^k)VAR(X^h)}}$$

Le coefficient de corrélation $\rho(X^k,X^h)$ prend ses valeurs entre -1 et 1. Il atteint les bornes 1 ou -1, si X^k est une fonction linéaire de X^h, $X^h=aX^k+b$.
De plus, comme ρ est défini à partir du coefficient de covariance, ρ possède également la propriété suivante: si X^k et X^h sont indépendantes, alors $\rho(X^k,X^h)=0$.
Si nous calculons les indices de corrélation entre toutes les variables de la matrice des données, nous obtenons la matrice des coefficients de corrélation de taille nxn. Cette matrice est symétrique et sa diagonale est formée de 1.

1.2. L'esprit de la méthode

Pour chaque variable, on peut calculer la moyenne, puis l'écart à la moyenne. Si la matrice des données n'est formée que de deux variables, on peut présenter ces écarts à la moyenne graphiquement dans le plan. Il suffit de dessiner

deux axes correspondant aux deux variables et de repérer les observations par leurs écarts à la moyenne et les variables par leurs écarts-types. Ce procédé est difficilement praticable pour trois variables, puisqu'il nécessite la construction d'une figure à 3 dimensions et il devient impossible pour un nombre de variables supérieur à 3. C'est ici qu'intervient l'analyse factorielle. Indépendamment du nombre de variables retenues pour l'analyse, on va montrer que les observations et les variables peuvent être représentées sur un graphique moyennant une certaine perte d'information.

Le plan de représentation sera défini de manière à ce que les positions mutuelles des observations soient reproduites le plus fidèlement possible et que les variables ayant des valeurs presque identiques pour presque toutes les observations soient représentées par des points proches. Au lieu de faire coïncider chaque axe avec une variable, un axe correspondra à une somme pondérée de variables (combinaison linéaire de variables). L'ensemble des coefficients de pondération constitue un facteur et l'axe associé est appelé un axe factoriel. La position d'une observation sur un axe est appelée son score ou la composante de l'observation. Cette position se calcule en faisant une somme pondérée des valeurs de l'observation pour les variables et en divisant le résultat par la racine carrée du nombre de variables.

La position d'une variable sur un axe sera donnée, à un coefficient de proportionnalité près, par sa covariance (si la matrice de données est centrée) ou par sa corrélation (si la matrice de données est centrée et réduite) avec les scores des observations sur ce même facteur. Cette valeur est encore égale au coefficient de corrélation entre la variable et les scores, multiplié par l'écart-type de la variable.

La moyenne des variables définit l'observation moyenne. Chaque observation peut être située par rapport à cette observation moyenne. Si la matrice des données a été préalablement centrée ou centrée et réduite, chaque observation peut être située par rapport à cette observation moyenne. Sur un axe factoriel l'observation moyenne est

représentée par l'origine.
La fidélité avec laquelle la position effective d'une observation est reproduite sur un axe est mesurée par la contribution du facteur à l'observation. Le cumul des contributions des deux premiers facteurs indique la qualité avec laquelle l'observation est représentée dans le plan factoriel. Pour l'ensemble des observations, on peut calculer la qualité moyenne de l'information qui est reproduite par le plan factoriel. La fidélité n'est pas égale à 100 pourcent, mais elle est supérieure à l'information qui pourrait être représentée sur un plan défini par deux variables. De plus, le pouvoir explicatif des facteurs est le même pour les positions des variables et pour celles des observations.
On peut, en plus des observations et des variables qui ont participé à déterminer les axes factoriels, représenter n'importe quelle autre observation ou variable dans l'espace factoriel Une telle observation ou variable est alors dite "supplémentaire".

1.3. Aspects mathématiques de la méthode

1.3.1. La transformation de la matrice des données

Comme nous le verrons plus loin, l'analyse factorielle en composantes principales mesure la distance entre deux observations par la distance euclidienne.
Avant d'effectuer l'analyse factorielle, on transforme souvent la matrice des données X. La nouvelle matrice ainsi obtenue "par standardisation" sera notée Y. Un élément de Y s'écrit selon l'option choisie dans le package:

$y_i^j = x_i^j$: matrice brute

$y_i^j = x_i^j - \bar{x}^j$: matrice centrée

$y_i^j = (x_i^j - \bar{x}^j)/s^j$: matrice centrée et réduite

1.3.2. La distance entre observations

La distance entre deux observations Y_i et Y_k est :

$$d^2(Y_i,Y_k) = \sum_j (y_i^j - y_k^j)^2$$

Une des caractéristiques du nuage des observations peut être mesurée par la moyenne des distances mutuelles :

$$d^2 = \frac{1}{m^2} \sum_{i,k} d^2(Y_i,Y_k)$$

Si nous considérons un plan P dans l'espace R^n des variables, nous pouvons projeter les observations dans ce plan et calculer les distances mutuelles entre les projections. Le plan P peut être défini par deux vecteurs orthonormés V(1) et V(2):

$$V(1) = (v^1(1)\ v^2(1) \ldots v^j(1) \ldots v^n(1))$$

$$V(2) = (v^1(2)\ v^2(2) \ldots v^j(2) \ldots v^n(2))$$

Ces deux vecteurs définissent des axes orthogonaux dans le plan P et on peut calculer les coordonnées ou les composantes des observations sur ces axes:

$$k=1,2 \quad F(k;Y_i) = \sum_j y_i^j\ v^j(k)$$ composante de l'observation Y_i sur l'axe k

Comme ces axes sont orthogonaux, on peut appliquer le théorème de Pythagore et calculer la distance entre les projections de deux observations:

$$d_p^2(Y_i,Y_k) = (F(1;Y_i)-F(1;Y_k))^2 + (F(2;Y_i)-F(2;Y_k))^2$$

ainsi que la moyenne des carrés des distances mutuelles des projections:

$$d_p^2 = \frac{1}{m^2} \sum_{i,k} d_p^2(Y_i, Y_k)$$

1.3.3. Définition des facteurs du nuage des observations

Rappelons que les observations constituent un nuage de points dans l'espace R^n. L'information représentée par le nuage sera mesurée par la moyenne des distances mutuelles d^2. Dès qu'on se donne un plan P dans l'espace R^n, on peut y représenter les observations par leurs coordonnées sur un système d'axes définissant ce plan. Avec les notations précédentes, un élément du nuage projeté sera représenté par le point :

$$(F(1;Y_i),\ F(2;Y_i))$$

Après projection, l'information représentée dans le plan sera d_p^2. Dans l'analyse factorielle en composantes principales, on cherche à déterminer le plan P tel que la perte d'information soit minimale ou, autrement dit, on cherche le plan P tel que :

$$d_p^2 \text{ soit maximal}$$

Au lieu de projeter dans un plan, on pourrait aussi projeter sur une droite. Si D(j) est la droite définie par le vecteur V(j), l'information résumée par la projection sur cette droite est :

$$d_{D(j)}^2 = \frac{1}{m^2} \sum_{i,k} (f(j;Y_i) - f(j;Y_k))^2$$

Pour des droites orthogonales on a :

$$d_p^2 = d_{D(1)}^2 + d_{D(2)}^2$$

Au lieu de choisir des vecteurs arbitraires, on choisira la droite D(1) telle que $d_{D(1)}^2$ soit maximum, ce sera le premier

axe factoriel, puis une droite D(2) orthogonale à D(1) et telle que $d^2_{D(2)}$ soit maximum, appelée deuxième axe factoriel. On peut montrer que cette optimisation séquentielle permet de déterminer le plan factoriel qui, a priori, a été défini par une optimisation globale.
Pour une raison, que nous verrons plus loin, le score de l'observation Y_i sur le facteur k ne sera pas $F(k;Y_i)$, mais:

$$f(k;Y_i) = F(k;Y_i)/\sqrt{n}$$

1.3.4. Calcul analytique des facteurs du nuage des observations

Soit C, la matrice de variances-covariances des variables, respectivement des coefficients de corrélation dans le cas où on a réduit les données. On peut montrer que la maximisation de $d^2_{D(1)}$ nous amène à résoudre le problème :

$$\text{maximiser } V(1)CV(1)'$$

et la maximisation de d^2_p à :

$$\text{maximiser } V(1)CV(1)' + V(2)CV(2)'$$
$$\text{sous la contrainte } V(1) \perp V(2)$$

Remarquons que le signe apostrophe désigne la transposée de la matrice.
Ce problème est bien connu en analyse numérique. Sa solution est donnée par les deux vecteurs propres V(1) et V(2) correspondant aux plus grandes valeurs propres α_1 et α_2 ($\alpha_1 > \alpha_2$) de la matrice C.
L'information totale d^2 du nuage est égale à la trace de la matrice C et la qualité de la représentation dans le plan factoriel est donc mesurée par le rapport:

$$(\alpha_1 + \alpha_2)/ \text{ trace}(C)$$

1.3.5. Qualité de la représentation des observations

La variance de l'observation Y_i par rapport à l'observation moyenne est :

$$VAR(Y_i) = \frac{1}{n}\sum_j (y_i^j - \bar{y}_i)^2$$

Sur l'axe factoriel k, la distance de l'observation Y_i à l'origine est:

$$(f(k;Y_i))^2$$

Le rapport

$$(f(k;Y_i))^2/VAR(Y_i)$$

appelé contribution du facteur k à l'observation Y_i mesure le pourcentage de la variance entre l'observation Y_l et l'observation moyenne représentée par la distance de l'observation Y_i sur le facteur k.
La qualité globale de la représentation du nuage sur le facteur k est mesurée par le rapport

$$\sum_i (f(k;Y_i)^2) \; / \sum_i VAR(Y_i)$$

qui est égal à

$$\alpha_k \; / \; traca(C)$$

1.3.6. Le nuage des variables

Il existe deux approches possibles pour projeter les variables dans le plan factoriel. On peut tout d'abord remarquer que les variables peuvent être représentées dans l'espace R^m, dont les axes canoniques sont définis par les observations. Une approche similaire à celle menée pour les observations permet donc de déterminer des axes factoriels

du nuage des variables ainsi que leurs composantes, notées $g(k,Y^j)$ dans la suite. Pour justifier la superposition des deux systèmes d'axes, ceux du nuage des observations et ceux du nuage des variables, on démontre alors algébriquement les relations :

$$f(k;Y_i) = \frac{1}{\sqrt{n \cdot \alpha_k}} \sum_j y_i^j g(k;Y^j)$$

$$g(k;Y_j) = \frac{1}{m}\sqrt{\frac{n}{\alpha_k}} \sum_i y_i^j f(k;Y_i)$$

qui font apparaître les points "observations" comme centres d'inertie des points "variables" et réciproquement.
On constate alors que les composantes des variables peuvent aussi être calculées à partir des coefficients de corrélation entre les variables et les scores des observations sur les facteurs. Notons en effet par $c(j,k)$, le coefficient de corrélation entre les vecteurs Y^j et $f(k)$. La composante de la variable j sur le facteur k peut s'écrire :

$g(k;Y^j) = s^j\, c(j,k)$ cas non réduit

ou

$g(k;Y^j) = c(j,k)$ cas réduit

La contribution de la variable Y^j au facteur k est alors naturellement définie par le rapport :

$$(g(k;Y^j))^2 \Big/ \sum_m (g(k;Y^m))^2$$

Du point de vue du calcul, on se sert de la relation suivante, moins onéreuse, pour calculer les composantes des variables:

$$g(k;Y^j) = \sqrt{\alpha_k}\, v^j(k)$$

1.3.7. Remarque concernant la méthode

Souvent on reproche à la méthode, d'être sensible "à l'effet de taille des observations". On constate alors que les observations sont rangées selon "leur taille" le long du premier (ou parfois deuxième) axe factoriel. Si, par exemple, les observations de la matrice des données sont des entreprises et les variables, le chiffre d'affaires, le nombre d'employés,... le premier axe factoriel oppose les grandes entreprises aux petites entreprises. Cet effet de taille a tendance à masquer la structure plus profonde des données.

2. L'ANALYSE FACTORIELLE DES CORRESPONDANCES (Benzécri)

Afin de sensibiliser le lecteur aux principes mathématiques de l'analyse factorielle, nous avons présenté en détail l'analyse factorielle en composantes principales. Puisque l'analyse factorielle des correspondances peut être considérée comme un cas particulier de cette méthode, seules les particularités et quelques relations mathématiques vont être données.

2.1. La matrice des données

La méthode utilise la distance du CHI2 pour mesurer la proximité entre deux observations. De ce fait, la matrice des données doit vérifier quelques conditions. Tous les éléments x_i^j doivent être positifs. De plus la somme des éléments qui forment une observation ou une variable doit être strictement positive,

$$\forall i \; x_i = \sum_j x_i^j > 0 \quad \text{et} \quad \forall j \; x^j = \sum_i x_i^j > 0$$

et posséder un sens sémantique.
La méthode s'applique le plus souvent aux tableaux de

contingence, mais son champ d'application peut s'étendre à toute matrice de données positives et homogènes.

2.2. Résumé des principales caractéristiques de la méthode

La méthode donne aux observations et aux variables un rôle symétrique. La dénomination ne subsiste que pour faciliter l'identification des données initiales.
Au lieu d'analyser les observations et les variables brutes, la méthode va prendre en considération les profils des observations et les profils des variables.
Si nous notons $x = \sum_i \sum_j x_i^j$, l'effectif total de la matrice des données, nous pouvons définir les fréquences relatives suivantes:

$$p_i^j = \frac{x_i^j}{x}$$

$p_i = \sum_j p_i^j$ fréquence relative marginale des colonnes

$p^j = \sum_i p_i^j$ fréquence relative marginale des lignes

Le vecteur P_o, formé des éléments p_i, $i \in \{1, 2, \ldots, m\}$, correspond au profil de la variable moyenne.

Le vecteur P^o, formé des éléments p^j, $j \in \{1, 2, \ldots, n\}$, correspond au profil de l'observation moyenne.

Le profil de l'observation X_i, noté R_i, peut être représenté à l'aide des fréquences conditionnelles:

$$R_i = \frac{p_i^1}{p_i}, \frac{p_i^2}{p_i}, \ldots, \frac{p_i^j}{p_i}, \ldots, \frac{p_i^n}{p_i}$$

$$R_i = \frac{x_i^1}{x_i}, \frac{x_i^2}{x_i}, \ldots, \frac{x_i^j}{x_i}, \ldots, \frac{x_i^n}{x_i}$$

Le j-ième élément du vecteur R_i sera noté, $r_i^j = p_i^j / p_i = x_i^j / x_i$.

De même le profil de la j-ième variable, noté Q^j, s'écrit:

$$Q^j = \frac{p_1^j}{p^j}, \frac{p_2^j}{p^j}, \ldots, \frac{p_i^j}{p^j}, \ldots, \frac{p_m^j}{p^j}$$

$$= \frac{x_1^j}{x^j}, \frac{x_2^j}{x^j}, \ldots, \frac{x_i^j}{x^j}, \ldots, \frac{x_m^j}{x^j}$$

Le i-ième élément du vecteur Q^j sera noté $q_i^j = p_i^j / p^j$

Cette transformation laisse déjà transparaître une caractéristique de la méthode. L'effet de taille, reproché à la méthode d'analyse factorielle en composantes principales, est supprimé, puisque deux points auront des profils identiques, si leurs coordonnées sont proportionnelles. Ainsi si nous comparons, par exemple, les causes d'invalidité dans différents pays, un grand pays tel que les Etats-Unis pourra posséder le même profil qu'un petit pays, tel que la Suisse. En anticipant sur la suite, signalons dès maintenant que des points de même profil ont même position dans l'espace factoriel.
Le but de l'analyse des correspondances peut se résumer alors à l'étude des profils des observations et des variables.

Les profils des observations R_i peuvent être représentés dans R^n comme un nuage de points munis des poids p_i. Les m points seront situés dans un sous-espace affine à n-1 dimensions, puisque les m profils vérifient la relation:

$$\sum_j \frac{p_i^j}{p_i} = 1 \qquad \forall i=\{1, \ldots, m\}$$

Les profils des variables Q^j peuvent également être considérés comme un nuage de points dans l'espace R^m où chaque point est muni d'un poids p^j. Ces n points seront situés dans un sous-espace à m-1 dimensions, puisqu'ils vérifient la relation:

$$\sum_i \frac{p_i^j}{p^j} = 1 \qquad \forall j=\{1, \ldots, m\}$$

2.3. La distance

La distance entre deux observations est mesurée par la distance du CHI2 entre leurs profils:

$$d(R_1,R_k) = \sum_j \frac{1}{p^j} \left(\frac{p_i^j}{p_i} - \frac{p_k^j}{p_k}\right)^2 = \sum_j \frac{x}{x^j} \left(\frac{x_i^j}{x_1} - \frac{x_k^j}{x_k}\right)^2$$

et la distance entre deux variables par la distance du CHI2 entre leurs profils:

$$d(Q^j,Q^k) = \sum_i \frac{1}{p_i} \left(\frac{p_i^j}{p^j} - \frac{p_i^k}{p^k}\right)^2$$

La distance du CHI2 vérifie le principe d'équivalence distributionnelle qui s'énonce de la façon suivante.
Si l'on agrège deux observations (ou variables) qui possèdent le même profil, la distance entre ces deux observations agrégées (ou variables) et les autres observations (ou variables) reste invariante.
Cette propriété garantit donc une certaine invariance des résultats par rapport à la nomenclature de la matrice des données.

Si nous réussissons à transformer la distance du CHI2 en une distance euclidienne, les calculs de l'analyse factorielle des correspondances découleront de ceux de l'analyse factorielle en composantes principales et nous pourrons nous contenter de donner les résultats finaux.
Le passage de la distance du CHI2 à la distance euclidienne se fait par le développement mathématique suivant:
Définissons la famille des vecteurs Z_i, munis des poids p_i, par la relation:

$$z_i^j = \frac{p_i^j}{p_i \sqrt{p^j}}$$

Les coordonnées du centre de gravité de la famille, noté Z_o, sont:

$$z_o^j = \sum_{i=1}^{m} p_i \frac{p_i^j}{p_i \sqrt{p^j}} = \sqrt{p^j}$$

En notant d_1 (le carré de) la distance du CHI2 et d_2 (le carré de) la distance euclidienne, on prouve en explicitant les calculs, que:

$$d_1(R_i,R_k) = d_2(Z_i,Z_k) = d_2(Z_i-Z_o,Z_k-Z_o)$$

Par un artifice mathématique, nous venons de montrer que l'analyse des correspondances peut être ramenée à une analyse en composantes principales sur les vecteurs transformés Z_i.

2.4. Quelques résultats

Le formulaire suivant résume les principales formules de l'analyse factorielle des correspondances.
La k-ième valeur propre qui correspond à l'inertie du k-ième facteur, est défini par la relation:

$$\alpha_k = \sum_j (g(k;Y^j)^2)/p^j = \sum_i (f(k;Y_1)^2)/p_i$$

Les projections (ou composantes) des varisbles et des observations sont données par les formules symétriques suivantes:

$$g(k;Y_j) = \sum_i q_i^j f(k;Y_i)/\sqrt{\alpha_k}$$ projection de la j-ième varisble sur le k-ième sxe factoriel

$$f(k;Y_i) = \sum_j r_i^j g(k;Y^j)/\sqrt{\alpha_k}$$ projection de la i-ième observation sur le k-ième sxe factoriel

La k-ième composante de la vsriable j apparaît donc comme la somme pondérée des composantes de même rang des observations, les coefficients de pondérstion étant les éléments du profil de la variable. Réciproquement la k-ième composante de l'observation i eat la somme pondérée des k-ième composantes des variables, les coefficients étant les éléments du profil de l'observation.

3. LES ETAPES OE L'ALGORITHME

Noua allons résumer les étapes de l'algorithme des deux méthodes d'analyse factorielle. Lorsque l'étape diffère selon le type d'analyse, une explication séparée sera donnée; dsns le css contrsire on donnera une explicstion unique. L'algorithme se déroule en 8 étapes.

1) Tranaformatlon de la matrice des donnéea

a) Cas de l'analyae fsctorielle en composantes principalea:
Selon l'optlon choisie par l'utilisateur, il fsut créer une matrice Y qui correspond à la matrice de données brute, centrée ou centrée et réduite.

b) Cas de l'analyae factorielle des correspondances:
Il faut tout d'abord contrôler que la matrice des données vérifie toutes les conditions nécessaires à l'exécution de la méthode et former la matrice Y dont l'élément (i,j)

sst donné psr:

$$y_i^j = \frac{p_i^j}{p_i \sqrt{p^j}} - \sqrt{p^j}$$

2) Cslcul de ls mstrice à disgonsliser

s) Css de l'snslyse fsctorielle en compossntes principsles:
Cslculer ls mstrice C=(1/n)Y'Y.
C correspond à :
- ls mstrice des produits scslsires si Y correspond à ls mstrice de données brute
- ls mstrice de covsrisnce si Y correspond à ls mstrice des données centrée
- ls mstrice de corrélation si Y correspond à ls mstrice des données centrée réduite.

b) Css de l'snslyse fsctorielle dss correspondsnces:
Cslculer ls mstrice C dont l'élément (j,j') est défini psr:

$$c_j^{j'} = \sum_i (p_i^j/(p_i\sqrt{p^j} - \sqrt{p^j}))(p_i^{j'}/(p_i \sqrt{p^{j'}} - \sqrt{p^{j'}}))$$

3) Chsrchsr les vsleurs et les vscteurs propres de ls mstrice C

4) Clssser les vsleurs propres psr ordre décroissant et "renuméroter" les vscteurs propres

5) Sélectionner les 5 premières vslsurs et vecteurs propres

6) Projeter les observations sur les sxes fsctoriels

7) Projster les vsrisbles sur les sxss fsctorisls

8) Cslculer les contributions des fscteurs sux observstions et sux vsrisbles

Les formules qui ont été utilisées dans CLASFAC proviennent de /8/.

4. LES RESULTATS OBTENUS A LA SUITE D'UNE ANALYSE FACTORIELLE

Nous allons donner les résultats fournis par CLASFAC à la suite d'une analyse factorielle. Ces résultats sont donnés dans le but de faciliter l'interprétation des axes factoriels. Les sorties d'ordinateur sont identiques pour les deux types de méthodes.

4.1. La matrice-input

L'analyse factorielle ne peut être exécutée que sur la matrice des données.

4.2. Les résultats de l'analyse factorielle

L'analyse factorielle calcule les projections (ou composantes) des observations et des variables.

4.3. Les sorties d'ordinateur qui permettent l'interprétation des résultats

Remarque préliminaire:
Si le nombre de variables est supérieur à 5, CLASFAC ne prend en compte que les 5 premiers axes factoriels.

La statistique sur les valeurs propres et vecteurs propres

CLASFAC imprime la valeur propre et l'inertie représentée par chaque axe factoriel.
Lors de l'interprétation des axes factoriels, aucune méthode ne permet de connaître a priori le nombre de facteurs qui doit être retenu, car celui-ci dépend du nombre de variables qui participent à l'analyse et de la forme du nuage des points dans l'espace. Souvent on se fixe un seuil qui correspond à un pourcentage minimum de l'inertie que l'on

veut restituer et on dessine les plans factoriels avec les axes sélectionnés. Si l'utilisateur désire, par exemple, restituer le 80 pourcent de l'inertie du nuage de points et que le premier facteur restitue 40 pourcent de l'information, le deuxième le 30 pourcent et le troisième le 10 pourcent, l'interprétation des résultats devra se baser sur les plans factoriels formés des axes 1,2 2,3 et 1,3.

Les contributions des variables et des observations

CLASFAC donne tout d'abord quelques caractéristiques pour chaque observation (et variable) de la matrice des données: sa moyenne, sa variance et le pourcentage de la variance de l'observation (ou de la variable) par rapport à la variance totale du nuage de points.
La contribution des observations (ou des variables) aux facteurs exprime la part prise par chaque observation (ou variable) dans l'inertie (ou la variance) expliquée par un facteur; cette contribution mesure donc l'influence de chacune des observations (ou variables) dans la détermination du facteur.
La contribution d'un facteur à une observation (ou une variable) exprime la part de la dispersion de l'observation (ou de la variable) représentée par la position sur le facteur. Elle mesure donc la capacité d'un facteur à rendre compte de la position "réelle" du point dans l'espace.

5. LE DEROULEMENT D'UNE ANALYSE FACTORIELLE DANS CLASFAC

Les modules d'analyse factorielle en composantes principales /A/ et d'analyse des correspondances /C/ réalisent tous les calculs numériques relatifs aux deux méthodes. Le module de représentation graphique /O/ permet alors de représenter les variables et/ou les observations dans le (ou les) plan(s) factoriel(s).
Si de plus, l'utilisateur veut créer des observations et/ou des variables supplémentaires, il devra tout d'abord les définir en appelant le module /F/. Leurs composantes sur les facteurs sont calculées par le module /U/, dit module de

projections des observations et/ou des variables supplémentaires. Le même module /D/, déjà cité permet enfin de représenter graphiquement les positions des éléments supplémentaires dans un espace factoriel.

6. LE MODULE DE PROJECTIONS DES OBSERVATIONS OU DES VARIABLES SUPPLEMENTAIRES DANS L'ESPACE FACTORIEL /U/

Ce module ne peut être exécuté que s'il a été précédé d'une analyse factorielle et donc d'un appel à un des modules /A/ ou /C/. En effet les projections des observations supplémentaires sont obtenues à partir des projections des variables et les projections des variables supplémentaires à partir des projections des observations. Lors de l'exécution du module /U/, l'utilisateur n'a pas besoin de spécifier le type d'analyse factorielle qui a été précédemment effectué, car CLASFAC l'a mémorisé et effectue automatiquement les "bons calculs". Les formules utilisées sont également tirées de /8/.

6.1. Les matrices-input

Les matrices dont les objets peuvent être projetées sont les suivantes:

- les observations supplémentaires créées par le module /F/
- les centres d'inertie de classes d'observations créés par un module de partitionnement /B/ ou /P/.
- les variables supplémentaires créées par le module /F/
- les centres d'inertie de classes de variables créés par un module de partitionnement /B/ ou /P/.

6.2. Les matrices-résultat

Le tableau suivant indique pour chaque matrice-input, le nom attribué à la matrice-résultat:

matrice-input	matrice résultat
observations supplémentaires	projections des observations supplémentaires
variables supplémentaires	projections des variables supplémentaires
centre d'inertie des observations de la matrice des données	projection des centres d'inertie des observations de la matrice des données
centres d'inertie des variables de la matrice des données	projections des centres d'inertie des variables de la matrice des données

6.3. Les aides à l'interprétation

CLASFAC imprime les coordonnées des projections des points supplémentaires sur les facteurs ainsi que leurs contributions.

7. EXEMPLES:

Un exemple complet d'analyse factorielle est présenté. Après l'exécution du module d'analyse factorielle des correspondances (module /C/), deux observations supplémentaires appelées BEAUTE et DURABLE, seront définies (module /F/). BEAUTE représente la somme des observations QUALITE, ESTHETIQUE et ELEGANT et DURABLE la somme des observations FIABILITE, PRECISION, QUALITE et SOLIDITE.
Puis on imprimera (module /I/), on projettera (module /U/) et dessinera (module /D/) ces observations supplémentaires dans le plan factoriel. Finalement on appellera le module trace (module /T/).

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←↵→ N
MODULE CHOISI ? /E/P/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←↵→ C

*******************************************************
*MODULE D'ANALYSE FACTORIELLE DES CORRESPONDANCES /C/*
*******************************************************


LES VECTEURS SONT ORTHONORMES
LA TRANSFORMEE PAR LA MATRICE EST PARALLELE AUX VECTEURS PROPRES


PARAMETRES POUR L'ANALYSE FACTORIELLE DES CORRESPONDANCES /C/ (27.12.1982 3H 3)
*************************************************************
LA MATRICE DES DONNEES

STATISTIQUE SUR VALEURS PROPRES ? /O/N/ ←↵→ O

STATISTIQUE SUR VALEURS PROPRES

FACTEURS                1FACTEUR    2FACTEUR    3FACTEUR    4FACTEUR    5FACTEUR
VALEURS PROPRES          0.07610     0.02311     0.01137     0.00199     0.00062
POURCENTAGE              0.66957     0.20331     0.10003     0.01747     0.00547
POURCENTAGE CUMULE       0.66957     0.87288     0.97290     0.99037     0.99584
TEST                    1.1067E3    4.2577E2    9.0753E1    3.2255E1    1.3948E1

TEST TOTAL   3.3492E3          TRACE          0.114

IMPRESSION DES COMPOSANTES DES VARIABLES ? /O/N/ ←↵→ O

COMPOSANTES DES VARIABLES

NO  NOM          1FACTEUR    2FACTEUR    3FACTEUR    4FACTEUR    5FACTEUR
 1  1MARQUE      -0.4840      0.3032      0.1642     -0.0088     -0.0353
 2  2MARQUE      -0.1805     -0.0699     -0.1340      0.0325     -0.0089
 3  3MARQUE       0.1306     -0.0052      0.1755      0.2124      0.0070
 4  4MARQUE       0.3887     -0.0690      0.2186     -0.0103     -0.0171
 5  5MARQUE       0.0056     -0.0796     -0.0441     -0.0417     -0.0045
 6  6MARQUE       0.3351     -0.1056      0.1307     -0.0015     -0.0045
 7  7MARQUE      -0.0709     -0.0573      0.0139      0.0744     -0.0263
 8  8MARQUE       0.2921      0.0401      0.0420      0.0831      0.0580
 9  9MARQUE      -0.1193     -0.0046     -0.0810      0.1112     -0.0032
10  10MARQUE     -0.2307      0.0852      0.0300     -0.0184      0.0489
11  11MARQUE      0.7890      0.5366     -0.2086     -0.0085     -0.0131
```

IMPRESSION DES COMPOSANTES DES OBSERVATIONS ? /O/N/ ←⊞→ O

COMPOSANTES DES OBSERVATIONS

NO	*NOM*	*1FACTEUR*	*2FACTEUR*	*3FACTEUR*	*4FACTEUR*	*5FACTEUR*
1	*QUALITE*	¯0.1574	¯0.1110	0.0067	0.0902	¯0.0166
2	*TECHNIQUE*	0.1701	¯0.1933	0.1113	¯0.0272	0.0407
3	*PRECISION*	¯0.0515	¯0.1573	¯0.0516	¯0.0387	¯0.0078
4	*ESTHETIQUE*	¯0.2647	0.1340	0.1421	0.0108	0.0288
5	*FIABILITE*	¯0.0393	¯0.0526	¯0.0987	¯0.0426	¯0.0187
6	*VALEUR*	0.2413	0.2898	¯0.1506	0.0257	0.0252
7	*SOLIDITE*	¯0.0800	¯0.0556	¯0.1519	¯0.0070	0.0148
8	*ELEGANT*	¯0.3411	0.1863	0.0829	¯0.0470	¯0.0300
9	*ELECTRON*	0.6476	0.0469	0.1013	0.0033	¯0.0293

IMPRESSION DES CONTRIBUTIONS RELATIVES DES VARIABLES /O/N/ ←⊞→ O

CONTRIBUTIONS RELATIVES DES VARIABLES

NO	*NOM*	*EFFECTIF*	*PROB. MARG.*	*VARIANCE*	*POURC. VAR.*
1	*1MARQUE*	2.1400E3	0.073	3.5461E¯1	0.227
2	*2MARQUE*	4.5150E3	0.153	5.6637E¯2	0.076
3	*3MARQUE*	2.2100E2	0.007	1.0637E¯1	0.007
4	*4MARQUE*	1.0600E3	0.036	2.0450E¯1	0.065
5	*5MARQUE*	9.2180E3	0.313	1.0134E¯2	0.028
6	*6MARQUE*	3.6600E3	0.124	1.4072E¯1	0.154
7	*7MARQUE*	1.9250E3	0.065	1.5613E¯2	0.009
8	*8MARQUE*	1.1790E3	0.040	1.0087E¯1	0.036
9	*9MARQUE*	5.8600E2	0.020	3.9616E¯2	0.007
10	*10MARQUE*	3.8480E3	0.131	6.4353E¯2	0.074
11	*11MARQUE*	1.1170E3	0.038	9.5443E¯1	0.318
	TOTAL	2.9469E4		1.1365E¯1	

CONTRIBUTIONS DES VARIABLES AUX FACTEURS

NO	*NOM*	*1FACTEUR*	*2FACTEUR*	*3FACTEUR*	*4FACTEUR*	*5FACTEUR*
1	*1MARQUE*	¯0.2235	0.2888	0.1723	¯0.0029	¯0.1454
2	*2MARQUE*	¯0.0656	¯0.0324	¯0.2420	0.0814	¯0.0194
3	*3MARQUE*	0.0017	0.0000	0.0203	0.1705	0.0006
4	*4MARQUE*	0.0714	¯0.0074	0.1511	¯0.0019	¯0.0169
5	*5MARQUE*	0.0001	¯0.0858	¯0.0535	¯0.2741	¯0.0100
6	*6MARQUE*	0.1833	¯0.0599	0.1866	¯0.0001	¯0.0040
7	*7MARQUE*	¯0.0043	¯0.0093	0.0011	0.1824	¯0.0728
8	*8MARQUE*	0.0449	0.0028	0.0062	0.1391	0.2168
9	*9MARQUE*	¯0.0037	0.0000	¯0.0115	0.1240	¯0.0003
10	*10MARQUE*	¯0.0913	0.0411	0.0103	¯0.0222	0.5033
11	*11MARQUE*	0.3101	0.4724	¯0.1450	¯0.0014	¯0.0105

CONTRIBUTIONS DES FACTEURS AUX VARIABLES

NO	NOM	1FACTEUR	2FACTEUR	3FACTEUR	4FACTEUR	5FACTEUR
1	1MARQUE	0.6606	0.2592	0.0761	0.0002	0.0035
2	2MARQUE	0.5752	0.0864	0.3171	0.0186	0.0014
3	3MARQUE	0.1604	0.0003	0.2894	0.4243	0.0005
4	4MARQUE	0.7387	0.0233	0.2336	0.0005	0.0014
5	5MARQUE	0.0031	0.6256	0.1918	0.1717	0.0020
6	6MARQUE	0.7981	0.0792	0.1214	0.0000	0.0001
7	7MARQUE	0.3222	0.2103	0.0123	0.3550	0.0444
8	8MARQUE	0.8458	0.0159	0.0175	0.0684	0.0334
9	9MARQUE	0.3594	0.0005	0.1656	0.3124	0.0003
10	10MARQUE	0.8273	0.1129	0.0140	0.0052	0.0372
11	11MARQUE	0.6523	0.3017	0.0456	0.0001	0.0002

IMPRESSION DES CONTRIBUTIONS RELATIVES DES OBSERVATIONS ? /O/N/ ↩ *O*

CONTRIBUTIONS RELATIVES DES OBSERVATIONS

NO	NOM	EFFECTIF	PROB. MARG.	VARIANCE	POURC VAR.
1	QUALITE	4.4370E3	0.151	4.5606E-2	0.060
2	TECHNIQUE	3.1460E3	0.107	8.1394E-2	0.076
3	PRECISION	3.2850E3	0.111	3.2533E-2	0.032
4	ESTHETIQUE	3.2610E3	0.111	1.0954E-1	0.107
5	FIABILITE	3.2640E3	0.111	1.7649E-2	0.017
6	VALEUR	2.8230E3	0.096	1.6634E-1	0.140
7	SOLIDITE	3.1910E3	0.108	3.3538E-2	0.032
8	ELEGANT	3.0620E3	0.104	1.6133E-1	0.147
9	ELECTRON	3.0000E3	0.102	4.3286E-1	0.388
	TOTAL	2.9469E4		1.1365E-1	

CONTRIBUTIONS DES OBSERVATIONS AUX FACTEURS

NO	NOM	1FACTEUR	2FACTEUR	3FACTEUR	4FACTEUR	5FACTEUR
1	QUALITE	-0.0490	-0.0803	0.0006	0.6174	-0.0670
2	TECHNIQUE	0.0406	-0.1726	0.1163	-0.0397	0.2846
3	PRECISION	-0.0039	-0.1193	-0.0261	-0.0841	-0.0108
4	ESTBETIQUE	-0.1019	0.0860	0.1965	0.0065	0.1473
5	FIABILITE	-0.0023	-0.0133	-0.0949	-0.1014	-0.0627
6	VALEUR	0.0733	0.3482	-0.1911	0.0319	0.0979
7	SOLIDITE	-0.0091	-0.0145	-0.2197	-0.0027	0.0384
8	ELEGANT	-0.1589	0.1560	0.0629	-0.1158	-0.1504
9	ELECTRON	0.5611	0.0097	0.0919	0.0006	-0.1409

CONTRIBUTIONS DES FACTEURS AUX OBSERVATIONS

NO NOM	*1FACTEUR*	*2FACTEUR*	*3FACTEUR*	*4FACTEUR*	*5FACTEUR*
1 *QUALITE*	0.5434	0.2703	0.0010	0.1785	0.0061
2 *TECHNIQUE*	0.3555	0.4589	0.1522	0.0091	0.0203
3 *PRECISION*	0.0815	0.7604	0.0818	0.0460	0.0019
4 *ESTHETIQUE*	0.6398	0.1640	0.1843	0.0011	0.0076
5 *FIABILITE*	0.0877	0.1568	0.5517	0.1030	0.0199
6 *VALEUR*	0.3501	0.5050	0.1364	0.0040	0.0038
7 *SOLIDITE*	0.1910	0.0921	0.6877	0.0015	0.0066
8 *ELEGANT*	0.7212	0.2151	0.0426	0.0137	0.0056
9 *ELECTRON*	0.9689	0.0051	0.0237	0.0000	0.0020
TOTAL	2.9469E4	1.113653275			

QUEL BRANCHEMENT ? /IST/IPV/IPO/ICV/ICO/ ←⊞→ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE D'ANALYSE FACTORIELLE DES CORRESPONDANCES:
/IST/ : IMPRESSION DE LA STATISTIQUE SUR VALEURS PROPRES
/IPV/ : IMPRESSION DES COMPOSANTES DES VARIABLES
/IPO/ : IMPRESSION DES COMPOSANTES DES OBSERVATIONS
/ICV/ : IMPRESSION DES CONTRIBUTIONS RELATIVES DES VARIABLES
/ICO/ : IMPRESSION DES CONTRIBUTIONS RELATIVES DES OBSERVATIONS

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/P/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←⊞→ N

QUEL BRANCHEMENT ? /IST/IPV/IPO/ICV/ICO/ ←⊞→ P

****
MODULE DE DEFINITION DE MATRICES SUBSIDIAIRES /P/
****

QUELLE MATRICE SUBSIDIAIRE ? /OSU/VSU/ ←⊞→ ?

CHOIX DES MATRICES SUBSIDIAIRES:
/OSU/ : MATRICE DES OBSERVATIONS SUPPLEMENTAIRES
/VSU/ : MATRICE DES VARIABLES SUPPLEMENTAIRES

QUELLE MATRICE SUBSIDIAIRE ? /OSU/VSU/ ←⊞→ OSU
QUELLE OPERATION ? /CRE/DEL/ ←⊞→ ?

```
POSSIBILITES DE TRAITEMENT POUR UNE MATRICE SUBSIDIAIRE:
  /CRE/ : CREATION DE LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES, SI ELLE N'EST
          PAS ENCORE DEFINIE, EXTENSION DE LA MATRICE DES
          OBSERVATIONS SUPPLEMENTAIRES SI ELLE EST DEJA DEFINIE
          (AUGMENTATION DU NOMBRE DES OBSERVATIONS SUPPLEMENTAIRES)
  /DEL/ : EFFACER UNE OU PLUSIEURS OBSERVATIONS SUPPLEMENTAIRES

QUELLE OPERATION ? /CRE/DEL/ <⊞> CRE
DE QUELLE FACON ? /ENT/MAT/DIS/WSP/ <⊞> ?

POSSIBILITES (DE CREER OU D'ETENDRE LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES):
  /ENT/ : ENTRER LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES AU CLAVIER
  /MAT/ : EN OPERANT SUR LA MATRICE DES DONNEES
  /DIS/ : LECTURE DE LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES SUR DISQUE
  /WSP/ : AFFECTATION DE VARIABLES DEFINIES DANS LA WORKSPACE
          IL FAUT DEFINIR 3 VARIABLES DANS LA WORKSPACE:
             - LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES
             - LES NOMS DES OBSERVATIONS SUPPLEMENTAIRES
             - LES POIDS DES OBSERVATIONS SUPPLEMENTAIRES

DE QUELLE FACON ? /ENT/MAT/DIS/WSP/ <⊞> MAT
DE QUELLE MANIERE ? /DUP/ENU/INT/FIS/FII/ <⊞> ?

POSSIBILITES
  /DUP/ : PAR DUPLICATION DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /ENU/ : OBSERVATION SUPPLEMENTAIRE SERA LA SOMME DES OBSERVATIONS DE
          LA MATRICE DES DONNEES QU'ON ENUMERE.
  /INT/ : OBSERVATION SUPPLEMENTAIRE SERA LA SOMME DES ELEMENTS DONT
          LES NOMS SONT COMPRIS ENTRE DEUX BORNES (BORNES Y COMPRISES).
  /FIS/ : OBSERVATION SUPPLEMENTAIRE SERA LA SOMME DES ELEMENTS SELECTIONNES PAR
          UN FILTRE.
  /FII/ : LES OBSERVATIONS SUPPLEMENTAIRES SERONT LES ELEMENTS SELECTIONNES PAR
          UN FILTRE

DE QUELLE MANIERE ? /DUP/ENU/INT/FIS/FII/ <⊞> ENU
→ NOMS DES OBSERVATIONS A SOMMER <⊞> QUALITE ESTHETIQUE ELEGANT
→ NOM DE LA NOUVELLE OBSERVATION SUPPLEMENTAIRE <⊞> BEAUTE
QUEL BRANCHEMENT ? /CRE/DEL/ <⊞> CRE
DE QUELLE FACON ? /ENT/MAT/DIS/WSP/ <⊞> MAT
DE QUELLE MANIERE ? /DUP/ENU/INT/FIS/FII/ <⊞> ENU [FIABILITE PRECISION *
QUALITE  SOLIDITE] [DURABLE]
QUEL BRANCHEMENT ? /CRE/DEL/ <⊞> I

*************************
*MODULE D'IMPRESSION /I/*
*************************

QUELLE DONNEE ? /DON/POB/PVA/SOB/SVA/PCO/PCV/IVO/IVV/IFO/IFV/OSU/VSU/?/ <⊞> OSU
MATRICE DES DONNEES ? /O/N/ <⊞> O [O]
```

IMPRESSION DES OBSERVATIONS SUPPLEMENTAIRES

	1MARQUE	2MARQUE	3MARQUE	4MARQUE	5MARQUE	6MARQUE
BEAU	1275	1766	91	301	3088	1012
DURA	895	2593	94	401	4742	1550

	7MARQUE	8MARQUE	9MARQUE	10MARQUE	11MARQUE
BEAU	777	353	239	1720	138
DURA	992	479	317	1822	292

NOMS DES VARIABLES /O/N/ ←⊞→ U

****
MODULE DE PROJECTIONS DE POINTS /U/
****

QUELLE DONNEE ? /OSU/VSU/IVO/IVV/TOU/?/ ←⊞→ ?

L'UTILISATEUR PEUT CALCULER LES PROJECTIONS DES MATRICES CI-DESSOUS:
/OSU/ : LES OBSERVATIONS SUPPLEMENTAIRES
/VSU/ : LES VARIABLES SUPPLEMENTAIRES
/IVO/ : LES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
/IVV/ : LES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES
/TOU/ : POUR TOUTES LES MATRICES : /OSU/VSU/IVO/IVV/ QUI SONT DEFINIES.

QUELLE DONNEE ? /OSU/VSU/IVO/IVV/TOU/?/ ←⊞→ OSU

PARAMETRES POUR LA PROJECTION DE POINTS /U/ (27.12.1982 3H 11)
****
OBSERVATIONS SUPPLEMENTAIRES
PROJECTION APRES L'ANALYSE FACTORIELLE DES CORRESPONDANCES /C/

IMPRESSION DES COMPOSANTES DES OBSERVATIONS SUPPLEMENTAIRES ? /O/N/ ←⊞→ O
IMPRESSION DES CONTRIBUTIONS DES OBSERVATIONS SUPPLEMENTAIRES ? /O/N/ ←⊞→ O

COMPOSANTES DES OBSERVATIONS SUPPLEMENTAIRES

NO NOM	1FACTEUR	2FACTEUR	3FACTEUR	4FACTEUR	5FACTEUR
1 *BEAUTE*	¯0.2422	0.0478	¯0.0694	0.0271	¯0.0067
2 *DURABLE*	¯0.0883	¯0.0958	¯0.0667	0.0079	¯0.0080

CONTRIBUTIONS RELATIVES DES OBSERVATIONS SUPPLEMENTAIRES

NONOM	*EFFECTIF*	*PROB. MARG.*	*VARIANCE*	*POURC. VAR.*
1 *BEAUTE*	1.0760E4	0.365	6.6557E¯2	0.214
2 *DURABLE*	1.4177E4	0.481	2.1560E¯2	0.091

CONTRIBUTION DES OBSERVATIONS SUPPLEMENTAIRES AUX FACTEURS

NO NOM	1*FACTEUR*	2*FACTEUR*	3*FACTEUR*	4*FACTEUR*	5*FACTEUR*
1 *BEAUTE*	-0.2815	0.0362	-0.1549	0.1350	-0.0262
2 *DURABLE*	-0.0493	-0.1912	-0.1885	0.0151	-0.0493

CONTRIBUTION DES FACTEURS AUX OBSERVATIONS SUPPLEMENTAIRES

NO NOM	1*FACTEUR*	2*FACTEUR*	3*FACTEUR*	4*FACTEUR*	5*FACTEUR*
1 *BEAUTE*	0.8815	0.0344	0.0724	0.0110	0.0007
2 *DURABLE*	0.3614	0.4258	0.2066	0.0029	0.0030

QUEL BRANCHEMENT /OSU/VSU/IVO/IVV/?/ ↵ *D*

****
MODULE DE DESSIN
****

PLAN DEFINI PAR ? /OBV/VRB/FAC/?/ ↵ *?*

POSSIBILITES DE DEFINIR LES AXES DU PLAN PAR:
/OBV/ : DES OBSERVATIONS.
EXEMPLE DES POINTS A REPRESENTER:
- VARIABLES DE LA MATRICE DES DONNEES
- CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES
/VRB/ : DES VARIABLES.
EXEMPLE DES POINTS A REPRESENTER:
- OBSERVATIONS DE LA MATRICE DES DONNEES
- CENTRE D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
/FAC/ : DES FACTEURS.
EXEMPLE DES POINTS A REPRESENTER:
- PROJECTIONS DES OBSERVATIONS
- PROJECTIONS DES VARIABLES
- PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
- PROJECTIONS DES VARIABLES SUPPLEMENTAIRES
- CENTRES D'INERTIE DES PROJECTIONS DES OBSERVATIONS
- CENTRES D'INERTIE DES PROJECTIONS DES VARIABLES
- PROJECTIONS DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
- PROJECTIONS DES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES

LISTE DES MATRICES QUI SONT DEFINIES
- LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
- LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
- LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
- LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/
- LA MATRICE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES /SOB/

PLAN DEFINI PAR ? /OBV/VRB/FAC/?/ ←⊞→ *FAC*
→ *QUELS POINTS ? MULTI REPONSES POSSIBLES /POB/PVA/SOB/SVA/PCO/PCV/IFO/IFV/* ←⊞→ *POB*
PVA SOB
→ *NUMERO DE LA COLONNE REPRESENTANT L'ABSCISSE ?* ←⊞→ *?*

NUMERO DE LA VARIABLE, OU DU FACTEUR, OU DE L'OBSERVATION QUI REPRESENTE L'ABSCISSE.
LE NUMERO CHOISI DOIT ETRE COMPRIS ENTRE 1 *ET* 5.

→ *NUMERO DE LA COLONNE REPRESENTANT L'ABSCISSE ?* ←⊞→ 1
→ *NUMERO DE LA COLONNE REPRESENTANT L'ORDONNEE ?* ←⊞→ *?*

NUMERO DE LA VARIABLE, OU DU FACTEUR, OU DE L'OBSERVATION QUI REPRESENTE L'ORDONNEE.
LE NUMERO CHOISI DOIT ETRE COMPRIS ENTRE 1 *ET* 5.

→ *NUMERO DE LA COLONNE REPRESENTANT L'ORDONNEE ?* ←⊞→ 2
MEME ECHELLE SUR LES DEUX AXES ? /O/N/ ←⊞→ *?*

/O/ : LA MEME ECHELLE EST UTILISEE SUR LES DEUX AXES
/N/ : ECHELLES DIFFERENTES SUR LES DEUX AXES, SURTOUT UTILES POUR /OBV/ ET /VRB/

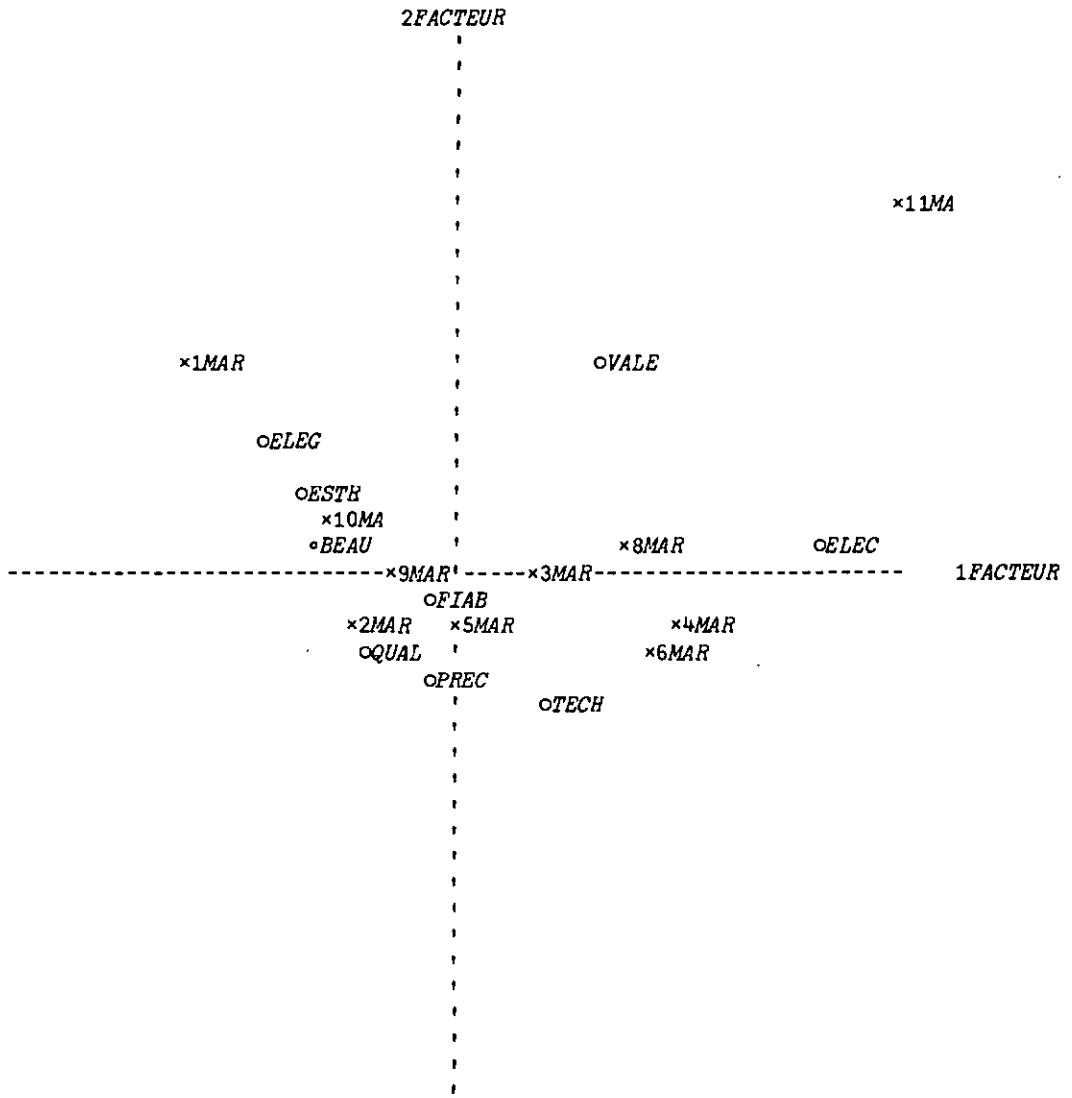
MEME ECHELLE SUR LES DEUX AXES ? /O/N/ ←⊞→ *O*
ECHELLE ? /ANC/AUT/MAN/?/ ←⊞→ *?*

L'ECHELLE PEUT ETRE DEFINIE:
/ANC/ : MEME ECHELLE QUE LORS DE LA REPRESENTATION GRAPHIQUE ANTERIEURE
/AUT/ : ECHELLE CALCULEE AUTOMATIQUEMENT PAR LE PROGRAMME
/MAN/ : ECHELLE DEFINIE PAR L'UTILISATEUR

ECHELLE ? /ANC/AUT/MAN/?/ ←⊞→ *AUT*

VALEUR DU MAXIMUM : ABSCISSE 0.78905 *ORDONNEE* 0.78905

<--------> *VAUT* 0.23371 *UNITES* 0.23371 *UNITES*



+ *CACHE* + 27.12.1982 3H 16
FIAB *SOLI*
FIAB *7MAR*
2MAR *DURA*

QUEL BRANCHEMENT ? /OBV/VRB/FAC/NOA/NOO/?/ ←⊞→ ?

LISTE DES BRANCHEMENTS DANS LE MODULE DE DESSIN:
/OBV/ : CHOIX D'UN NOUVEAU PLAN DEFINI PAR LES OBSERVATIONS
/VRB/ : CHOIX D'UN NOUVEAU PLAN DEFINI PAR LES VARIABLES
/FAC/ : CHOIX D'UN NOUVEAU PLAN DEFINI PAR LES FACTEURS
/NOA/ : CHOIX D'UN NOUVEAU NUMERO POUR L'ABSCISSE
/NOO/ : CHOIX D'UN NOUVEAU NUMERO POUR L'ORDONNEE

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←⊞→ N

QUEL BRANCHEMENT ? /OBV/VRB/FAC/NOA/NOO/?/ ←⊞→ T S O

LE SIGNE × REPRESENTE LES VARIABLES
o LES OBSERVATIONS
∘ LES OBSERVATIONS SUPPLEMENTAIRES
+ LES VARIABLES SUPPLEMENTAIRES
□o LES CENTRES D'INERTIE DES OBSERVATIONS
□× LES CENTRES D'INERTIE DES VARIABLES
⊞o LES PROJECTIONS DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
⊞× LES PROJECTIONS DES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES

****
MODULE DE TRACE AUTOMATIQUE /T/
****

ANALYSES DEJA EFFECTUEES
****
ANALYSE FACTORIELLE DES CORRESPONDANCES SUR LA MATRICE DES DONNEES
DEFINITION DE LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES
IMPRESSION DE LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES
PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES OBSERVATIONS
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES VARIABLES
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES OBSERVATIONS
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES VARIABLES
REPRESENTATION GRAPHIQUE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES

FIN DE L'EXECUTION DU PACKAGE

LA CLASSIFICATION AUTOMATIQUE

1. PRELIMINAIRES

Nous allons commencer par expliquer ce que nous entendons par "pondération des objets", car toutes les méthodes de classification, à l'exception de la classification hiérarchique descendante, font appel à cette notion.
La pondération a pour effet d'attribuer une plus ou moins grande importance à chaque objet à classer. Parmi les nombreuses pondérations possibles, nous allons décrire uniquement celles qui sont proposées dans CLASFAC.
Il est très difficile de donner des conseils en recommandant l'emploi d'une pondération plutôt que d'une autre. Seul l'utilisateur peut en décider, en tenant compte de considérations a priori sur l'objet de l'étude. Afin de choisir la meilleure pondération possible, l'utilisateur doit cependant comprendre quel usage on en fait dans la méthode statistique.

1.1. Les poids des objets dans CLASFAC

1) L'utilisateur peut donner la même importance à chaque objet. Dans ce cas, les poids des objets seront tous égaux à 1.
2) L'utilisateur peut attribuer librement des poids aux objets.
3) Les poids des objets peuvent être égaux aux marges de la matrice des données :

$$p_i = p_{X_i} = \sum_j x_i^j \Big/ \sum_i \sum_j x_i^j$$

4) les poids des objets correspondent aux poids des classes qu'ils représentent. Cette option n'est proposée que pour classer des centres d'inertie.

1.2. Typologie des méthodes

On distingue parmi les méthodes de classification automatique les méthodes polythétiques et monothétiques.
Une méthode est dite polythétique si le critère de division en classes prend en condidération l'ensemble des variables de la matrice des données. Elle est dite monothétique si une seule variable est utillsée comme critère de division.

2. GENERALITES CONCERNANT LES METHODES DE PARTITIONNEMENT

2.1. Introduction

Conformément à la définition mathématique de la partition, les méthodes de partitionnement cherchent à regrouper m objets en k classes, de telle façon qu'aucune classe ne soit vide et qu'un objet n'appartienne qu'à une et une seule classe. Parmi l'ensemble des partitions possibles, la partition sélectionnée doit maximiser ou minimiser un critère. De manière générale on cherche des classes aussi homogènes que possibles et les plus différenciées les unes des autres. Les méthodes de partitionnement appartiennent aux méthodes polythétiques, car le critère prend en considération l'ensemble des variables.
Le nombre de partitions possibles de n objets en k classes est donné par le nombre de Stirling de deuxième ordre :

$$S(m,k) = \frac{1}{k!} \sum_{j=1}^{k} C_k^j (-1)^j (k-j)^m$$

La croissance du nombre de partitions est extrêmement rapide, comme le montre le tableau suivant :

nombre d'objets	nombre de classes	nombre de partitions possibles
10	3	9330
10	5	179487
100	3	$8590\ 10^{43}$
100	5	$2316\ 10^{66}$

En aucun cas, on ne pourra donc travailler par énumération complète. Pour un petit nombre d'objets, ou des critères particuliers, la partition optimale peut être trouvée par la programmation dynamique, une des méthodes de la recherche opérationnelle. Si le nombre d'objets augmente, cette méthode devient inapplicable, car elle nécessite un temps de calcul trop important. De ce fait, la plupart des algorithmes de partitionnement sont des heuristiques. C'est ainsi que par les méthodes d'échange, on trouve une partition qui n'est qu'un optimal local.

2.2. Notations et formulaire

Dans ce paragraphe, nous introduisons les notations utilisées pour décrire les méthodes de partitionnement et rappelons quelques formules de statistique.

m : nombre d'objets à classer

k : nombre de classes

C : nuage des objets à classer, $C=\{X_1, X_2, X_3, \ldots, X_m\}$

C_i, C_j : deux classes, c'est-à-dire deux sous-ensembles de C disjoints et non vides, soit : $C_i \neq \emptyset$, $C_j \neq \emptyset$, $C_i \cap C_j = \emptyset$

A chaque objet de C, on a associé un poids noté n_x. Le poids d'une classe C_i de C est noté n_i et se définit de la façon suivante:

$$n_{C_i} = n_i = \sum_{X \in C_i} n_x$$

Remarquons que si le poids de chaque objet de C est égal à 1, le poids de la classe C_i, n_i, correspond au nombre d'objets appartenant à la classe.

Le centre d'inertie (ou de gravité) du nuage de points des objets est noté G_o et est défini par la formule:

$$G_o = \frac{1}{n_C} \sum_{X \in C} n_x \cdot X$$

Le centre d'inertie (ou de gravité) d'une classe C_i est défini par la relation :

$$G(C_i) = G(i) = \frac{1}{n_i} \sum_{X \in C_i} n_x \cdot X$$

Nous allons maintenant décrire les algorithmes de partitionnement implantés dans CLASFAC : KMEAN et la méthode séquentielle adaptative, abrégée MSA. La première méthode appartient aux méthodes d'échange et la deuxième aux méthodes rapides.

3. LE PARTITIONNEMENT PAR LA METHODE DE KMEAN (Régnier 1965, Mac Queen 1967)

3.1. Résumé de la méthode

Le nombre de classes est fixé a priori. De plus il faut se donner une partition initiale (répartition des objets dans les classes). Cette partition initiale sert d'input pour la première itération. Lors de chaque itération on améliore la partition en enlevant un élément à sa classe et en le mettant dans une autre classe.

3.2. Le choix du critère à optimiser pour KMEAN

La variation totale d'un nuage de points C représente une de ses caractéristiques.

$$V^2_{total} = \sum_{X \in C} n_x \, d(X,G_o)^2$$

La variation totale est indépendante de la partition choisie. Le théorème de Huyghens donne une relation intéressante entre la variation totale et les caractéristiques d'une partition. Cette relation n'est vérifiée que pour une distance quadratique. Elle peut s'exprimer de la façon suivante : la variation totale est égale à la somme de la variation entre les classes et la somme des variations dans les classes.

$$V^2_{total} = V^2_{inter\text{-}classes} + V^2_{intra\text{-}classes}$$

$$= \sum_{i=1}^{k} n_i \, d(G(i),G_o)^2 + \sum_{i=1}^{k} \sum_{X \in C_i} n_x \, d(X,G(i))^2$$

La variation inter-classes est une mesure de la différenciation entre les classes et la somme des variations intra-classes est une mesure de l'homogénéité moyenne des classes. Comme la variation totale est constante pour un nuage de points C, il revient au même de minimiser la variation intra-classes -choix de KMEAN- ou de maximiser la variation inter-classes.

3.3. Les conditions de transfert d'un objet

La méthode de KMEAN consiste, étant donné une partition de départ, à enlever un objet de sa classe et de l'affecter à une autre, afin d'obtenir une partition meilleure. Un point X appartenant à la classe C_i est transféré dans la classe

C_j si les conditions de transfert sont vérifiées. La démarche se fait en deux étapes: la première consiste à chercher si un transfert est possible et la deuxième à sélectionner le transfert optimal parmi les transferts possibles. L'objet X, appartenant à la classe C_i peut être transféré dans la classe C_j, si la variation du critère, notée $\Delta f(X, i \to j)$ est négatif. Comme la fonction f représente la variation intra-classes, sa variation peut être donnée par la formule suivante :

$$\Delta f(X, i \to j) = f(C_1, C_2, \ldots, C_i - \{X\}, \ldots, C_j \cup \{X\}, \ldots, C_k) - f(C_1, C_2, \ldots, C_i, \ldots, C_j, \ldots, C_k)$$

$$= \frac{n_j \cdot n_x}{n_j + n_x} d(X, G(j))^2 - \frac{n_i \cdot n_x}{n_i + n_x} d(X, G(i))^2$$

Le transfert optimal est défini par la classe k(x) qui donne la plus petite variation négative pour le critère:

$$\Delta f(X, i \to k(x)) = \min_j \Delta f(X, i \to j)$$

Remarque:
Si le poids de tout objet isolé est petit par rapport au poids d'une classe, la variation du critère vaut approximativement:

$$\Delta f(X, i \to j) \sim n_x \, [d(X, G(j))^2 - d(X, G(i))^2]$$

Il en résulte qu'on affectera alors l'objet X à la classe dont le centre d'inertie est le plus proche. Ce critère est utilisé strictement par la méthode MMEAN qui est une autre méthode d'échange.

3.4. L'algorithme de KMEAN

Les principes de KMEAN ayant été définis, il est facile de résumer brièvement une itération de l'algorithme.

Données initiales

k : nombre de classes

$C_1^o, C_2^o, \ldots, C_k^o$: partition initiale

T-ième itération

Au début de la t-ième itération, la partition est :

$$C_1^{t-1}, C_2^{t-1}, \ldots, C_j^{t-1}, \ldots, C_k^{t-1}$$

Lors de l'itération, on parcourt tous les objets de C. Pour chaque objet, on vérifie si l'objet peut être affecté à une autre classe. Le transfert est effectué si les conditions décrites au paragraphe 3.3. sont remplies.
Dans l'algorithme, le transfert de l'objet X de la classe C_i dans la classe $C_{k(x)}$ est effectué en adaptant la liste des objets contenus dans la classe C_i et $C_{k(x)}$ et en calculant les nouveaux centres d'inertie de ces deux classes. En termes mathématiques:

$$C_p^t = C_p^{t-1} \quad p \neq i \text{ et } p \neq k(x)$$

$$C_i^t = C_i^{t-1} - \{X\}$$

$$C_{k(x)}^t = C_{k(x)}^{t-1} \cup \{X\}$$

$$G(i) = (n_i.G(i) - n_x.X) \frac{1}{n_i - n_x}$$

$$n_i = n_i - n_x$$

$$G(k(x)) = (n_{k(x)}.G((x(k)) + n_x.X) \frac{1}{n_{k(x)} + n_x}$$

$$n_{k(x)} = n_{k(x)} + n_x$$

Fin de l'algorithme

L'algorithme s'arrête si en parcourant tous les objets de C

aucun objet ne peut être affecté à une autre classe, c'est-à-dire:

$\forall X \in C, \quad j = (1, 2, \ldots, m) \quad \Delta f(X, 1 \to j) \geq 0$

3.5. Les options de l'algorithme de KMEAN

3.5.1. La matrice-input

L'utilisateur a le choix de classer les objets suivants :

- les observations de la matrice des données
- les variables de la matrice des données
- les projections des observations sur un espace factoriel
- les projections des variables sur un espace factoriel

3.5.2. Le nombre de classes

Le nombre de classes peut être choisi en se basant sur l'analyse des résultats d'une analyse factorielle ou en analysant le dendrogramme obtenu par une classification hiérarchique.

3.5.3. La partition initiale

CLASFAC permet de définir la partition initiale de deux façons différentes. La première se base sur la méthode de Mac Queen (1967) qui consiste à affecter les m objets à k classes de la manière suivante :

Classe	Objets
C_1^o	$= (X_1 , X_{k+1} , X_{2k+1} , \ldots)$
C_2^o	$= (X_2 , X_{k+2} , X_{2k+2} , \ldots)$
C_3^o	$= (X_3 , X_{k+3} , X_{2k+3} , \ldots)$
$\ldots$	$= (\ldots , \ldots , \ldots , \ldots)$
C_k^o	$= (X_k , X_{2k} , X_{3k} , \ldots)$

La deuxième conaiate à demander à l'utiliaateur de choiair, d'après aes connaissances a priori, la partition initiale qu'il juge la meilleure.
Les deux exemples auivanta montrent comment l'utiliaateur peut, à l'aide dea réaultata obtenus par d'autres méthodea atatiatiquea da CLASFAC, définir la partition initiale:

a) Nagy (1969) propose d'utiliser la partition initiale obtenue en coupant le dendrogramme d'une claaaification hiérarchique.
b) Il eat également poaaible d'utiliaer les réaultats d'une analyae factorielle pour définir une partition initiale. Il suffira de regrouper les objets en tenant compte da leur proximité géométrique dans le plan factoriel.
c) Il eat poaaible d'utiliaer lea réaultats obtenus par MSA, dana le but d'améliorer une telle partition.

3.5.4. Les diatances
L'utiliaateur a le choix entre les deux distances quadratiquea auivantea:

- la diatance euclidienne
- la diatance du CHI2

3.5.5. Lea teats d'arrêt
L'algorithme a'arrête si l'une ou l'autre des deux conditions est vérifiées :

a) durant une itération aucun objet ne peut être transféré dans une autre claaae
b) durant troia itérationa conaécutives, la valeur d suivante eat toujoura plus grande que -0.0001, et ceci pour tout objet X.

$$d = \frac{\Delta f(X,\ i \to k(x))}{\text{variation intra-claaaea}}$$

3.5.6. Lea réaultats de KMEAN
L'algorithme génère une matrice de centrea d'inertie aaaociée à la partition optimale. Ainai, la matrice réaultant du

partitionnement des observations de la matrice des données, est la matrice des centres d'inertie de ces observations.

3.6. Conseils complémentaires pour l'utilisateur

Les méthodes de partitionnement permettent d'atteindre un optimum local. Aucun moyen ne permet de savoir si l'optimum global est atteint. Ainsi pour s'assurer de la stabilité de la partition trouvée par rapport à la partition initiale, l'algorithme de KMEAN doit être exécuté sur plusieurs partitions initiales. De plus pour s'assurer que le nombre de classes k choisi a priori est judicieux, on peut conseiller l'exécution de KMEAN avec k, k-1 et k+1 classes.

3.7. Exemple

L'exemple a été exécuté après l'analyse factorielle (cf. chapitre 3, paragraphe 7.). Plusieurs jeux de données sont donc déjà définis. Avant d'exécuter le module de partitionnement (module /P/), on fait appel au menu dynamique (module /Q/) qui permet de connaître l'ensemble des modules qui pourraient être exécutés.

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ N
MODULE CHOISI ? /E/F/A/C/U/P/B/R/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ Q

******************
*MODULE MENU /Q/*
******************

LISTE DES ANALYSES STATISTIQUES POSSIBLES
*****************************************

ANALYSE FACTORIELLE EN COMPOSANTES PRINCIPALES /A/
 - LA MATRICE DES DONNEES /DON/

ANALYSE FACTORIELLE DES CORRESPONDANCES /C/
 - LA MATRICE DES DONNEES /DON/
```

CALCUL DE PROJECTIONS DE POINTS /U/
- *LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES /OSU/*

PARTITIONNEMENT PAR LA METHODE DE KMEAN /P/
- *LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/*
- *LES VARIABLES DE LA MATRICE DES DONNEES /VAR/*
- *LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/*
- *LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/*

PARTITIONNEMENT PAR LA METHODE SEQUENTIELLE ADAPTATIVE (MSA) /B/
- *LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/*
- *LES VARIABLES DE LA MATRICE DES DONNEES /VAR/*
- *LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/*
- *LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/*

CLASSIFICATION HIERARCHIQUE ASCENDANTE /H/
- *LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/*
- *LES VARIABLES DE LA MATRICE DES DONNEES /VAR/*
- *LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/*
- *LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/*

CLASSIPICATION NIERARCHIQUE DESCENDANTE POLYTHETIQUE /X/
- *LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/*
- *LES VARIABLES DE LA MATRICE DES DONNEES /VAR/*
- *LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/*
- *LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/*

CLASSIFICATION NIERARCRIQUE DESCENDANTE MONOTHETIQUE (SEGMENTATION) /W/
- *LA MATRICE DES DONNEES /DON/*

LISTE DES MODULES D'ENTREE-SORTIE
****

CORRECTIONS, EXTENSIONS, MEMORISATION DE LA MATRICE DES DONNEES /E/

DEPINITION DE MATRICES SUBSIDIAIRES /P/
- *MATRICE DES VARIABLES SUPPLEMENTAIRES /VSU/*

EXTENSION, REDUCTION /F/
- *MATRICE DES OBSERVATIONS SUPPLEMENTAIRES /OSU/*

MASQUER /M/
- *LA MATRICE DES DONNEES /DON/*
- *LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/*
- *LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/*
- *LA MATRICE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES /SOB/*

```
IMPRIMER /I/
 - LA MATRICE DES DONNEES /DON/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/
 - LA MATRICE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES /SDB/
 - LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES /OSU/

REPRESENTATION GRAPNIQUE /D/
 - LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
 - LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/
 - LA MATRICE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES /SDB/

EFFACER /V/
 - LA MATRICE DES DONNEES /DON/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/
 - LA MATRICE DES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES /SOB/
 - LA MATRICE DES OBSERVATIONS SUPPLEMENTAIRES /OSU/

MODULES UTILITAIRES
*******************

PLACE DISPONIBLE /R/
COMMENTAIRES /J/
MODULE D'ARRET D'EXECUTION DU PACKAGE /S/

MODULE CNOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ P

*******************************************************
*MODULE DE PARTITIONNEMENT PAR LA METHODE DE KMEAN /P/*
*******************************************************

QUELLE DONNEE ? /OBS/VAR/POB/PVA/?/ ←⊞→ ?

LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE UTILISEES POUR LE PARTITIONNEMENT:
  /POB/ : PROJECTIONS DES OBSERVATIONS
  /PVA/ : PROJECTIONS DES VARIABLES
  /OBS/ : OBSERVATIONS DE LA MATRICE DES DONNEES
  /VAR/ : VARIABLES DE LA MATRICE DES DONNEES

SEULES LES MATRICES CI-DESSOUS SONT DEJA DEFINIES:
 - LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
 - LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/
```

```
QUELLE DONNEE ? /OBS/VAR/POB/PVA/?/ ←⊞→ OBS
QUELLE DISTANCE ? /DEU/DCH/ ←⊞→ ?

CHOIX DES DISTANCES POUR L'ALGORITHME:
  /DEU/ : DISTANCE EUCLIDIENNE
  /DCH/ : DISTANCE DU CHI2

QUELLE DISTANCE ? /DEU/DCH/ ←⊞→ DCH
POIDS DES OBJETS (POUR PONDERER LES DISTANCES ENTRES OBJETS) ? /POI/POC/POU/
←⊞→ ?

LES POIDS DES OBJETS SUIVANTS PEUVENT ETRE UTILISES:
  /POI/ : POIDS IDENTIQUES TOUS EGAUX A 1
  /POC/ : POIDS EGAUX AUX MARGES
  /POU/ : POIDS CHOISIS PAR L'UTILISATEUR

POIDS DES OBJETS (POUR PONDERER LES DISTANCES ENTRES OBJETS) ? /POI/POC/POU/
←⊞→ POC
→ NOMBRE DE CLASSES ? ←⊞→ ?

LE NOMBRE DE CLASSES POUR LE PARTITIONNEMENT DOIT ETRE COMPRIS ENTRE 1 ET 9.

→ NOMBRE DE CLASSES ? ←⊞→ 3
PARTITION INITIALE ? /AUT/MAN/ ←⊞→ ?

POSSIBILITES :
  /AUT/ : LA PARTITION INITIALE EST DONNEE PAR LE PROGRAMME (PARTITION
          ALEATOIRE)
  /MAN/ : LA PARTITION INITIALE EST DONNEE PAR L'UTILISATEUR

PARTITION INITIALE ? /AUT/MAN/ ←⊞→ AUT

PARAMETRES UTILISES POUR LE PARTITIONNEMENT SELON KMEAN /P/ (27.12.1982 3H 20)
************************************************************
OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
POIDS DES OBJETS EGAUX AUX MARGES
NOMBRE DE CLASSES: 3
PARTITION DE DEPART ALEATOIRE
DISTANCE DU CHI2

IMPRESSION DES CENTRES D'INERTIE ? /O/N/ ←⊞→ O

DE QUELLE FACON ? /VAL/POU/ ←⊞→ ?

/VAL/ : IMPRESSION DES CENTRES D'INERTIE EN VALEUR
/POU/ : IMPRESSION DES CENTRES D'INERTIE EN POURCENTAGE

DE QUELLE FACON ? /VAL/POU/ ←⊞→ VAL
```

MATRICE DES CENTRES D'INERTIE

	1MARQUE	2MARQUE	3MARQUE	4MARQUE	5MARQUE	6MARQUE
1	2.238E2	6.483E2	2.350E1	1.003E2	1.186E3	3.875E2
2	4.650E2	4.690E2	2.250E1	8.450E1	8.700E2	2.600E2
3	1.050E2	3.280E2	2.733E1	1.633E2	9.120E2	5.300E2

	7MARQUE	8MARQUE	9MARQUE	10MARQUE	11MARQUE	POIDS
1	2.480E2	1.197E2	7.925E1	4.555E2	7.300E1	1.418E4
2	2.060E2	9.300E1	6.400E1	5.700E2	5.750E1	6.323E3
3	1.737E2	1.713E2	4.700E1	2.953E2	2.367E2	8.969E3

IMPRESSION DE LA COMPOSITION DES CLASSES ? /O/N/ ↵ O

IMPRESSION DE LA COMPOSITION DES CLASSES

CLASSE 1

1 *QUALITE* 3 *PRECISION* 5 *FIABILITE* 7 *SOLIDITE*

CLASSE 2

4 *ESTHETIQUE* 8 *ELEGANT*

CLASSE 3

2 *TECHNIQUE* 6 *VALEUR* 9 *ELECTRON*

QUEL BRANCHEMENT ? /DIS/NBC/PTI/CEI/CLA/ ↵ ?

LISTES DES BRANCHEMENTS POSSIBLES DANS LE MODULE DE PARTITIONNEMENT:
/DIS/ : CHOIX D'UNE NOUVELLE DISTANCE POUR L'ALGORITHME (MEME JEU DE DONNEES)
/NBC/ : CHANGEMENT DU NOMBRE DE CLASSES POUR LE PARTITIONNEMENT (MEME JEU DE DONNEES)
/PTI/ : DEFINITION D'UNE NOUVELLE PARTITION DE DEPART (MEME JEU DE DONNEES)
/CEI/ : IMPRESSION DES CENTRES D'INERTIE (DEJA CALCULES)
/CLA/ : IMPRESSION DE LA COMPOSITION DES CLASSES (DEJA CALCULEES)

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/N/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ↵ N

QUEL BRANCHEMENT ? /DIS/NBC/PTI/CEI/CLA/ ↵ NBC [4] AUT O VAL O S O

PARAMETRES UTILISES POUR LE PARTITIONNEMENT SELON KMEAN /P/ (27.12.1982 3// 22)

****

OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
POIDS DES OBJETS EGAUX AUX MARGES
NOMBRE DE CLASSES: 4
PARTITION DE DEPART ALEATOIRE
DISTANCE DU CHI2

MATRICE DES CENTRES D'INERTIE

	1*MARQUE*	2*MARQUE*	3*MARQUE*	4*MARQUE*	5*MARQUE*	6*MARQUE*
1	6.700*E*1	2.940*E*2	3.100*E*1	2.010*E*2	9.785*E*2	6.360*E*2
2	1.810*E*2	3.960*E*2	2.000*E*1	8.800*E*1	7.790*E*2	3.180*E*2
3	2.238*E*2	6.483*E*2	2.350*E*1	1.003*E*2	1.186*E*3	3.875*E*2
4	4.650*E*2	4.690*E*2	2.250*E*1	8.450*E*1	8.700*E*2	2.600*E*2

	7*MARQUE*	8*MARQUE*	9*MARQUE*	10*MARQUE*	11*MARQUE*	*POIDS*
1	1.830*E*2	1.785*E*2	4.250*E*1	2.665*E*2	1.950*E*2	6.146*E*3
2	1.550*E*2	1.570*E*2	5.600*E*1	3.530*E*2	3.200*E*2	2.823*E*3
3	2.480*E*2	1.198*E*2	7.925*E*1	4.555*E*2	7.300*E*1	1.418*E*4
4	2.060*E*2	9.300*E*1	6.400*E*1	5.700*E*2	5.750*E*1	6.323*E*3

IMPRESSION DE LA COMPOSITION DES CLASSES

CLASSE 1

2 *TECHNIQUE* 9 *ELECTRON*

CLASSE 2

6 *VALEUR*

CLASSE 3

1 *QUALITE* 3 *PRECISION* 5 *FIABILITE* 7 *SOLIDITE*

CLASSE 4

4 *ESTHETIQUE* 8 *ELEGANT*

FIN DE L'EXECUTION DU PACKAGE

4. LE PARTITIONNEMENT PAR LA METHODE SEQUENTIELLE ADAPTATIVE (MSA) (Sebestyen, Bock)

Avant de décrire la méthode MSA, nous allons indiquer quelques-unes des propriétés des méthodes rapides, famille à laquelle appartient MSA.
Comme leur nom l'indique, ces méthodes sont plus rapides que les méthodes d'échange, par le fait qu'elles ne sont pas itératives. L'affectation d'un objet à une classe est déterminé par une règle d'attribution.
Ces méthodes aboutissent à une partition sans atteindre pour autant un optimum local. A l'encontre des méthodes d'échange, elles n'ont pas besoin d'une partition initiale.

4.1. Présentation de la méthode MSA

4.1.1. Le premier parcours séquentiel

Le nombre de classes k est fixé a priori. L'algorithme commence à former une classe C_1 avec le premier objet, X_1. Puis pour chaque objet X_t, t={2, 3, ..., m}, - le parcours se fait en séquence - on applique la règle d'attribution.

4.1.2. La règle d'attribution appliquée à l'objet X_t

Situons-nous en cours de l'algorithme. h classes sont déjà créées et elles sont représentées par leurs centres d'inertie. De plus, la condition h≤k est vérifiée, ce qui signifie que le nombre de classes déjà construites est plus petit ou égal au nombre k souhaité.
La règle d'attribution se base sur deux paramètres : ρ_a le seuil d'acceptation et ρ_s le seuil de séparation. Le premier paramètre contrôle l'homogénéité des classes et le deuxième leur écartement.
Notons d_o la distance minimale entre l'objet X_t et les centres de gravité des classes déjà créées. La règle d'attribution peut être énoncée de la façon suivante:

- si la distance de l'objet X_t au centre de gravité le plus proche est inférieur au seuil d'acceptation

$(d_o \leq \rho_a)$, l'objet X_t est affecté à la classe correspondante.

- si la distance à tous les centres de gravité est supérieur au seuil de séparation et si le nombre de classes déjà créées h est inférieur à k, l'objet devient le centre d'une nouvelle classe ($d_o > \rho_s$ et $h<k$).
  Dès qu'on a créé k classes, l'objet est attribué à une classe particulière, appelée la classe des objets non-classables.
- si la distance de l'objet X_t au centre de gravité le plus proche est supérieur au seuil d'acceptation mais inférieur au seuil de séparation, l'objet est attribué à la classe des objets non-classables ($\rho_a < d_o \leq \rho_s$).

4.1.3. Le deuxième parcours séquentiel

Une fois la liste des objets épuisée, on applique la règle d'attribution une deuxième fois à tous les objets non-classables. A la fin de ce parcours, ceux d'entre eux qui ne peuvent être attribués à une classe, restent des objets non-classés.

4.2 Définition et remarques

1) La création de la classe des objets non-classés demande de définir deux partitions, la partition partielle et la partition complète. La partition partielle est formée de h classes qui ont été créées par l'algorithme, $P_t = \{C_1, C_2, \ldots, C_h\}$. La partition complète est formée de la partition partielle et de $Card(C_o)$ classes dont chacune contient un objet non-classé. Remarquons que si la classe des objets non-classés est vide, nous parlerons alors de partition complète.

2) Le résultat final est influencé par l'ordre des objets dans la matrice-input.

3) Le nombre de classes k, fixé a priori n'est par forcément atteint. Cela peut provenir soit du choix des seuils ρ_a

et ρ_a, soit de la nature même des objets.

4.3 L'algorithme de MSA

Avant de décrire l'algorithme, nous allons fixer quelques notations. Il est utile de se rappeler que la t-ième itération traite le t-ième objet, noté X_t.

h_t : nombre de classes déjà créées après le traitement de l'objet X_t, ou ce qui revient au même : à la fin de la t-ième itération

C_i^t : i-ième classe à la fin de la t-ième itération, soit après le traitement de X_t

G_i^t : centre de gravité (ou d'inertie) de la classe C_i^t

n_i^t : poids de la classe C_i^t, égal à la somme des poids des objets de cette classe

C_o^t : classe des objets non-classables, à la fin de la t-ième itération

P_t : partition partielle à la fin de la t-ième itération

Les principes de MSA ayant été définis au paragraphe précédent, il est facile de résumer brièvement l'algorithme.

Données initiales

k : nombre de classes
ρ_a : seuil d'acceptation
ρ_s : seuil de séparation

condition: $0 \leq \rho_a \leq \rho_s$

Partition partielle initiale

$h_1 = 1$

$$C_1^1 = \{X_1\}$$

$$G_1^1 = X_1$$

$$n_1 = n_{X_1}$$

$$C_o^1 = \emptyset$$

$$P_1 = \{C_1\}$$

I-ième itération

Considérons, $B=\{1, 2, \ldots, h_{t-1}\}$, l'ensemble des indices des classes déjà créées à la fin de l'itération t-1.
Nous allons décrire les opérations mathématiques effectuées lors de l'application de la règle d'attribution.

Cas 1 $\min_{j\in B} d(X_t, G_j^{t-1}) > \rho_s$ et $h_{t-1} < k$

$$h_t = h_{t-1} + 1$$

$$C_{h_t}^t = \{X_t\}$$

$$G_{h_t}^t = X_t$$

$$n_{h_t}^t = n_{X_t}$$

$$C_o^t = C_o^{t-1}$$

$$P_t = P_{t-1} \cup \{C_{h_t}^t\}$$

Cas 2 $\min_{j\in B} d(X_t, G_j^{t-1}) \leq \rho_a$

Notons w l'indice de la classe pour laquelle le minimum ci-dessus est atteint :

$$d(X_t, G_w) = \min_{j\in B} d(X_t, G_j^{t-1})$$

Nous obtenons alors:

$$h_t = h_{t-1}$$

$$C_w^t = C_w^{t-1} \cup \{X_t\}$$

$$G_w^t = (n_w^{t-1}.G_w^{t-1} + n_{x_t}.X_t)\ \frac{1}{n_w^{t-1}+n_x}$$

$$n_w^t = n_w^{t-1} + n_{x_t}$$

$$C_o^t = C_o^{t-1}$$

$$P_t - C_w^t = P_{t-1} - C_w^{t-1}$$

Css 3 $\rho_s < \min_{j\in B} d(X_t, G_j) \leq \rho_s$

ou $\min_{j\in B} d(X_t, G_j) > \rho_s$ et $k = h_{t-1}$

$$h_t = h_{t-1}$$

$$C_j^t = C_j^{t-1} \quad \forall\ j \in B$$

$$G_j^t = G_j^{t-1} \quad \forall\ j \in B$$

$$n_j^t = n_j^{t-1} \quad \forall\ j \in B$$

$$C_o^t = C_o^{t-1} \cup \{X_t\}$$

$$P_t = P_{t-1}$$

4.4 Les options de l'algorithme de MSA

4.4.1. La matrice-input

L'utilisateur a le choix de classer les objets suivants:
- les observations de la matrice des données
- les variables de la matrice des données

- les projections des observations sur un espace factoriel
- les projections des variables sur un espace factoriel

4.4.2. Le nombre de classes

Le nombre de classes peut être choisi en se basant sur les résultats d'une analyse factorielle ou en analysant le dendrogramme obtenu par une classification hiérarchique.

4.4.3. Les distances

L'utilisateur a le choix entre les deux distances quadratiques suivantes :

- la distance euclidienne
- la distance du CHI2

4.4.4. Les seuils d'acceptation et de séparation (ρ_a et ρ_s)

Afin d'implanter une autre méthode rapide qui n'est qu'un cas particulier de la méthode de MSA, deux options sont proposées. Elles se distinguent par la condition que doivent vérifier les deux seuils, $\rho_a \leq \rho_s$. Le choix $\rho_a < \rho_s$ correspond à la méthode de MSA que nous venons de décrire. Par contre si $\rho_a = \rho_s$, nous obtenons une autre méthode rapide, appelée "quick cluster leader" par Hartigan /16/ et "leader" par Spaeth /22/. Dans cette méthode la règle d'attribution se résume de la façon suivante :

si $d_o \leq \rho_o = \rho_a = \rho_s$, l'objet est classé dans la classe la plus proche

si $d_o > \rho_o$ et $h < k$, une nouvelle classe est créée

si $d_o > \rho_o$ et $k = h$, l'objet est attribué à la classe des objets non-classables

Dans les deux cas, deux possibilités sont offertes pour définir les seuils:

a) Le seuil est calculé automatiquement par le programme. ρ_a correspond à la distance moyenne entre les 15 premiers objets de la matrice-input (choix arbitraire). Si l'utilisateur a choisi l'option $\rho_a = \rho_s$, ρ_s est égal à ρ_a. Par contre si $\rho_a < \rho_s$, on pose:

$$\rho_s = \rho_a + \min_{i,j} d(X_i, X_j), \quad (i,j \in \{1, 2, \ldots, 15\}).$$

Remarque:
Si la matrice-input contient moins de 15 objets à classer, le calcul de ρ_a et ρ_a se fait sur ses m objets.

b) Les seuils sont choisis par l'utilisateur.

4.4.5. Les résultats de MSA

L'algorithme génère comme pour KMEAN, une matrice des centres d'inertie associée à la partition. Pour une analyse ultérieure, l'utilisateur peut sélectionner les centres d'inertie associés soit à la partition partielle $P_m = \{C_1, C_2, \ldots, C_h\}$, soit à la partition complète (cf. paragraphe 4.2).

4.5 Exemples

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←↵→ N
MODULE CHOISI ? /E/P/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←↵→ B

*****************************************************
*MODULE DE PARTITIONNEMENT PAR LA METHODE DE MSA /B/*
*****************************************************

QUELLE DONNEE ? /OBS/VAR/POB/PVA/?/ ←↵→ ?

LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE UTILISEES POUR LE PARTITIONNEMENT:
  /POB/ : PROJECTIONS DES OBSERVATIONS
  /PVA/ : PROJECTIONS DES VARIABLES
  /OBS/ : OBSERVATIONS DE LA MATRICE DES DONNEES
  /VAR/ : VARIABLES DE LA MATRICE DES DONNEES

SEULES LES MATRICES CI-DESSOUS SONT DEJA DEFINIES:
 - LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
 - LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/


QUELLE DONNEE ? /OBS/VAR/POB/PVA/?/ ←↵→ OBS
QUELLE DISTANCE ? /DEU/DCH/ ←↵→ ?
```

```
CHOIX DES DISTANCES POUR L'ALGORITHME:
  /DEU/ : DISTANCE EUCLIDIENNE
  /DCH/ : DISTANCE DU CHI2

QUELLE DISTANCE ? /DEU/DCH/ ←⊞→ DCH
POIDS DES OBJETS (POUR PONDERER LES DISTANCES ENTRE OBJETS) ? /POI/POC/POU/
←⊞→ ?

LES POIDS DES OBJETS SUIVANTS PEUVENT ETRE UTILISES:
  /POI/ : POIDS IDENTIQUES TOUS EGAUX A 1
  /POC/ : POIDS EGAUX AUX MARGES
  /POU/ : POIDS CHOISIS PAR L'UTILISATEUR

POIDS DES OBJETS (POUR PONDERER LES DISTANCES ENTRE OBJETS) ? /POI/POC/POU/
←⊞→ POC
→ NOMBRE DE CLASSES ? ←⊞→ ?

LE NOMBRE DE CLASSES POUR LE PARTITIONNEMENT (MSA) DOIT ETRE COMPHIS ENTRE 1
ET 9.

→ NOMBRE DE CLASSES ? ←⊞→ 4
MEME SEUIL D'ACCEPTATION ET DE SEPARATION ?./O/N/ ←⊞→ ?

POSSIBILITES POUR LES SEUILS :
  /O/ : SEUIL D'ACCEPTATION = SEUIL DE SEPARATION
  /N/ : SEUIL D'ACCEPTATION ≠ SEUIL DE SEPARATION

MEME SEUIL D'ACCEPTATION ET DE SEPARATION ? /O/N/ ←⊞→ N
MODE DE DEFINITION DU (OU DES) SEUIL(S) ? /AUT/MAN/ANC/ ←⊞→ ?

POSSIBILITES POUR DEFINIR LES SEUILS :
  /AUT/ : CALCULES AUTOMATIQUEMENT PAR LE PROGRAMME
  /MAN/ : ENTRES AU CLAVIER PAR L'UTILISATEUR
  /ANC/ : MEMES SEUILS QUE LORS D'UN PARTITIONNEMENT ANTERIEUR,
          DISPONIBLES TANT QUE L'UTILISATEUR N'EST PAS SORTI DU MODULE /B/

MODE DE DEFINITION DU (OU DES) SEUIL(S) ? /AUT/MAN/ANC/ ←⊞→ AUT

PARAMETRES POUR LE PARTITIONNEMENT SELON MSA /B/ (27.12.1982 3H 24)
************************************************
OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
POIDS DES OBJETS EGAUX AUX MARGES
NOMBRE MAXIMUM DE CLASSES 4
DISTANCE DU CHI2
SEUIL D'ACCEPTATION = 0.2695175278
SEUIL DE SEPARATION = 0.2799734458

RESULTATS
---------
NOMBRE DE CLASSES CREEES    2
NOMBRE D'OBJETS NON-CLASSES 0
```

IMPRESSION DES CENTRES D'INERTIE ? /O/N/ ↵ O
DE QUELLE FACON ? /VAL/POU/ ↵ ?

/VAL/ : IMPRESSION DES CENTRES D'INERTIE EN VALEUR
/POU/ : IMPRESSION DES CENTRES D'INERTIE EN POURCENTAGE

DE QUELLE FACON ? /VAL/POU/ ↵ VAL

MATRICE DES CENTRES D'INERTIE

	1MARQUE	2MARQUE	3MARQUE	4MARQUE	5MARQUE	6MARQUE
1	2.633E2	5.393E2	2.350E1	1.037E2	1.041E3	3.699E2
2	3.400E1	2.010E2	3.300E1	2.300E2	8.920E2	7.010E2

	7MARQUE	8MARQUE	9MARQUE	10MARQUE	11MARQUE	POIDS
1	2.199E2	1.223E2	6.788E1	4.589E2	9.938E1	2.647E4
2	1.660E2	2.010E2	4.300E1	1.770E2	3.220E2	3.000E3

IMPRESSION DE LA COMPOSITION DES CLASSES ? /O/N/ ↵ O

IMPRESSION DE LA COMPOSITION DES CLASSES

CLASSE 1

1 *QUALITE* 2 *TECHNIQUE* 3 *PRECISION* 4 *ESTHETIQUE*

5 *FIABILITE* 6 *VALEUR* 7 *SOLIDITE* 8 *ELEGANT*

CLASSE 2

9 *ELECTRON*

QUEL BRANCHEMENT ? /DIS/NBC/SEU/CEI/CLA/AFF/ ↵ ?

LISTES DES BRANCHEMENTS POSSIBLES DANS LE MODULE DE PARTITIONNEMENT (MSA):
/DIS/ CHOIX D'UNE NOUVELLE DISTANCE POUR L'ALGORITHME (SUR LE MEME JEU DE DONNEES)
/NBC/ CHANGEMENT DU NOMBRE DE CLASSES (SUR LE MEME JEU DE DONNEES)
/SEU/ CHOIX DES SEUILS D'ACCEPTATION ET DE SEPARATION (SUR LE MEME JEU DE DONNEES)
/CEI/ IMPRESSION DES CENTRES D'INERTIE (DEJA CALCULES)
/CLA/ IMPRESSION DE LA CONSTITUTION DES CLASSES (DEJA CALCULEES)
/AFF/ CONSTITUTION DE NOUVELLES CLASSES AVEC LES ELEMENTS NON-CLASSES

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/R/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ↵ N

QUEL BRANCHEMENT ? /DIS/NBC/SEU/CEI/CLA/AFF/ ←⊞→ SEU
MEME SEUIL D'ACCEPTATION ET DE SEPARATION ? /O/N/ ←⊞→ NON
MODE DE DEFINITION DU (OU DES) SEUIL(S) ? /AUT/MAN/ANC/ ←⊞→ MAN
→ ENTREZ DANS L'ORDRE LES SEUILS SUIVANTS (CONDITION: SEUIL D'ACCEPTATION ≤ SEUIL DE SEPARATION)
ACCEPTATION SEPARATION
.25 .27

PARAMETRES POUR LE PARTITIONNEMENT SELON MSA /B/ (27.12.1982 3H 26)
****
OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
POIDS DES OBJETS EGAUX AUX MARGES
NOMBRE MAXIMUM DE CLASSES 2
DISTANCE DU CHI2
SEUIL D'ACCEPTATION = 0.25
SEUIL DE SEPARATION = 0.27

RESULTATS

NOMBRE DE CLASSES CREEES 2
NOMBRE D'OBJETS NON-CLASSES 1

IMPRESSION DES CENTRES D'INERTIE ? /O/N/ ←⊞→ N O

IMPRESSION DE LA COMPOSITION DES CLASSES

CLASSE 1

1	*QUALITE*	2	*TECHNIQUE*	3	*PRECISION*	4	*ESTHETIQUE*
5	*FIABILITE*	7	*SOLIDITE*	8	*ELEGANT*		

CLASSE 2

9 *ELECTRON*

LISTE DES ELEMENTS NON-CLASSES

6 *VALEUR*

FAIRE UNE CLASSE AVEC CHAQUE ELEMENT NON-CLASSE ? /O/N/ ←⊞→ ?

/O/ : ON FORMERA AVEC CHAQUE ELEMENT NON-CLASSE UNE NOUVELLE CLASSE
/N/ : ON NE TIENT PAS COMPTE DES ELEMENTS NON-CLASSES

FAIRE UNE CLASSE AVEC CHAQUE ELEMENT NON-CLASSE ? /O/N/ ←⊞→ N
QUEL BRANCHEMENT ? /DIS/NBC/SEU/CEI/CLA/APF/ ←⊞→ S O
FIN DE L'EXECUTION DU PACKAGE

5. INTRODUCTION AUX METHODES DE CLASSIFICATION HIERARCHIQUE

Les techniques de classification hiérarchique cherchent à structurer m objets en une hiérarchie de classes (objets isolés, sous-sous-classes, sous-classes, classes, classe principale) de telle façon que la similitude entre deux objets appartenant à une même sous-classe soit plus grande que celle entre objets d'une même classe. Une sous-classe est donc plus homogène que sa classe. Pour construire une hiérarchie de classes on peut procéder de deux façons différentes. On part des objets isolés qu'on regroupe peu à peu en des classes de plus en plus importantes jusqu'à ce qu'on aboutisse à une seule classe contenant tous les objets. On peut aussi partir de cette "grande" classe, la diviser en deux jusqu'à ce que toutes les classes ne contiennent encore qu'un seul objet. Ces deux procédures conduisent à deux méthodes de classification hiérarchique, la première correspond à la classification hiérarchique ascendante et la deuxième à la classification hiérarchique descendante. Le schéma ci-dessous illustre, par un exemple, les principes de ces deux méthodes.
(Remarque: les objets sont tous numérotés de façon à ce qu'on regroupe toujours des classes de numéro contiguës.)

nombre de classes	itération	classes	itération
m	0	$(X_1)(X_2)\ldots(X_{i-1})(X_i)(X_{i+1})\ldots\ldots(X_k)\ldots(X_m)$	m-1
m-1	1	$(X_1)(X_2)\ldots(X_{i-1},X_i)(X_{i+1})\ldots\ldots\ldots\ldots(X_m)$	m-2
m-2	2	$(X_1,X_2)\ldots(X_{i-1},X_i)(X_{i+1})\ldots\ldots\ldots\ldots(X_m)$	m-3
...	...	.....	...
3	m-3	$(X_1,X_2,\ldots,X_{i-1},X_i)(X_{i+1},\ldots,X_k)(X_{k+1},\ldots,X_m)$	2
2	m-2	$(X_1,X_2,\ldots,X_{i-1},X_i)(X_{i+1},\ldots\ldots\ldots\ldots,X_m)$	1
1	m-1	$(X_1,X_2,X_3,\ldots,X_{i-1},X_i,X_{i+1},\ldots\ldots\ldots\ldots,X_m)$	0

Légende: ↓ : méthode ascendante ↑ : méthode descendante

Remarque: (X_i) représente la classe formée de l'unique objet X_i.

Le nombre de hiérarchies de classes est évidemment beaucoup plus élevé que le nombre de partitions de m objets en k classes. Faisons le calcul pour une classification hiérarchique ascendante.

A l'itération j, nous avons $\binom{m-j}{2}$ possibilités de fusionner deux classes en une seule. Le nombre de hiérarchies pour m objets s'élève donc à:

$$H(m) = \prod_{j=1}^{m} \binom{j}{2} = [m!.(m-1)!]/2^{m-1}$$

Par exemple, le nombre de hiérarchies possibles pour 4 objets H(4), est égal à 18 et pour 20 objets H(20), vaut $5.64\ 10^{29}$.

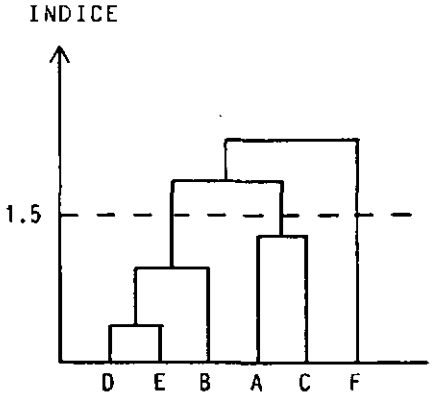
Toutes les méthodes de classification hiérarchique sont donc des heuristiques qui cherchent à optimiser à chaque itération un critère.

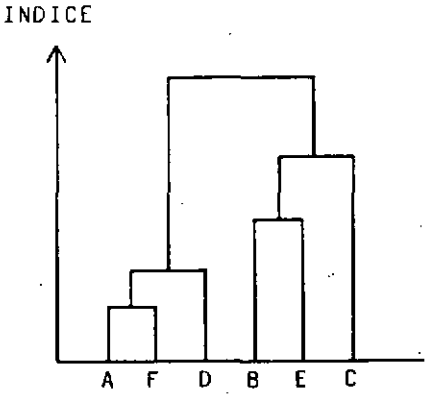
Les résultats d'une classification hiérarchique sont représentés sous la forme d'un dendrogramme, diagramme à deux dimensions qui illustre les fusions ou les divisions obtenues à chaque itération. Parfois, le dendrogramme est également appelé arbre.

Nous allons donner deux exemples de dendrogrammes. Notre but est de classer les objets A, B, C, D, E et F. Afin de bien faire comprendre le lien qui existe entre les fusions ou les divisions obtenues à chaque itération et le dendrogramme, nous allons donner tout d'abord un résumé des itérations, puis nous dessinerons le dendrogramme qui leur est associé.

Exemple 1
Classification hiérarchique ascendante
(D) (E) (B) (A) (C) (F)
(D E)(B)(A)(C)(F)
(D E B) (A) (C) (F)
(D E B) (A C) (F)
(D E B A C) (F)
(D E B A C F)

Exemple 2
Classification hiérarchique descendante
(A F D B E C)
(A F D) (B E C)
(A F D) (B E) (C)
(A F D) (B) (E) (C)
(A F) (D) (B) (E) (C)
(A) (F) (D) (B) (E) (C)





A chaque itération ou, ce qui revient au même, à chaque niveau de l'arbre, on associe un nombre, appelé son indice. Pour les méthodes ascendantes, cet indice est égal à la distance entre les deux classes qui ont été réunies, et pour les méthodes descendantes à la valeur d'un critère qui a été choisi pour sélectionner la meilleure partition. On dit qu'une hiérarchie est monotone si son indice est croissant. Autrement on dit qu'il y a une inversion.
En coupant le dendrogramme à un certain niveau, on obtient une division de l'ensemble des objets. Pour le dendrogramme de l'exemple 1, si nous le coupons au niveau 1.5 de la

valeur de son indice, nous obtenons la partition suivante : (D E B)) (A C) (F).

Puisqu'en classification hiérarchique ascendante les fusions sont basées sur des calculs de distances, la méthode est polythétique. Par contre, les méthodes descendantes peuvent être polythétiques ou monothétiques. CLASFAC contient deux méthodes descendantes, la première est une méthode polythétique (HIERKMEAN et POLYDIV) et la deuxième est une méthode monothétique, la segmentation. Cette méthode est particulière, car à l'encontre des autres méthodes de classification hiérarchique de CLASFAC, elle ne peut être exécutée que sur une matrice de données qualitatives comprenant un critère externe booléen.

Nous allons présenter dans l'ordre la méthode de classification hiérarchique ascendante, les méthodes de classification hiérarchiques descendantes polythétiques, puis la méthode de segmentation.

6. LA CLASSIFICATION HIERARCHIQUE ASCENDANTE, ABREGEE CHA

Avant de résumer l'algorithme de CHA, nous allons expliquer les principes de la méthode en développant les notions statistiques les plus importantes.

6.1. Présentation de la méthode

6.1.1. Présentation générale

Pareille à toutes les méthodes de classification hiérarchique, CHA recherche une hiérarchie de partitions emboîtées. Le point de départ de la méthode est une partition formée de m classes où chaque classe ne contient qu'un élément. Cette partition initiale sera notée $C^o = (C_1^o, C_2^o, \ldots, C_i^o, \ldots C_m^o)$. A chaque itération, le nombre de classes de la partition diminue d'une unité, car les deux classes qui minimisent un critère vont être réunies en une seule. Après m-1

itérations, la partition n'est plus formée que d'une seule classe qui contient tous les objets.
La méthode implantée est basée sur le calcul de distance entre les différentes classes.

6.1.2. Le critère optimisé à chaque itération

Si nous nous situons au début de la t-ième itération, la partition est $C_1^{t-1}, C_2^{t-1}, \ldots, C_{m-t+1}^{t-1}$.
Parmi les m-t+1 classes de la partition, les deux classes qui se trouvent à la plus petite distance l'une de l'autre (les classes les plus similaires) doivent être réunies. Mathématiquement cela revient à minimiser:

$$\min_{p \neq q,\ p,q \in \{1,2,\ldots,m-t+1\}} d(C_p^{t-1}, C_q^{t-1})$$

Notons i et j les indices des classes pour lesquelles ce minimum est atteint. Les classes C_i et C_j vont donc être réunies en une seule.

La nouvelle partition obtenue est donc:

$$C_p^t = C_p^{t-1} \quad p \neq i \text{ et } q \neq j$$

$$C_{i \cup j}^t = C_i^{t-1} \cup C_j^{t-1}$$

L'itération suivante ne peut être effectuée que si l'on sait calculer la distance entre la nouvelle classe et les autres classes. La définition de la notion de distance entre objets est donc essentielle.

6.1.3. La stratégie d'agrégation

6.1.3.1. Définition

En classification hiérarchique ascendante, la notion de stratégie d'agrégation (nous omettrons volontairement la stratégie basée sur la théorie de l'information, non implantée dans CLASFAC) coïncide avec la notion de distance

entre classes. Les différentes méthodes de la classification hiérarchique ne se distinguent que par le choix de la stratégie d'agrégation.
Selon la stratégie d'agrégation choisie, l'espace peut être déformé par l'opération d'agrégation. Une stratégie d'agrégation aura la propriété de dilater, de contracter ou de conserver l'espace.
Par la suite, nous noterons C_i, C_j et C_k, trois classes (sous-ensembles disjoints, non vides) quelconques.

Une distance est <u>contractante</u> si l'on a:

$$\forall\ i,j,k \quad d(C_i \cup C_j, C_k) \leq \min\ (d(C_i,C_k), d(C_j,C_k))$$

L'opération d'agrégation des classes C_i et C_j a donc pour effet de "rapprocher" la classe C_k. Une stratégie contractante aura tendance à intégrer les objets dans les classes déjà existantes plutôt que de former de nouvelles classes.

Une stratégie est <u>dilatante</u>, si on a:

$$\forall\ i,j,k \quad d(C_i \cup C_j, C_k) \geq \max\ (d(C_i,C_k), d(C_j,C_k))$$

Dans ce cas la classe C_k "s'éloigne" après l'opération d'agrégation des classes C_i et C_j. Une stratégie dilatante aura tendance à former beaucoup de classes artificiellement compactes.
Les stratégies pour lesquelles l'inégalité suivante est vérifiée sont dites <u>conservantes</u>:

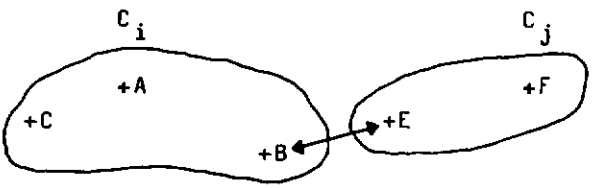
$$\forall\ i,j,k$$
$$\min\ (d(C_i,C_k), d(C_j,C_k)) < d(C_i \cup C_j, C_k) < \max\ (d(C_i,C_k), d(C_j,C_k))$$

Avant de poursuivre notre exposé, nous allons donner un exemple d'une des stratégies implantées dans CLASFAC, <u>le chaînage simple</u>; les autres stratégies seront expliquées lors de la présentation des options de CHA.

Example : Le chaînage simple
La stratégie du chaînage simple définit la distance entre deux classes C_i et C_j comme la distance minimale entre leurs objets respectifs. Elle peut être calculée comme suit:

$$d(C_i,C_j) = \min \{ d(X,Y) \mid X \in C_i,\ Y \in C_j \}$$

Par un schéma, nous allons illustrer cette stratégie.



La distance $d(C_i,C_j)$ entre les classes C_i et C_j correspond dans cet exemple à la longueur du segment BE.

6.1.3.2. L'algorithme combinatoire
L'algorithme de CHA serait d'une extrême complexité, si pour calculer la distance $d(C_k,C_i \cup C_j)$, on devait remonter aux objets de la matrice des données pour calculer la distance entre les classes $C_i \cup C_j$ et C_k. En 1967, Lance et Williams ont établi une formule de récurrence qui permet de calculer la distance $d(C_k,C_i \cup C_j)$ à partir des distances $d(C_i,C_j)$, $d(C_i,C_k)$ et $d(C_j,C_k)$ et ceci pour toutes les stratégies usuelles.
Cette formule de récurrence s'écrit:

$$d(C_k,C_i \cup C_j) = a_i.d(C_k,C_i) + a_j.d(C_k,C_j) + b.d(C_i,C_j)$$
$$+ g.|d(C_k,C_i) - d(C_k,C_j)|$$

Cette formule possède une très grande importance pour le calcul numérique, car il suffira, pour implanter différentes stratégies d'agrégation, de donner les valeurs de leurs coefficients dans la formule de récurrence.

Afin de bien cerner les étapes de l'algorithme, nous allons résumer l'algorithme sans entrer dans les détails informati-

quea.

6.2. Résumé de l'algorithme

But:

Former une hiérarchie de partitions emboîtées avec m objets, qui sont notés: X_1, X_2, ..., X_i, ..., X_m

Initialisation

k=m : nombre de classes de la partition

$C = \{C_1^o, C_2^o, \ldots, C_k^o\}$: partition initiale telle que $\forall i$, $C_i^o = \{X_i\}$

$\forall\ i,j\ \ d(C_i^o, C_j^o) = d(X_i, X_j)$: les distances entre les classes sont égales aux distances entre les uniques objets qu'elles contiennent.

La t-ième itération

Au début de la t-ième itération, la partition est :

$C_1^{t-1}, C_2^{t-1}, \ldots, C_k^{t-1}$ et le nombre de classes k=m-t+1
Quand aucune confusion n'est à craindre, nous omettrons dans la suite l'indice supérieur t-1.

Les étapes

1) Chercher :

$i \neq j$ tel que $d(C_i, C_j) = \min \{d(C_p, C_q) \mid p,q \in \{1,2,\ldots,k\}, p \neq q\}$

2) Mémoriser $d(C_i, C_j)$: indice de l'itération pour le dendrogramme

3) Agréger C_i et C_j et former $C_i \cup C_j$

4) Définir une nouvelle partition :
(($C_p^t = C_p^{t-1}$, $p \neq i$ et $p \neq j$), $C_i \cup C_j$)

5) Calculer les distances en tenant compte de la stratégie d'agrégation choisie:

$d(C_p^t, C_i \cup C_j)$ $\forall$ $p \neq i$, $p \neq j$

6) Ajuster le nombre de classes : k=(m-t+1)-1=m-t

Arrêt de l'algorithme

L'algorithme s'arrête après m-1 itérations, c'est-à-dire quand k=1 et $C=\{X_1, X_2, \ldots, X_i, \ldots, X_m\}$.

6.3. Les options de CHA

6.3.1. La matrice-input
L'utilisateur a le choix de classer les objets suivants:
- les observations de la matrice des données
- les variables de la matrice des données
- les projections des observations
- les projections des variables
- les centres d'inertie des observations de la matrice des données
- les centres d'inertie des variables de la matrice des données

6.3.2. Les distances entre objets
L'utilisateur a le choix entre les distances suivantes:
- la distance euclidienne
- la distance de CM12
- la distance de Minkowski d'ordre 1
- la distance de Canberra
- la distance de Minkowski d'ordre infini

6.3.3. Les stratégies d'agrégation
Nous allons présenter tout d'abord les différentes stratégies en soulignant leur effet sur la hiérarchie finale, puis

nous donnerons les coefficients de la formule de récurrence pour chacune d'elles sous forme d'un tableau. Dans la littérature, il existe une multitude de synonymes pour désigner chaque stratégie, nous avons uniquement retenu le nom le plus usité.

6.3.3.1. Le chaînage simple ("single linkage")

Seulement l'effet de la stratégie du chaînage simple sur le résultat final va être décrit, car sa définition a déjà été donnée à titre d'exemple au paragraphe 6.1. Par son nom, cette stratégie évoque l'effet qu'elle provoque sur la hiérarchie: elle tend à regrouper comme les maillons d'une chaîne les objets dans des groupes déjà existants. En d'autres termes, elle a donc tendance à créer des classes très allongées.
Remarquons qu'elle est une des rares méthodes qui ne forme pas des classes ellipsoïdales.
Cette stratégie a la propriété de laisser invariante la hiérarchie créée par rapport à une transformation monotone de la distance entre objets.

6.3.3.2. Le chaînage complet ("complete linkage")

Pour la stratégie du chaînage complet la distance entre deux classes C_i et C_j est définie comme étant la distance maximale entre les objets des deux classes. Elle peut s'écrire:

$$d(C_i,C_j) = \max \{d(X,Y) \mid X \in C_i,\ Y \in C_j\}$$

Si l'on se réfère à l'exemple du paragraphe 6.1, la distance $d(C_i,C_j)$ correspond à la longueur du segment CF.
Cette stratégie tend à former des groupes très compacts et à créer des classes "artificielles". Elle possède également la propriété que la hiérarchie créée est invariante par rapport à une transformation monotone de la distance entre objets.

6.3.3.3. Ls distsnce moysnne ("group sversge")

Cette strstégie eat un compromis entre ls strstégie du chsînsge simple et du chaînags complet. Ls distance entre deux classes C_i, C_j est définie comme ls moyenne des distsnces entre les objets des deux classes :

$$d(C_i,C_j) = \frac{1}{n_j+n_i} \sum_{\substack{X \in C_i \\ Y \in C_j}} n_x \, n_y \, d(X,Y)$$

Les objets d'une même clsase sont donc proches "en moyenne".

6.3.3.4. Les centres de gravité ("centroid method")

Ls distsnce entre les clsases C_i et C_j est définie par la distance entre leurs centres de grsvité G(i) et G(j) :

$$d(C_i,C_j) = d(G(i),G(j))$$

Cette méthode ne peut être utilisée que si la distance entre objets est une distsnce quadrstique. Rsmarquons que si les classes ne sont pss convexes, ls méthode des centres de gravité eat une mauvaise meaure de la diasimilarité. Cette méthode peut psrfois conduire à des inversions dans le dendrogramme. Elle n'eat donc pss monotone.

6.3.3.5. Ls méthode de Ward ou ls méthode de ls varisnce inter-classes ("Ward method")

Ce n'est que moyennant un artifice que ls strstégie de Ward peut être réduite au calcul de ls distsnce entre classes. Conaidérons laa deux partitions:

$$P_o = (C_1, C_2, \ldots, C_i, \ldots, C_j, \ldots, C_k, \ldots)$$

$$P_1 = \{C_1, C_2, \ldots, C_i \cup C_j, \ldots\ldots\ldots, C_k, \ldots\}$$

La partition P_1 est donc obtenue à partir de la partition P_o en réunissant les classes C_i et C_j en une seule.
Notons $V^2_{\text{inter-classes}}$ la variation totale entre les classes associée à une partition (cf. paragraphe 3.2).

$$V^2_{\text{inter-classes}} = \sum_{i=1}^{k} n_i \; d^2(G(i),G_o)$$

où G_o désigne le centre d'inertie de tous les objets.
Définissons maintenant la distance entre les classes C_i et C_j par:

$$d(C_i,C_j) = V^2_{\text{inter-classes}}(P_o) - V^2_{\text{inter-classes}}(P_1)$$

La distance entre les classes C_i et C_j est donc égale à la variation de la variation inter-classes quand on a réduit ces deux classes en une seule.
Un calcul simple montre que:

$$d(C_i,C_j) = \frac{n_i \cdot n_j}{n_i+n_j} \; d(G(i),G(j))$$

Comme on réunit à chaque itération les deux classes les plus proches, la méthode de Ward maximise donc séquentiellement la variation totale entre les classes ou, ce qui revient au même, la variance inter-classes.

6.3.3.6. La méthode flexible ("flexible method")

Cette méthode proposée par Lance et Williams permet à l'utilisateur de choisir la valeur du paramètre b dans la formule de récurrence. Selon la valeur attribuée à b, l'espace est soit conservé, dilaté ou contracté (cf. la figure 2.4). Usuellement on utilise b=-0.25

Les valeurs des coefficients dans la formule de récurrence associées à chaque stratégie d'agrégation

Rappelons la formule de récurrence : avec $k=i \cup j$

$$d(C_h,C_k) = a_i.d(C_h,C_i)+a_j.d(C_h,C_j)+b.d(C_i,C_j) + g.|d(C_h,C_i)-d(C_h,C_j)|$$

stratégie	a_i	a_j	b	g	monotone	déformation de l'espace
chaînage simple	0.5	0.5	0	-.5	oui	fortement contractante
chaînage complet	0.5	0.5	0	0.5	oui	fortement dilatante
distance moyenne	n_i/n_k	n_j/n_k	0	0	oui	conservante
*centre de gravité	n_i/n_k	n_j/n_k	$-a_i.a_j$	0	non	conservante
*méthode de Ward	$(n_h+n_i)/(n_h+n_k)$	$(n_h+n_j)/(n_j+n_k)$	$-n_h/(n_h+n_k)$	0	oui	dilatante
méthode flexible	Condition: $a_i+a_j+b=1$, $a_i=a_j$, $b<1$, $g=0$				oui	b>0: contractante b=0 conservante b<0 dilatante

Figure 2.4

Légende: * : la stratégie ne doit être utilisée qu'avec une distance quadratique.

Remarque:
Les poids des objets n'interviennent que dans les stratégies de la distance moyenne, de la méthode des centres de gravité et de la méthode de Ward.

6.3.4. L'impression des agrégations successives
Afin de faciliter la lecture du dendrogramme, CLASFAC imprime les agrégations successives de l'algorithme. Pour comprendre les résultats qui sont imprimés pour chaque itération, un point de l'implantation informatique doit être précisé. Lors de l'agrégation de deux classes, C_i et C_j, le programme doit supprimer une classe et remplacer l'autre par la réunion des deux classes $C_i \cup C_j$. Le choix effectué est le suivant: la classe contenant le plus petit numéro d'ordre contiendra la réunion des deux classes. Cette classe portera le nom d'AINE, et la classe qui va être supprimée s'appellera BENJAMIN. Dans la suite de l'algorithme, la classe AINE sera le représentant de la classe effective $C_i \cup C_j$.
CLASFAC n'imprime pas le numéro d'ordre i ou j, mais les noms des objets associés à chaque ligne de la matrice-input.

6.4. Remarques concernant CHA

Même si la méthode optimise un critère à chaque itération, rien ne prouve que la hiérarchie obtenue soit la meilleure. De plus par sa construction itérative, une erreur commise lors de la fusion de deux classes se transmettra jusqu'à la fin de la construction de la hiérarchie.
Pour dégager "les formes fortes" (les objets qui restent attribués aux mêmes classes), l'utilisateur doit exécuter plusieurs fois CHA en choisissant différentes stratégies.

6.5. Exemples

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ N
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ H
```

```
*******************************************************
*MODULE DE CLASSIFICATION HIERARCHIQUE ASCENDANTE /H/*
*******************************************************

QUELLE DONNEE ? /OBS/VAR/POB/PVA/IVO/IVV/?/ <⊞> ?

LISTE DE TOUTES LES MATRICES DONT LES OBJETS PEUVENT ETRE CLASSES:
  /OBS/ : LES OBSERVATIONS DE LA MATRICE DES DONNEES
  /VAR/ : LES VARIABLES DE LA MATRICE DES DONNEES
  /POB/ : LES PROJECTIONS DES OBSERVATIONS
  /PVA/ : LES PROJECTIONS DES VARIABLES
  /IVO/ : LES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /IVV/ : LES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES

SEULES LES MATRICES CI-DESSOUS SONT DEFINIES:
 - LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
 - LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
 - LA MATRICE DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES /I
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/


QUELLE DONNEE ? /OBS/VAR/POB/PVA/IVO/IVV/?/ <⊞> OBS
DISTANCE CHOISIE ? /DEU/DMI/DM1/DCN/DCH/?/ <⊞> ?

LES DISTANCES SUIVANTES PEUVENT ETRE UTILISEES:
  /DEU/ : DISTANCE EUCLIDIENNE
  /DMI/ : DISTANCE DE MINKOWSKI D'ORDRE INFINI
  /DM1/ : DISTANCE DE MINKOWSKI D'ORDRE 1
  /DNC/ : DISTANCE DE CANBERRA
  /DCH/ : DISTANCE DU CHI-DEUX

DISTANCE CHOISIE ? /DEU/DMI/DM1/DCN/DCH/?/ <⊞> DCH
STRATEGIE ? /SCS/SCC/SDM/SCG/SWA/SPL/?/ <⊞> ?

LES STRATEGIES SUIVANTES PEUVENT ETRE UTILISEES:
  /SCS/ : CHAINAGE SIMPLE
  /SCC/ : CHAINAGE COMPLET
  /SDM/ : DISTANCE MOYENNE
  /SCG/ : CENTRE DE GRAVITE
  /SWA/ : METHODE DE WARD
  /SPL/ : METHODE FLEXIBLE

STRATEGIE ? /SCS/SCC/SDM/SCG/SWA/SPL/?/ <⊞> SDM
POIDS DES OBJETS ? (INTERVIENT DANS LE CALCUL DE LA DISTANCE ENTRE CLASSES)
/POI/POC/POU/?/ <⊞> ?
```

LES POIDS DES OBJETS SUIVANTS PEUVENT ETRE UTILISES:
/POI/ : POIDS IDENTIQUES TOUS EGAUX A 1
/POC/ : POIDS EGAUX AUX MARGES
/POU/ : POIDS CHOISIS PAR L'UTILISATEUR

POIDS DES OBJETS ? (INTERVIENT DANS LE CALCUL DE LA DISTANCE ENTRE CLASSES)
/POI/POC/POU/?/ ↵ *POC*

PARAMETRES POUR LA CLASSIFICATION HIERARCHIQUE ASCENDANTE /H/ (27.12.1982 3H 28)
****
OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
L'UTILISATEUR A FAIT LES CHOIX SUIVANTS
- *DISTANCE DU CHI-DEUX*
- *POIDS DES OBJETS EGAUX AUX MARGES*
- *POIDS DES CARACTERISTIQUES EGAUX A 1*
- *STRATEGIE DISTANCE MOYENNE*

IMPRESSION DE LA MATRICE DES DISTANCES ? /O/N/ ↵ *O*

MATRICE DES DISTANCES

	QUALITE	*TECHNIQUE*	*PRECISION*	*ESTNETIQUE*	*FIABILITE*	*VALEUR*
QUAL	0.0000*E*0	1.4219*E*-1	3.4643*E*-2	9.8808*E*-2	4.7356*E*-2	3.5040*E*-1
TECH	1.4219*E*-1	0.0000*E*0	8.1455*E*-2	2.9996*E*-1	1.1269*E*-1	3.1028*E*-1
PREC	3.4643*E*-2	8.1455*E*-2	0.0000*E*0	1.7258*E*-1	1.7502*E*-2	3.0172*E*-1
ESTH	9.8808*E*-2	2.9996*E*-1	1.7258*E*-1	0.0000*E*0	1.5023*E*-1	3.6697*E*-1
FIAB	4.7356*E*-2	1.1269*E*-1	1.7502*E*-2	1.5023*E*-1	0.0000*E*0	2.0670*E*-1
VALE	3.5040*E*-1	3.1028*E*-1	3.0172*E*-1	3.6697*E*-1	2.0670*E*-1	0.0000*E*0
SOLI	4.5590*E*-2	1.5329*E*-1	2.4020*E*-2	1.5807*E*-1	1.0456*E*-2	2.2507*E*-1
ELEG	1.4720*E*-1	4.1183*E*-1	2.2188*E*-1	2.0090*E*-2	1.8361*E*-1	4.1311*E*-1
ELEC	6.8997*E*-1	2.9243*E*-1	5.5676*E*-1	8.4532*E*-1	5.2607*E*-1	2.9151*E*-1

	SOLIDITE	*ELEGANT*	*ELECTRON*
QUAL	4.5590*E*-2	1.4720*E*-1	6.8997*E*-1
TECH	1.5329*E*-1	4.1183*E*-1	2.9243*E*-1
PREC	2.4020*E*-2	2.2188*E*-1	5.5676*E*-1
ESTH	1.5807*E*-1	2.0090*E*-2	8.4532*E*-1
FIAB	1.0456*E*-2	1.8361*E*-1	5.2607*E*-1
VALE	2.2507*E*-1	4.1311*E*-1	2.9151*E*-1
SOLI	0.0000*E*0	1.8606*E*-1	6.0648*E*-1
ELEG	1.8606*E*-1	0.0000*E*0	1.0003*E*0
ELEC	6.0648*E*-1	1.0003*E*0	0.0000*E*0

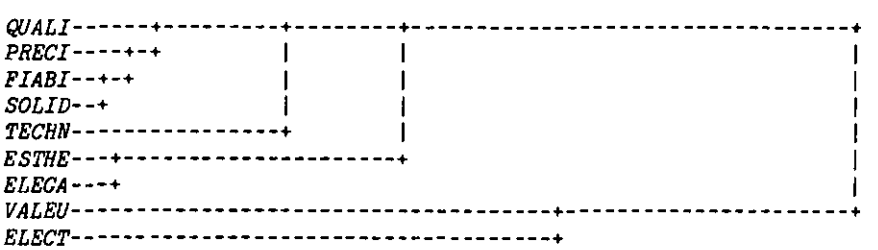
IMPRESSION DES AGREGATIONS SUCCESSIVES ? /O/N/ ↵ *O*

AGREGATIONS SUCCESSIVES

ITERATION	DIST.	AINE	BENJAMIN	POIDS AINE	POIDS BENJAMIN
1	1.0456E-2	FIABILITE	SOLIDITE	1.1076E-1	1.0828E-1
2	2.0090E-2	ESTRETIQUE	ELEGANT	1.1066E-1	1.0391E-1
3	2.0724E-2	PRECISION	FIABILITE	1.1147E-1	2.1904E-1
4	4.2490E-2	QUALITE	PRECISION	1.5057E-1	3.3052E-1
5	1.2382E-1	QUALITE	TECHNIQUE	4.8108E-1	1.0676E-1
6	1.9585E-1	QUALITE	ESTHETIQUE	5.8784E-1	2.1456E-1
7	2.9151E-1	VALEUR	ELECTRON	9.5796E-2	1.0180E-1
8	4.8421E-1	QUALITE	VALEUR	8.0240E-1	1.9760E-1

IMPRESSION DE L'ARBRE ? /O/N/ ↵ O

<--------> *VAUT* 8.0701E-2 *INDICE=DISTANCE*



QUEL BRANCHEMENT ? /STR/IAC/IAR/?/ ↵ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE DE CLASSIFICATION HIERARCHIQUE ASCENDANTE:
/STR/ : DEFINITION D'UNE NOUVELLE STRATEGIE
/IAC/ : IMPRESSION DES AGREGATIONS SUCCESSIVES (MEME JEU DE DONNEES)
/IAR/ : IMPRESSION DE L'ARBRE (MEME JEU DE DONNEES)

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ↵ N

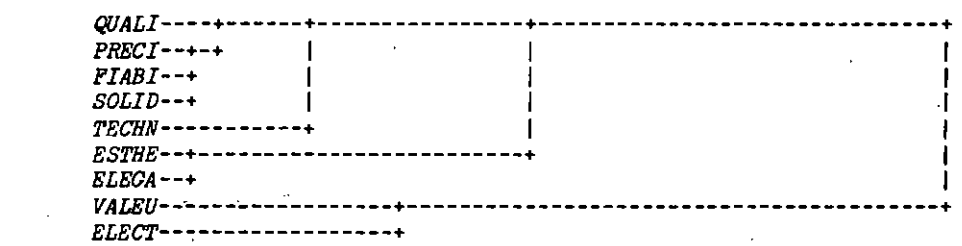
QUEL BRANCHEMENT ? /STR/IAC/IAR/?/ ↵ STR SWA
POIDS DES OBJETS ? (INTERVIENT DANS LE CALCUL DE LA DISTANCE ENTRE CLASSES)
/POI/POC/POU/?/ ↵ POC N O O

PARAMETRES POUR LA CLASSIFICATION HIERARCHIQUE ASCENDANTE /H/ (27.12.1982 3H 30)
****
OBJETS : LES OBSERVATIONS DE LA MATRICE DES DONNEES
L'UTILISATEUR A FAIT LES CHOIX SUIVANTS
- *DISTANCE DU CHI-DEUX*
- *POIDS DES OBJETS EGAUX AUX MARGES*
- *POIDS DES CARACTERISTIQUES EGAUX A 1*
- *STRATEGIE METHODE DE WARD*

AGREGATIONS SUCCESSIVES

ITERATION	*DIST.*	*AINE*	*BENJAMIN*	*POIDS AINE*	*POIDS BENJAMIN*
1	1.0456E⁻2	*FIABILITE*	*SOLIDITE*	1.1076E⁻1	1.0828E⁻1
2	2.0090E⁻2	*ESTHETIQUE*	*ELEGANT*	1.1066E⁻1	1.0391E⁻1
3	2.4212E⁻2	*PRECISION*	*FIABILITE*	1.1147E⁻1	2.1904E⁻1
4	5.8274E⁻2	*QUALITE*	*PRECISION*	1.5057E⁻1	3.3052E⁻1
5	1.7338E⁻1	*QUALITE*	*TECHNIQUE*	4.8108E⁻1	1.0676E⁻1
6	2.9151E⁻1	*VALEUR*	*ELECTRON*	9.5796E⁻2	1.0180E⁻1
7	4.7256E⁻1	*QUALITE*	*ESTHETIQUE*	5.8784E⁻1	2.1456E⁻1
8	1.0532E0	*QUALITE*	*VALEUR*	8.0240E⁻1	1.9760E⁻1

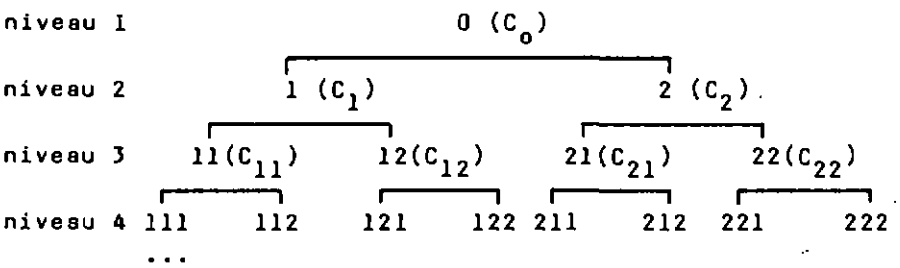
<--------> *VAUT* 1.7553E⁻1 *INDICE=DISTANCE*



QUEL BRANCHEMENT ? /STR/IAG/IAR/?/ S 0
FIN DE L'EXECUTION DU PACKAGE

7. INTRODUCTION A LA CLASSIFICATION HIERARCHIQUE DESCENDANTE, ABREGEE CHD

A l'inverse des méthodes de classification hiérarchique ascendante, on propose, dans les méthodes descendantes, d'arrêter l'algorithme avant la construction complète de la hiérarchie de classes afin de raccourcir le temps de calcul. Dès lors, une terminologie doit être adoptée, afin d'indiquer à quel niveau la subdivision des classes doit être interrompue.
L'arbre binaire suivant traduit les notations que nous avons adoptées.



Les numéros 0, 1, 2, 11, 12, 21, 22, ... désignent les noeuds de l'arbre, ils correspondent aux classes créées par la méthode de classification hiérarchique descendante.
La racine de l'arbre, le noeud 0 correspond à la classe $C_0=\{X_1, X_2, \ldots, X_m\}$, et se situe au niveau 1 de l'arbre.
Remarquons que les noms des noeuds sont toujours construits à partir de celui de leur prédécesseur. Par exemple, les noeuds 2121 et 2122 possèdant le prédécesseur 212. Le nombre de paliers de cet arbre est égal à 3. Le lecteur remarquera qu'on n'a pas compté la hauteur, mais le nombre de paliers de l'arbre.
Cette terminologie est utilisée dans CLASFAC à plusieurs reprises, aussi bien dans le dialogue que lors de l'impression des résultats. Ainsi l'utilisateur peut-il demander l'arrêt de la classification après un nombre n de paliers.
Par la suite, il est possible de poursuivre la classification en un noeud non encore exploré, soit par exemple au

noeud 211 de la figure ci-dessus.

8. LA CLASSIFICATION HIERARCHIQUE DESCENDANTE PAR UNE METHODE POLYTHETIQUE (HIERKMEAN, POLYDIV)

8.1. Présentation générale

Comme toutes les méthodes de classification hiérarchique, cette technique cherche une hiérarchie de partitions emboîtées. Le point de départ de la méthode correspond à la partition formée d'une seule classe qui contient les m objets : $C_o = \{X_1, X_2, \ldots, X_m\}$. A chaque itération le nombre de classes de la partition augmente d'une unité, car une classe est subdivisée en deux sous-classes plus homogènes. Après m-1 itérations, la partition est formée de m classes qui ne contiennent qu'un seul élément, $C_i = \{X_i\}$.
Souvent, on interrompt l'algorithme prématurément, c'est-à-dire avant l'exécution des m-1 itérations.

8.2. Le critère de division d'une classe en deux sous-classes

Les différentes méthodes de classification hiérarchique descendante se distinguent par la façon de diviser une classe en deux sous-classes. CLASFAC contient deux méthodes. La première, appelée HIERKMEAN subdivise une classe en deux sous-classes en utilisant l'algorithme de partitionnement de KMEAN, avec un nombre de classes fixé à deux. La deuxième, appelée POLYDIV consiste à projeter les objets d'une classe dans l'espace factoriel en utilisant la méthode de l'analyse factorielle en composantes principales, puis de les séparer en deux sous-classes selon leurs projections sur le premier axe factoriel. Il existe évidemment plusieurs variantes pour subdiviser les projections des objets sur le premier axe factoriel. Elles seront décrites lors de l'exposé des options de POLYDIV.
Les principes de la méthode de partitionnement KMEAN et de l'analyse factorielle en composantes principales ont déjà

été présentés en détail et le lecteur peut donc se référer aux paragraphes correspondants.

8.3. Réaumé des étapes de l'algorithme (HIERKMEAN ET POLYDIV)

On va décrire les étapes de l'algorithme sans préciser la façon de subdiviser une classe en deux sous-classes.

Initialisation

$C_o = \{X_1, X_2, \ldots, X_m\}$: partition initiale

LISTE = $\{C_o\}$, liste des noms des classes déjà créées et qui doivent être subdivisées en deux sous-classes

Itération

L'algorithme va, pour chaque classe contenue dans LISTE, subdiviser la classe en deux sous-classes.
Nous allons décrire une itération en nous basant sur un exemple. Admettons qu'au début de l'itération nous disposons de:

$C = \{C_1, C_2\}$: la partition

LISTE= $\{C_1, C_2\}$

$C_1 = \{X_a, X_b, X_c, X_y, X_z\}$

La classe C_1 est la première classe qui doit être subdivisée en deux sous-classes. En appliquant le critère utilisé par la méthode HIERKMEAN ou POLYDIV, on obtient, par exemple, les deux sous-classes $C_{11}= \{X_a, X_y, X_z\}$ et $C_{12}= \{X_b, X_c\}$.
Avant d'exécuter l'itération suivante qui consiste à subdiviser la classe C_2, la liste des noms des classes qui doivent encore être subdivisées doit être mise à jour. L'élément C_1 doit être supprimé de la liste et si

les conditions d'arrêt de l'algorithme ne sont pas vérifiées, les deux classes C_{11} et C_{12} doivent y être ajoutées. Dans notre exemple LISTE est formé après cette mise à jour de C_2, C_{11}, C_{12}.
A la fin de cette itération, la partition obtenue est : $C = [C_{11}, C_{12}, C_2]$.

Arrêt de l'algorithme

L'algorithme s'arrête lorsque chaque classe est formée d'un seul objet, $C_i = [X_i]$, ou lorsque les conditions d'arrêt prématuré sont vérifiées (cf. paragraphe 8.4)

8.4. Les options de l'algorithme de HIERKMEAN et POLYDIV

8.4.1. La matrice-input

L'utilisateur a le choix de classer les objets suivants:
- les observations de la matrice des données
- les variables de la matrice des données
- les projections des observations
- les projections des variables
- les centres d'inertie des observations de la matrice des données
- les centres d'inertie des variables de la matrice des données

8.4.2. La subdivision d'une classe en deux sous-classes

8.4.2.1. La méthode de KMEAN

La subdivision d'une classe par la méthode de partitionnement de KMEAN se fait automatiquement par le programme; elle ne nécessite donc aucune intervention de la part de l'utilisateur. Les options choisies pour l'algorithme sont les suivantes:
- nombre de classes fixé à 2
- distance euclidienne
- partition initiale selon la méthode de Mac Queen
- poids des objets égaux à 1

8.4.2.2. La méthode de POLYDIV

Avant de présenter les trois variantes qui permettront de séparer les projections des objets d'une classe sur le premier facteur, nous allons introduire les notations suivantes:

C : classe à subdiviser

F : les projections des objets de la classe C sur le premier axe factoriel

La division selon la variance inter-classes (méthode de DALE)

Cette méthode consiste à sélectionner parmi les subdivisions possibles de la classe C, les deux classes C_1 et C_2 qui maximisent la variance inter-classes. Mathématiquement, le critère à maximiser est le suivant:

$$\frac{n_1 \cdot n_2}{n_1+n_2} \left(\bar{F}_{C_1} - \bar{F}_{C_2}\right)^2$$

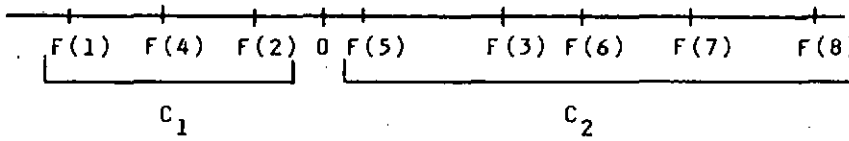
n_1, n_2 nombre d'éléments de la classe C_1 resp. C_2

avec $\bar{F}_{C_1} = \sum_{i \in C_1} F(i)$ et $\bar{F}_{C_2} = \sum_{i \in C_2} F(i)$

Cette méthode de division est appelée également méthode de Dale, du nom du chercheur qui a trouvé une solution informatique à ce problème. Son algorithme trouve la meilleure partition parmi les $2^{Card(C)-1}-1$ partitions possibles en Card(C)-1 itérations.

La division selon le centre de gravité

La méthode consiste à former deux sous-classes en séparant les projections des objets par rapport au centre de gravité de l'axe factoriel. L'exemple suivant illustre cette division :

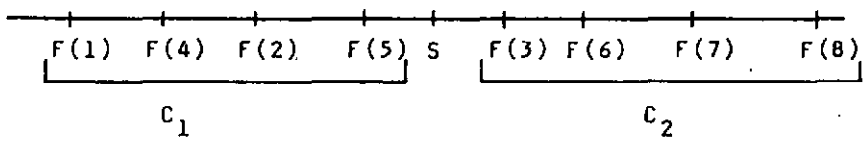


La règle d'affectation de l'objet X_i à l'une des classes C_1 ou C_2 peut être résumée de la façon suivante:

si $F(i) \leq 0$ alors mettre X_i dans C_1
si $F(i) > 0$ alors mettre X_i dans C_2

<u>La division qui affecte le même nombre d'objets à chaque sous-classe</u>

La division est obtenue en cherchant sur l'axe factoriel un point de séparation S qui sépare les éléments en deux sous-classes de même taille. Si le nombre d'objets de la classe à diviser est impair, une des deux sous-classes contiendra un élément de plus que l'autre. Si nous reprenons l'exemple ci-dessus le point S se situe entre les objets X_5 et X_3.



Le point X_i sera affecté à une sous-classe en tenant compte de la règle suivante:

si $F(i) \leq S$ alors mettre X_i dans C_1
si $F(i) > S$ alors mettre X_i dans C_2

8.4.3. <u>Les tests d'arrêt</u>

Si l'utilisateur s'intéresse à la construction d'une hiérarchie complète de classes, il n'imposera aucune condition d'arrêt. Par contre, si c'est seulement le début de la hiérarchie qui l'intéresse, il pourra faire cesser l'exploration d'un noeud, ou autrement dit la poursuite de la

division d'une classe, en imposant l'une ou l'autre des deux conditions suivantes:

- la classe contient un effectif inférieur ou égal à la valeur minimale choisie par l'utilisateur
- le niveau de la classe (dans l'arbre) dépasse une valeur maximale fixée par l'utilisateur.

8.4.4. Poursuite de la classification en un noeud non encore exploré

Si l'utilisateur a imposé l'arrêt de la classification après un nombre donné de niveaux, il peut reprendre celle-ci en un noeud non encore exploré.

8.4.5. Les résultats

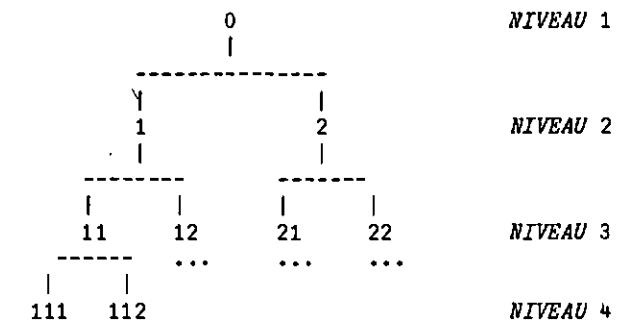
La première représentation des résultats est donnée sous forme d'identation. Elle permet de donner pour chaque classe toutes ses caractéristiques : nombre d'objets, variance de la classe, variance intra-classes, liste des objets qui la constituent.
La deuxième représentation est donnée sous forme de dendrogramme. L'indice de l'arbre, choisi par l'utilisateur, est soit la variance intra-classes, soit la variance de la classe.

8.5. Remarques concernant la méthode

Comme pour la méthode de CHA, une erreur commise lors de la division d'une classe se transmettra jusqu'à la fin de la construction de la hiérarchie.
Parfois, afin de connaître la stabilité des résultats obtenus par la méthode de CHA, il est utile de les comparer à ceux obtenus par la méthode HIERKMEAN ou POLYDIV et inversément.

8.6. Exemples

```
GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←┤├→ N
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←┤├→ X

**********************************************************************
*MODULE DE CLASSIFICATION HIERARCHIQUE DESCENDANTE POLYTHETIQUE /X/*
**********************************************************************

QUELLE DONNEE ? /OBS/VAR/POB/PVA/IVO/IVV/ ←┤├→ ?

LISTE DE TOUTES LES MATRICES DONT LES OBJETS PEUVENT ETRE CLASSES:
  /OBS/ : LES OBSERVATIONS DE LA MATRICE DES DONNEES
  /VAR/ : LES VARIABLES DE LA MATRICE DES DONNEES
  /POB/ : LES PROJECTIONS DES OBSERVATIONS
  /PVA/ : LES PROJECTIONS DES VARIABLES
  /IVO/ : LES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /IVV/ : LES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES

SEULES LES MATRICES CI-DESSOUS SONT DEFINIES:
 - LES OBSERVATIONS DE LA MATRICE DES DONNEES /OBS/
 - LES VARIABLES DE LA MATRICE DES DONNEES /VAR/
 - LA MATRICE DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES /IVO/
 - LA MATRICE DES COMPOSANTES DES OBSERVATIONS /POB/
 - LA MATRICE DES COMPOSANTES DES VARIABLES /PVA/


QUELLE DONNEE ? /OBS/VAR/POB/PVA/IVO/IVV/ ←┤├→ OBS
METHODE ? /KME/DAL/GRA/NOM/ ←┤├→ ?

LES ALGORITHMES DE CLASSIFICATION HIERARCHIQUE DESCENDANTE IMPLANTES
SONT LES SUIVANTS :
 - POLYDIV: CET ALGORITHME DIVISE UNE CLASSE EN DEUX SOUS-CLASSES EN CALCULANT
            LES PROJECTIONS DES OBJETS SUR LE PREMIER AXE FACTORIEL. PLUSIEURS
            METHODES EXISTENT POUR DEFINIR LES DEUX SOUS-CLASSES; ELLES SE
            DIFFERENCIENT PAR LA FACON DE COUPER LES PROJECTIONS DES OBJETS
            SUR L'AXE FACTORIEL.
            - /DAL/ : DIVISION SELON LA VARIANCE (METHODE DE DALE)
            - /GRA/ : DIVISION SELON LE CENTRE DE GRAVITE
            - /NOM/ : DIVISION AVEC UN MEME NOMBRE D'OBJETS DANS CHAQUE
                      SOUS-CLASSE
 - HIERKMEAN: - /KME/ : CHAQUE CLASSE EST DECOUPEE EN DEUX SOUS-CLASSES
                        SELON L'ALGORITHME DE PARTITIONNEMENT KMEAN

LES RESULTATS DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE SONT
RESUMES PAR UN ARBRE.
DANS CE MODULE LA TERMINOLOGIE EST LA SUIVANTE:
```



LES NUMEROS 0,1,2,11,12,21,22,111,112,... DESIGNENT LES NOEUDS DE L'ARBRE.
LES NIVEAUX CORRESPONDANT AUX NOEUDS SONT 1,2,3,4.
LE NOMBRE DE PALIERS DE CET ARBRE EST DE 3.

METHODE ? /KME/DAL/GRA/NOM/ ←⊞→ KME
FORME DES DONNEES ? /BRU/CEN/CRE/ ←⊞→ ?

FORMES POSSIBLES POUR LA MATRICE DES DONNEES:
/BRU/ : DONNEES BRUTES
/CEN/ : DONNEES CENTREES
/CRE/ : DONNEES CENTREES ET REDUITES

FORME DES DONNEES ? /BRU/CEN/CRE/ ←⊞→ BRU
TEST D'ARRET ? /O/N/ ←⊞→ ?

L'UTILISATEUR PEUT IMPOSER L'ARRET DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE:
- A UN NIVEAU DE L'ARBRE
ET/OU - LORSQU'UNE CLASSE CONTIENT UN EFFECTIF QUI EST INSUFFISANT POUR POURSUIVRE L'ALGORITHME

TEST D'ARRET ? /O/N/ ←⊞→ O
DE QUELLE FACON ? /EFF/NIV/EFN/ ←⊞→ ?

LES TESTS D'ARRET PEUVENT ETRE LES SUIVANTS:
/EFF/ : ARRET DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE POUR CHAQUE CLASSE DONT L'EFFECTIF D'UNE CLASSE EST INFERIEUR OU EGAL A UN NOMBRE DONNE.
/NIV/ : ARRET DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE LORSQU'UN NIVEAU DE L'ARBRE EST ATTEINT
/EFN/ : CORRESPOND A LA CONJONCTION DES DEUX CONDITIONS PRECEDENTES

DE QUELLE FACON ? /EFF/NIV/EFN/ ←⊞→ NIV
→ NOMBRE DE PALIERS A CALCULER ? ←⊞→ 2

```
PARAMETRES POUR LA CLASSIFICATION HIERARCHIQUE DESCENDANTE /X/ (27.12.1982 3H 34)
************************************************************
LA MATRICE DES OBSERVATIONS DE LA MATRICE DES DONNEES
BRUTES
PAR LA METHODE DE KMEANS
RACINE DE L'ARBRE: 0
TEST D'ARRET:
  -EFFECTIF MINIMUM D'UNE CLASSE: 2
  -NOMBRE DE PALIERS: 2

IMPRESSION DES ETAPES DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE ? /O/N/
+⊞+ O

REPRESENTATION DES ELEMENTS POUR CHAQUE NOEUD
*********************************************

   1 NIVEAU
      CLASSE:0
      NOMBRE D'OBJETS: 9
      VARIANCE DE LA CLASSE:  1.3453E5
      VARIANCE INTRA-CLASSES: 9.5334E4
      PART DE 1= 7.8011E4
      PART DE 2= 1.7324E4
      OBJETS DE LA CLASSE:
      QUALITE     TECHNIQUE   PRECISION   ESTHETIQUE  FIABILITE
      VALEUR      SOLIDITE    ELEGANT     ELECTRON

      2 NIVEAU
         CLASSE:1
         NOMBRE D'OBJETS: 6
         VARIANCE DE LA CLASSE:  1.1702E5
         VARIANCE INTRA-CLASSES: 4.4904E4
         PART DE 11= 2.9163E4
         PART DE 12= 1.5742E4
         OBJETS DE LA CLASSE:
         QUALITE     TECHNIQUE   PRECISION   FIABILITE   SOLIDITE
         ELECTRON

         3 NIVEAU
            CLASSE:11
            NOMBRE D'OBJETS: 4
            VARIANCE DE LA CLASSE:  4.3744E4
            OBJETS DE LA CLASSE:
            QUALITE     PRECISION   FIABILITE   SOLIDITE

         3 NIVEAU
            CLASSE:12
            NOMBRE D'OBJETS: 2
            VARIANCE DE LA CLASSE:  4.7225E4
            OBJETS DE LA CLASSE:
            TECHNIQUE   ELECTRON
```

```
     2 NIVEAU
        CLASSE:2
        NOMBRE D'OBJETS: 3
        VARIANCE DE LA CLASSE:  5.1971E4
        VARIANCE INTRA-CLASSES: 2.9918E3
        PART DE 21= 2.9918E3
        PART DE 22= 0.0000E0
        OBJETS DE LA CLASSE:
        ESTHETIQUE  VALEUR      ELEGANT

        3 NIVEAU
           CLASSE:21
           NOMBRE D'OBJETS: 2
           VARIANCE DE LA CLASSE:  4.4877E3
           OBJETS DE LA CLASSE:
           ESTHETIQUE  ELEGANT

        3 NIVEAU
           CLASSE:22
           NOMBRE D'OBJETS: 1
           OBJETS DE LA CLASSE:
           VALEUR


DESSIN DE L'ARBRE ? /O/N/ <RET> O
QUEL INDICE ? /VCL/VIN/ <RET> ?

LES DEUX INDICES SUIVANTS PEUVENT ETRE CHOISIS:
   /VCL/ : LA VARIANCE DE LA CLASSE
   /VIN/ : LA VARIANCE INTRA-CLASSES

QUEL INDICE ? /VCL/VIN/ <RET> VCL

REPRESENTATION DE L'ARBRE
*************************

      <---------> VAUT   2.2422E4   INDICE=VARIANCE DE LA CLASSE

QUALI-+-------------------------------------------------------+------+
PRECI-+                                                       |      |
FIABI-+                                                       |      |
SOLID-+                                                       |      |
TECHN-+-----------------------------------------------------+ |      |
ELECT-+                                                     | |      |
ESTHE-+---------------------+-------------------------------+-+------+
ELEGA-+                     |
VALEU-----------------------+
```

```
POURSUITE DE LA CLASSIFICATION A UN NOEUD ? /O/N/ ←⊞→ ?

L'UTILISATEUR PEUT DIVISER UNE CLASSE DONT LE NOEUD ASSOCIE N'A PAS
ENCORE ETE EXPLORE.
VOICI LA LISTE DES NOEUDS NON ENCORE EXPLORES:
11

POURSUITE DE LA CLASSIFICATION A UN NOEUD ? /O/N/ ←⊞→ O
→ NUMERO DE LA PROCHAINE CLASSE ? ←⊞→ 11
TEST D'ARRET ? /O/N/ ←⊞→ N

PARAMETRES POUR LA CLASSIFICATION HIERARCHIQUE DESCENDANTE /X/ (27.12.1982 3H 36)
**********************************************************
LA MATRICE DES OBSERVATIONS DE LA MATRICE DES DONNEES
BRUTES
PAR LA METHODE DE KMEANS
RACINE DE L'ARBRE: 11
TEST D'ARRET:
  -EFFECTIF MINIMUM D'UNE CLASSE: 2
  -NOMBRE DE PALIERS: 100

IMPRESSION DES ETAPES DE LA CLASSIFICATION HIERARCHIQUE DESCENDANTE ? /O/N/
←⊞→ O

REPRESENTATION DES ELEMENTS POUR CHAQUE NOEUD
*********************************************

        3 NIVEAU
          CLASSE:11
          NOMBRE D'OBJETS: 4
          VARIANCE DE LA CLASSE:  4.3744E4
          VARIANCE INTRA-CLASSES: 3.4832E3
          PART DE 111= 0.0000E0
          PART DE 112= 3.4832E3
          OBJETS DE LA CLASSE:
          QUALITE     PRECISION   FIABILITE   SOLIDITE

          4 NIVEAU
            CLASSE:111
            NOMBRE D'OBJETS: 1
            OBJETS DE LA CLASSE:
            QUALITE

          4 NIVEAU
            CLASSE:112
            NOMBRE D'OBJETS: 3
            VARIANCE DE LA CLASSE:  4.6442E3
            VARIANCE INTRA-CLASSES: 3.7773E3
            PART DE 1121= 3.7773E3
            PART DE 1122= 0.0000E0
            OBJETS DE LA CLASSE:
            PRECISION   FIABILITE   SOLIDITE
```

```
5 NIVEAU
   CLASSE:1121
   NOMBRE D'OBJETS: 2
   VARIANCE DE LA CLASSE:  5.6660E3
   OBJETS DE LA CLASSE:
   PRECISION   SOLIDITE

5 NIVEAU
   CLASSE:1122
   NOMBRE D'OBJETS: 1
   OBJETS DE LA CLASSE:
   FIABILITE


DESSIN DE L'ARBRE ? /O/N/ ←⊞→ O VCL

REPRESENTATION DE L'ARBRE
*************************


      <--------> VAUT   7.2907E3   INDICE=VARIANCE DE LA CLASSE

QUALI-----------------------------------------------------------+
PRECI-+------+--------------------------------------------------+
SOLID-+      |
FIABI-------+


QUEL BRANCHEMENT ? /MAT/MET/ARR/IMP/ARB/NOE/ ←⊞→ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE DE CLASSIFICATION HIERARCHIQUE
DESCENDANTE:

  1) /MAT/ : CHOIX D'UNE NOUVELLE MATRICE DE DONNEES
  2) EXECUTER SUR LA MEME MATRICE DES DONNEES UNE NOUVELLE
     CLASSIFICATION.
      /MET/ : EN CHOISISSANT UN NOUVEL ALGORITHME DE CLASSIFICATION HIERARCHIQUE
              DESCENDANTE
      /ARR/ : EN CONSERVANT L'ALGORITHME: DIVISION SELON HIERKMEAN, MAIS EN IMPOSANT
              UN NOUVEAU TEST D'ARRET

  3)  OBTENIR DES RENSEIGNEMENTS SUR LA CLASSIFICATION QUI VIENT D'ETRE EXECUTEE
      /IMP/ : IMPRESSION DES ETAPES DE LA CLASSIFICATION HIERARCHIQUE
              DESCENDANTE
      /ARB/ : DESSIN DE L'ARBRE

  4)  /NOE/ : CHOIX D'UNE SOUS-CLASSE NON ENCORE DIVISEE

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/F/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←⊞→ N

QUEL BRANCHEMENT ? /MAT/MET/ARR/IMP/ARB/NOE/ ←⊞→ S O
FIN DE L'EXECUTION DU PACKAGE
```

9. LA CLASSIFICATION HIERARCHIQUE DESCENDANTE MONOTHETIQUE PAR LA METHODE "AUTOMATIC INTERACTIVE DETECTOR (AID)", APPELEE EGALEMENT LA SEGMENTATION (Sonquist et Morgan)

Comme toutes les méthodes de classification hiérarchique, la segmentation cherche une hiérarchie de partitions emboîtées. Elle se distingue des autres méthodes que nous avons déjà présentées par la forme particulière de sa matrice de données, qui doit être qualitative et par le critère optimisé qui ne prend en compte qu'une seule variable. Cette dernière particularité permet de qualifier la méthode de monothétique.

9.1. La matrice des données

9.1.1. Introduction

La méthode travaille sur une matrice de données particulière, puisqu'une des variables est choisie comme variable expliquée et que les autres sont considérées comme des variables explicatives. La variable expliquée doit être de type booléen, vecteur formé de 0 et de 1. Chaque variable explicative pourra être soit nominale, soit ordinale. Le type de chaque variable explicative devra être spécifié, car l'algorithme en dépend. Dans une étude qui cherche à expliquer les raisons qui poussent un individu à posséder un appareil de télévision, le j-ième élément de la variable expliquée sera égal à 1 si la personne interrogée possède un appareil de TV et il vaudra 0 dans le cas contraire. Les variables explicatives seront, par exemple, le sexe, l'âge, la profession, le domicile, ...
En plus de la construction d'une hiérarchie de partitions, la segmentation cherche alors à établir une relation de dépendance entre un phénomène déterminé et des causes qui pourraient être à son origine. On essaie donc de mettre en relation la variable expliquée avec un ensemble de variables explicatives. Remarquons qu'au lieu de classer la segmentation parmi les méthodes de classification automatique, il aurait été également judicieux de la classer parmi les

méthodes explicatives qui recherchent une relation entre deux ensembles de variables. La segmentation peut donc être comparée à la régression. Elle diffère cependant de cette dernière par le fait qu'elle ne détermine pas une fonction mathématique entre la variable expliquée et les variables explicatives.
La segmentation permet uniquement de "prévoir avec précision la valeur prise par la variable expliquée quand les variables explicatives prennent des valeurs données et, inversément pour la variable expliquée, on peut prévoir la combinaison des valeurs des variables explicatives" (Bosa /41/).

9.1.2. La notation de la matrice des données

D'après ce que nous venons de voir, la matrice des données X doit pouvoir se décomposer en un vecteur booléen (la variable expliquée), noté Y, et en une matrice de variables explicatives, notée V. En renumérotant au besoin les colonnes de la matrice des données X on peut toujours supposer que Y en constitue la première colonne.
Nous pouvons alors écrire :

$$X = (Y\ V) = (Y\ V^1\ V^2\ \ldots\ V^{n-1})$$

ou

$$\begin{bmatrix} x_1^1 & x_1^2 & \ldots & x_1^n \\ x_2^1 & x_2^2 & \ldots & x_2^n \\ \ldots & \ldots & \ldots & \ldots \\ x_m^1 & x_m^2 & \ldots & x_m^n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_m \end{bmatrix} \begin{bmatrix} v_1^1 & v_1^2 & \ldots & v_1^{n-1} \\ v_2^1 & v_2^2 & \ldots & v_2^{n-1} \\ \ldots & \ldots & \ldots & \ldots \\ v_m^1 & v_m^2 & \ldots & v_m^{n-1} \end{bmatrix}$$

9.2. Présentation générale de la méthode

Le point de départ de la méthode est une partition formée d'une seule classe $C_o = \{X_1, X_2, \ldots, X_m\}$. A chaque itération le nombre de classes de la partition va augmenter

d'une unité, car une classe va être subdivisée en deux sous-classes. En segmentation, une classe est également appelée un segment.

9.2.1. La division d'une classe en deux sous-classes

Pour subdiviser une classe en deux sous-classes il faut d'abord choisir une variable explicative, puis on affectera les objets à l'une ou l'autre sous-classe en se basant uniquement sur les valeurs prises par la variable explicative retenue.

Dans notre exposé, nous expliquerons donc successivement:

a) comment on dichotomise (subdivise en deux parties) les états d'une variable explicative
b) comment on construit le tableau de contingence entre la variable expliquée Y et la variable explicative dichotomisée
c) comment on calcule le CHI2 associé au tableau de contingence qui sera le critère retenu pour choisir à la fois la variable expliquée et la dichotomie de ses états.

a) Admettons qu'il s'agit de diviser une classe notée C en deux sous-classes et que la variable explicative retenue soit V^j. Notons $E=\{e_1, e_2, \ldots, e_q\}$ l'ensemble des états possibles de la variable explicative V^j. Alors chaque fois que nous nous donnons une 2-partition propre de E (c'est-à-dire deux sous-ensembles E_1 et E_2 de E qui vérifient $E_1 \neq \emptyset$, $E_2 \neq \emptyset$, $E_1 \cap E_2 = \emptyset$ et $E_1 \cup E_2 = E$), nous pouvons diviser la classe C en deux sous-classes C_1 et C_2 par la règle suivante: si pour une observation X_i appartenant à la classe C, la valeur prise par la variable explicative V_j appartient à E_1, on affectera cette observation à la sous-classe C_1, alors qu'on l'affectera à la sous-classe C_2 dans le cas contraire.
Mathématiquement nous pouvons écrire:

$$C_1 = C \cap \{i \mid v_i^j \in E_1\} \text{ et } C_2 = C \cap \{i \mid v_i^j \in E_2\}$$

b) En se référant toujours à la 2-partition (E_1, E_2) des états de la variable explicative V^j, on peut calculer un tableau de contingence entre la variable expliquée Y et la variable explicative V^j.

Y \ V^j	$V^j \in E_1$	$V^j \in E_2$	total
Y=1	a_{11}	a_{12}	$a_{1.}$
Y=0	a_{21}	a_{22}	$a_{2.}$
total	$a_{.1}$	$a_{.2}$	a

L'élément a_{11} représente l'effectif des éléments de la classe C pour lesquels Y vaut 1 et $V^j \in E_1$.
Les notations a, $a_{i.}$ et $a_{.j}$ correspondent à:

$$a_{i.} = a_{i1} + a_{i2} \qquad i \in \{1, 2\}$$

$$a_{.j} = a_{1j} + a_{2j} \qquad j \in \{1, 2\}$$

$$a = a_{11} + a_{12} + a_{21} + a_{22}$$

c) On peut ensuite calculer le critère S, associé au tableau de contingence précédent, par la formule suivante:

$$S = \frac{\left(a_{11} - \frac{a_{1.}}{a} a_{.1}\right)^2}{\frac{a_{1.}}{a} a_{.1}} + \frac{\left(a_{11} - \frac{a_{1.}}{a} a_{.2}\right)^2}{\frac{a_{1.}}{a} a_{.2}} + \frac{\left(a_{21} - \frac{a_{2.}}{a} a_{1.}\right)^2}{\frac{a_{2.}}{a} a_{.1}} + \frac{\left(a_{22} - \frac{a_{2.}}{a} a_{.2}\right)^2}{\frac{a_{2.}}{a} a_{.2}}$$

Le critère S qui est égal à la somme des carrés des différences entre l'effectif observé et l'effectif théorique, pondéré par l'inverse de l'effectif théorique, est une mesure de la dépendance entre la variable expliquée et la variable explicative. L'effectif théorique correspond à l'effectif obtenu en faisant l'hypothèse que les deux variables sont indépendantes. Dans ce cas et si l'effectif de chaque case du tableau de contingence est au moins égal à 5, le critère S est distribué comme un CHI2 à un degré de liberté. Dans la suite de l'exposé nous appellerons par abus de langage le critère S critère du CHI2.

9.2.2. Le critère à maximiser

L'algorithme procède par subdivisions successives de classes en deux sous-classes. A chaque itération, on cherchera parmi toutes les variables explicatives et toutes leurs dichotomies admissibles celles (la variable explicative et la dichotomie) qui donnent la plus grande valeur du CHI2. Les dichotomies admissibles dépendent du type de la variable explicative. Si la variable explicative est ordinale, on n'admettra que les dichotomies qui respectent la relation d'ordre, soit les q-1 partitions suivantes:

$E_1 = \{e_1\}$ $E_2 = \{e_2, e_3, \ldots, e_q\}$
$E_1 = \{e_1, e_2\}$ $E_2 = \{e_3, e_4, \ldots, e_q\}$
$E_1 = \{e_1, e_2, e_3\}$ $E_2 = \{e_4, e_5, \ldots, e_q\}$
.....
$E_1 = \{e_1, e_2, \ldots, e_{q-1}\}$ $E_2 = \{e_q\}$

Si la variable explicative est nominale, toutes les 2-partitions propres sont les dichotomies admissibles et il y en a donc $2^{q-1}-1$. Présentons quelques exemples:

$E_1 = \{e_1, e_3\}$ $E_2 = \{e_2, e_4, \ldots, e_q\}$
$E_1 = \{e_1, e_{q-1}\}$ $E_2 = \{e_2, e_3, \ldots, e_q\}$

Il est important de se rappeler, lors de l'utilisation de la méthode, que le traitement d'une variable nominale nécessite

plus de temps de calcul que le traitement d'une variable ordinale.

9.3. Résumé de l'algorithme

Les bases de la méthode étant définies, il est facile de la résumer.

Initialisation

$C_o = \{X_1, X_2, \ldots, X_m\}$

LISTE = $\{C_o\}$: la liste des noms des classes déjà créées et qui doivent être subdivisées en deux sous-classes

Itération

A chaque itération on subdivise une classe appartenant à LISTE en deux sous-classes. Nous allons décrire une itération en nous basant sur un exemple. Admettons qu'au début de l'itération, nous soyons dans la situation suivante:

$C \in$ LISTE
$C = \{X_a, X_b, X_c, X_y, X_z\}$

L'itération peut être résumée par les opérations suivantes:

1) rechercher la meilleure dichotomie :

Max $\{CHI2 \mid V^j \in V, \{E_1, E_2\}$ est une dichotomie des états de $V^j\}$

Le maximum sera noté MCHI2.
Admettons que la meilleure dichotomie soit obtenue pour la variable explicative V^t et la partition de ses états, $\{E_1^t, E_2^t\}$.

2) mémoriser les caractéristiques associées à la meilleure dichotomie, MCHI2, V^t, E_1^t et E_2^t.

3) former deux sous-classes C_1 et C_2 de la façon suivante:

si $x_i^t \epsilon E_1$ alors $X_i \epsilon C_1$

si $x_i^t \epsilon E_2$ alors $X_i \epsilon C_2$

4) Mettre à jour la liste des noms des classes qui doivent être encore subdivisées. L'élément C de la liste devra être effacé et les deux nouvelles classes doivent y être ajoutées, à moins que les conditions d'arrêt ne soient vérifiées. Après ces modifications, LISTE contiendra entre autres, les noms des classes C_1 et C_2.

5) La partition obtenue à la fin de l'itération est :
$C = \{C_1, C_2\}$
avec, par exemple :
$C_1 = \{X_a, X_z\}$
$C_2 = \{X_b, X_c, X_y\}$

Arrêt de l'algorithme

L'algorithme s'arrête quand la liste est vide. Une sous-classe ne sera pas mise dans la liste, ou autrement dit elle ne sera plus subdivisée, si l'une ou l'autre des conditions suivantes est vérifiée:

- la classe ne contient qu'un seul élément
- l'une des conditions d'arrêt prématuré est vérifiée (cf. les options de la méthode)
- la classe est homogène, ce qui signifie que tous ses éléments ont la même valeur pour la variable expliquée ou ce qui revient au même, l'une des valeurs suivantes a_1. ou a_2. du tableau de contingence est nulle.

9.4. Les options de l'algorithme

Rappelons que la segmentation ne peut être exécutée que sur la matrice des données.

Nous pouvons classer les options de la méthode en deux niveaux; le premier correspond aux options relatives à l'exécution de l'algorithme et le deuxième à des facilités qui sont implantées dans le but d'alléger le dépouillement des résultats.

9.4.1. Les options relatives à l'exécution de l'algorithme

9.4.1.1. Exécution automatique ou avec intervention de l'utilisateur

L'utilisateur peut choisir entre deux variantes d'exécution de la segmentation. La première option permet d'exécuter l'algorithme automatiquement, alors que dans la deuxième variante l'utilisateur peut intervenir et privilégier une variable ou même une dichotomie pour une variable et ceci à n'importe quel stade du processus. Cette option donne une grande souplesse à la méthode. Si le but d'une étude est de comprendre les caractéristiques qui distinguent les fumeurs des non-fumeurs et que l'analyste suppose que le comportement des hommes est différent de celui des femmes, il peut en début d'analyse privilégier la variable sexe, ce qui permettra d'étudier le comportement des deux sous-populations. Il est clair que si l'on abuse de cette option, les résultats ne peuvent que refléter les préjugés de l'analyste.

9.4.1.2. Les tests d'arrêt

Si l'utilisateur n'impose aucune condition d'arrêt, l'algorithme va constituer une hiérarchie complète de classes. Cependant d'éventuelles classes homogènes ne sont pas subdivisées.
Par contre, si l'utilisateur ne s'intéresse qu'au début de la hiérarchie, l'algorithme peut être arrêté prématurément en imposant l'une et/ou l'autre des deux conditions suivantes:

- la classe contient un effectif inférieur ou égal à une valeur minimale choisie par l'utilisateur

- le nombre de paliers de l'arbre dépasse une valeur maximale fixée par l'utilisateur

9.4.1.3. Poursuite de la segmentation à un noeud non encore exploré

Si l'utilisateur a imposé l'arrêt de la segmentation après un nombre de paliers donné, il pourra continuer d'exécuter la segmentation à un noeud non encore exploré.

9.4.1.4. Les résultats

Les itérations de l'algorithme sont résumées par les résultats suivants:

a) Une première représentation est donnée sous forme d'identation. Pour chaque noeud de l'arbre, on donne le nom de la variable explicative, la dichotomie de ses états, la valeur du CHI2, l'effectif de la population subdivisée, les effectifs du segment pour lesquels la variable expliquée vaut 0 respectivement 1, ainsi que leurs pourcentages par rapport au segment et à la population totale

b) La deuxième représentation, également sous forme d'identation, donne, pour chaque segment, le nom de la variable explicative, la dichotomie de ses états ainsi que les noms des objets répartis selon les états de la variable expliquée.

c) Le dendrogramme.

Il est également possible d'obtenir les tableaux de contingence entre toutes les variables explicatives et la variable expliquée. De plus, on peut, pour un ou plusieurs segments, obtenir les n (nombre choisi par l'utilisateur) meilleures dichotomies. Cela permet d'effectuer des études de sensibilité par rapport à la variable explicative retenue et à la dichotomie associée. En particulier, si l'utilisateur a privilégié une dichotomie, il est important de situer la position de celle-ci parmi les meilleures dichotomies calculées par le programme.

9.4.2. Les options qui facilitent la lecture des réaultats

Afin de fsciliter le dépouillement dea réaultats, l'utiliss-teur peut associer des noms aux étata de toutes les variables explicatives ou d'une partie d'entre elles. Lors des impreasions, les noms aeront reproduits plutôt que les valeura. Dans les dialogues, lea étsts seront définis par leurs codes et non pss par leura noms.

9.5. Remsrques concernant la méthode

Pareille à toutes les méthodes qui optimisent un critère à chsque itération, la segmentation n'stteint pas forcément un optimum globsl.

9.6. Exemples

Puisque la méthode s'effectue uniquement aur une mstrice de données qualitstives, dont une variable eat choisie comme variable expliquée, une mstrice de données particulières est utilisée.
Les vsrisbles explicatives aont X1, X2, X3, X4 et X5. Les quatres premières sont de type ordinal, alors que la dernière est de type nominsle. Ls vsrisble Y est à expliquer.
Les modules suivsnts sont sppelés (dsns l'ordre) : /E/, /I/, /W/ et /T/.

GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ O
MODULE CHOISI ? /E/P/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ E

****
MODULE DE DEFINITION, CORRECTION, MEMORISATION DE LA MATRICE DES DONNEES /E/
****

→ QUEL EST VOTRE NOM ? (5 LETTRES) ←⊞→ SEGME
QUELLE OPERATION ? /CRE/ETE/MOD/DEL/SAV/?/ ←⊞→ CRE DIS I DON O [O] W
LECTURE CORRECTE

****
MODULE D'IMPRESSION /I/
****

IMPRESSION DE LA MATRICE DES DONNEES

	Y	X1	X2	X3	X4	X5
1	1	5	4	2	3	1
2	0	2	4	3	2	4
3	1	1	2	3	1	5
4	1	4	5	2	2	4
5	1	3	1	2	4	5
6	0	3	5	5	3	3
7	1	5	3	2	5	3
8	1	3	3	2	4	3
9	0	3	5	2	4	2
10	1	4	5	2	2	2
11	0	1	2	4	4	3
12	0	3	1	4	4	4
13	1	1	1	5	3	3
14	1	5	1	3	5	3
15	0	4	3	2	3	2
16	1	4	2	3	4	4
17	1	3	4	3	2	3
18	1	4	1	5	4	4
19	1	3	2	2	2	4
20	0	2	4	5	3	3

****
MODULE DE CLASSIFICATION HIERARCHIQUE DESCENDANTE MONOTHETIQUE: SEGMENTATION /W/
****

TOUTES LES VARIABLES SONT-ELLES QUALITATIVES ? /O/N/ ←⊞→ ?

POUR EXECUTER L'ALGORITHME DE SEGMENTATION TOUTES LES VARIABLES DOIVENT ETRE DE TYPE QUALITATIP. SI LA MATRICE DES DONNEES NE REPOND PAS A CETTE CONDITION, L'UTILISATEUR DOIT APPELER LE MODULE /M/ APIN DE MASQUER LES VARIABLES QUANTITATIVES.
SI TOUTES LES VARIABLES SONT QUANTITATIVES, L'UTILISATEUR NE PEUT PAS UTILISER CE MODULE.

TOUTES LES VARIABLES SONT-ELLES QUALITATIVES ? /O/N/ ←⊞→ O
→ NOM DE LA VARIABLE EXPLIQUEE ? ←⊞→ ?

```
LA VARIABLE EXPLIQUEE EST UNE VARIABLE QUI DOIT ETRE DE TYPE BOOLEEN,
C'EST-A-DIRE DONT TOUS LES ELEMENTS VALENT 0 OU 1.

→ NOM DE LA VARIABLE EXPLIQUEE ? ←↵→ Y
TOUTES LES VARIABLES SONT-ELLES ORDINALES ? /O/N/ ←↵→ ?

L'ALGORITHME IMPLANTE PERMET DE TRAITER DES VARIABLES NOMINALES OU ORDINALES.
LE TRAITEMENT DES VARIABLES NOMINALES EST DIPPERENT DE CELUI DES VARIABLES
ORDINALES.
L'UTILISATEUR DOIT DONC INDIQUER S'IL EXISTE DES VARIABLES NOMINALES.
REMARQUE:
LE TEMPS DE TRAITEMENT DES VARIABLES NOMINALES EST NETTEMENT PLUS LONG QUE
CELUI DES VARIABLES ORDINALES.

TOUTES LES VARIABLES SONT-ELLES ORDINALES ? /O/N/ ←↵→ N
TOUTES LES VARIABLES SONT-ELLES NOMINALES ? /O/N/ ←↵→ N
→ NOMS DES VARIABLES NOMINALES ? ←↵→ X5
DONNER DES NOMS AUX ETATS DES VARIABLES EXPLICATIVES ? /O/N/ ←↵→ ?

L'UTILISATEUR PEUT ASSOCIER DES NOMS AUX ETATS DE TOUTES OU PARTIE DES
VARIABLES EXPLICATIVES.
CETTE OPTION A POUR BUT DE PACILITER LA LECTURE DES RESULTATS.

DONNER DES NOMS AUX ETATS DES VARIABLES EXPLICATIVES ? /O/N/ ←↵→ O
A TOUTES LES VARIABLES EXPLICATIVES ? /O/N/ ←↵→ O
→ ENTREZ AU-DESSOUS DES ETATS DE LA VARIABLE X1 LEURS NOMS ? ←↵→
1          2          3          4          5
R1         R2         R3         R4         R5
→ ENTREZ AU-DESSOUS DES ETATS DE LA VARIABLE X2 LEURS NOMS ? ←↵→
1          2          3          4          5
INSA       SATI       MOY        BON        EXC
→ ENTREZ AU-DESSOUS DES ETATS DE LA VARIABLE X3 LEURS NOMS ? ←↵→
2          3          4          5
VILA       VA         ASSE       BEAU
→ ENTREZ AU-DESSOUS DES ETATS DE LA VARIABLE X4 LEURS NOMS ? ←↵→
1          2          3          4          5
X41        X42        X43        X44        X45
→ ENTREZ AU-DESSOUS DES ETATS DE LA VARIABLE X5 LEURS NOMS ? ←↵→
1          2          3          4          5
CELI       MARI       DIV        SEPA       VEUP

IMPRESSION DES TABLEAUX DE CONTINGENCE ? /O/N/ ←↵→ ?

L'UTILISATEUR PEUT IMPRIMER DES TABLEAUX DE CONTINGENCE (EN POURCENTAGE)
ENTRE TOUTES LES VARIABLES EXPLICATIVES ET LA VARIABLE EXPLIQUEE Y.

IMPRESSION DES TABLEAUX DE CONTINGENCE ? /O/N/ ←↵→ O
```

EFFECTIFS DU SEGMENT 0 *ET REPARTITION SELON*
LES ETATS DES VARIABLES EXPLICATIVES
LE SEGMENT CONTIENT 20 *UNITES STATISTIQUES*
****

ETATS DE LA VARIABLE EXPLICATIVE X1

	R1	*R2*	*R3*	*R4*	*R5*	*EFFECTIFS*
Y =0	14.29	28.57	42.86	14.29	0.00	7.00
Y =1	15.38	0.00	30.77	30.77	23.08	13.00
TOTA	15.00	10.00	35.00	25.00	15.00	20.00

ETATS DE LA VARIABLE EXPLICATIVE X2

	INSA	*SATI*	*MOY*	*BON*	*EXC*	*EFFECTIFS*
Y =0	14.29	14.29	14.29	28.57	28.57	7.00
Y =1	30.77	23.08	15.38	15.38	15.38	13.00
TOTA	25.00	20.00	15.00	20.00	20.00	20.00

ETATS DE LA VARIABLE EXPLICATIVE X3

	VILA	*VA*	*ASSE*	*BEAU*	*EFFECTIFS*
Y =0	28.57	14.29	28.57	28.57	7.00
Y =1	53.85	30.77	0.00	15.38	13.00
TOTA	45.00	25.00	10.00	20.00	20.00

ETATS DE LA VARIABLE EXPLICATIVE X4

	X41	*X42*	*X43*	*X44*	*X45*	*EFFECTIFS*
Y =0	0.00	14.29	42.86	42.86	0.00	7.00
Y =1	7.69	30.77	15.38	30.77	15.38	13.00
TOTA	5.00	25.00	25.00	35.00	10.00	20.00

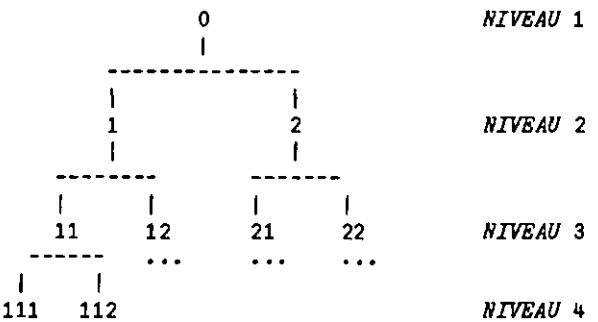
ETATS DE LA VARIABLE EXPLICATIVE X5

	CELI	*MARI*	*DIV*	*SEPA*	*VEUF*	*EFFECTIFS*
Y =0	0.00	28.57	42.86	28.57	0.00	7.00
Y =1	7.69	7.69	38.46	30.77	15.38	13.00
TOTA	5.00	15.00	40.00	30.00	10.00	20.00

OPTION POUR LA SEGMENTATION ? /AUT/INT/ ⊞ ?

/AUT/ : LA SEGMENTATION EST EXECUTEE AUTOMATIQUEMENT
/INT/ : CETTE OPTION TRES SOUPLE PERMET DE PRIVILEGIER A N'IMPORTE QUEL SEGMENT (NOEUD) N'IMPORTE QUELLE VARIABLE.
L'UTILISATEUR INTERVIENT A TOUS LES NOEUDS LORS DE LA SEGMENTATION.
IL CHOISIRA A CHAQUE NOEUD UNE DES OPTIONS SUIVANTES:
- *EXECUTION AUTOMATIQUE DE LA PROCHAINE DICHOTOMIE*
- *CHOIX IMPERATIF D'UNE VARIABLE EXPLICATIVE*
- *CHOIX IMPERATIF D'UNE VARIABLE EXPLICATIVE ET DE SES ETATS SUR L'UN DES DEUX SEGMENTS.*
- *ABANDON DU NOEUD*

LES RESULTATS DE LA SEGMENTATION SONT RESUMES PAR UN ARBRE.
DANS CE MODULE, LA TERMINOLOGIE EST LA SUIVANTE:



LES NUMEROS 0 1 2 11 12 21 22 111 112 DESIGNENT LES SEGMENTS OU LES NOEUDS DE L'ARBRE.
LES NIVEAUX CORRESPONDANT AUX SEGMENTS CI-DESSUS SONT 1 2 3 4.
LA RACINE DE L'ARBRE EST 0.
LE NOMBRE DE PALIERS DE CET ARBRE EST DE 3.

OPTION POUR LA SEGMENTATION ? /AUT/INT/ ↵ AUT
TEST D'ARRET ? /O/N/ ↵ ?

L'UTILISATEUR PEUT IMPOSER L'ARRET DE LA SEGMENTATION :
- *A UN NIVEAU DE L'ARBRE*
- */ET/OU/ LORSQU'UNE SOUS-POPULATION CONTIENT UN EFFECTIF QUI PARAIT INSUFFISANT POUR POURSUIVRE LA SEGMENTATION.*

TEST D'ARRET ? /O/N/ ↵ O
DE QUELLE FACON ? /EFF/NIV/EFN/ ↵ ?

LES TESTS D'ARRET PEUVENT ETRE LES SUIVANTS:
/EFF/ : ARRET DE LA SEGMENTATION LORSQUE L'EFFECTIF D'UNE SOUS-POPULATION EST INFERIEUR OU EGAL A UN EFFECTIF DONNE.
/NIV/ : ARRET DE LA SEGMENTATION LORSQU'UN NIVEAU DE L'ARBRE EST ATTEINT
/EFN/ : CORRESPOND A LA CONJONCTION DES DEUX CONDITIONS PRECEDENTES

```
DE QUELLE FACON ? /EPP/NIV/EPN/ ←⊞→ NIV
→ NOMBRE DE PALIERS A CALCULER ? ←⊞→ 2

PARAMETRES POUR LA SEGMENTATION /W/ (27.12.1982 3H 50)
**************************************
MATRICE DES DONNEES
VARIABLE EXPLIQUEE: Y
EFFECTIF DE LA POPULATION: 20
EFFECTIF DE LA POPULATION POUR LAQUELLE Y=1 13 POURC= 65.00
EFFECTIF DE LA POPULATION POUR LAQUELLE Y=0 7 POURC= 35.00
TEST D'ARRET
EFFECTIF MINIMUM POUR UNE SOUS-POPULATION: 1
NOMBRE MAXIMUM DE PALIERS: 2
LES VARIABLES SUIVANTES SONT NOMINALES:
X5


REPRESENTATION DE L'ARBRE
*************************

   1 NIVEAU
      SEGMENT:0
      EFFECTIF:20
      EFFECTIF Y=1 13
      EFFECTIF Y=0 7
      POURC. S.POP:Y=1  65.00
      POURC. S.POP:Y=0  35.00
      POURC. POP:Y=1  65.00
      POURC. POP:Y=0  35.00

      2 NIVEAU
         SEGMENT:1
         VARIABLE: X3            ETATS: VILA VA
         EFFECTIF:14
         EFFECTIF Y=1 11
         EFFECTIF Y=0 3
         POURC. S.POP:Y=1  78.57
         POURC. S.POP:Y=0  21.43
         POURC. POP:Y=1  55.00
         POURC. POP:Y=0  15.00
         CHI2=3.778

         3 NIVEAU
            SEGMENT:11
            VARIABLE: X5            ETATS: MARI
            EFFECTIF:3
            EFFECTIF Y=1 1
            EFFECTIF Y=0 2
            POURC. S.POP:Y=1  33.33
            POURC. S.POP:Y=0  66.67
            POURC. POP:Y=1   5.00
            POURC. POP:Y=0  10.00
            CHI2=4.641
```

```
   3 NIVEAU
      SEGMENT:12
      VARIABLE: X5            ETATS: CELI DIV  SEPA VEUF
      EFFECTIF:11
      EFFECTIF Y=1 10
      EFFECTIF Y=0 1
      POURC. S.POP:Y=1   90.91
      POURC. S.POP:Y=0    9.09
      POURC. POP:Y=1   50.00
      POURC. POP:Y=0    5.00
      CHI2=4.641

2 NIVEAU
   SEGMENT:2
   VARIABLE: X3            ETATS: ASSE BEAU
   EFFECTIF:6
   EFFECTIF Y=1 2
   EFFECTIF Y=0 4
   POURC. S.POP:Y=1   33.33
   POURC. S.POP:Y=0   66.67
   POURC. POP:Y=1   10.00
   POURC. POP:Y=0   20.00
   CHI2=3.778

   3 NIVEAU
      SEGMENT:21
      VARIABLE: X2            ETATS: INSA
      EFFECTIF:3
      EFFECTIF Y=1 2
      EFFECTIF Y=0 1
      POURC. S.POP:Y=1   66.67
      POURC. S.POP:Y=0   33.33
      POURC. POP:Y=1   10.00
      POURC. POP:Y=0    5.00
      CHI2=3

   3 NIVEAU
      SEGMENT:22
      VARIABLE: X2            ETATS: SATI BON  EXC
      EFFECTIF:3
      EFFECTIF Y=1 0
      EFFECTIF Y=0 3
      POURC. S.POP:Y=1    0.00
      POURC. S.POP:Y=0  100.00
      POURC. POP:Y=1    0.00
      POURC. POP:Y=0   15.00
      CHI2=3
```

```
LISTE DES ELEMENTS DE CHAQUE SEGMENT ? /O/N/ ←→ O

LISTE DES ELEMENTS DES SEGMENTS
*******************************

   1 NIVEAU
      SEGMENT:0
      OBJETS POUR LESQUELS Y=1
      1          3          4          5          7
      8          10         13         14         16
      17         18         19
      OBJETS POUR LESQUELS Y=0
      2          6          9          11         12
      15         20

      2 NIVEAU
         SEGMENT:1
         VARIABLE:X3          ETATS:VILA VA
         OBJETS POUR LESQUELS Y=1
         1          3          4          5          7
         8          10         14         16         17
         19
         OBJETS POUR LESQUELS Y=0
         2          9          15

         3 NIVEAU
            SEGMENT:11
            VARIABLE:X5          ETATS:MARI
            OBJETS POUR LESQUELS Y=1
            10
            OBJETS POUR LESQUELS Y=0
            9          15

         3 NIVEAU
            SEGMENT:12
            VARIABLE:X5          ETATS:CELI DIV  SEPA VEUF
            OBJETS POUR LESQUELS Y=1
            1          3          4          5          7
            8          14         16         17         19

            OBJETS POUR LESQUELS Y=0
            2

      2 NIVEAU
         SEGMENT:2
         VARIABLE:X3          ETATS:ASSE BEAU
         OBJETS POUR LESQUELS Y=1
         13         18
         OBJETS POUR LESQUELS Y=0
         6          11         12         20
```

```
3 NIVEAU
   SEGMENT:21
   VARIABLE:X2            ETATS:INSA
   OBJETS POUR LESQUELS Y=1
   13          18
   OBJETS POUR LESQUELS Y=0
   12

3 NIVEAU
   SEGMENT:22
   VARIABLE:X2            ETATS:SATI BON  EXC
   OBJETS POUR LESQUELS Y=0
   6           11          20
```

IMPRESSION DES TABLEAUX DE CONTINGENCE POUR UN OU PLUSIEURS SEGMENTS ? /O/N/
←⊞→ ?

L'UTILISATEUR PEUT IMPRIMER POUR UN OU PLUSIEURS SEGMENTS, LES TABLEAUX DE CONTINGENCE (EN POURCENTAGE) ENTRE TOUTES LES VARIABLES EXPLICATIVES ET LA VARIABLE EXPLIQUEE Y.

IMPRESSION DES TABLEAUX DE CONTINGENCE POUR UN OU PLUSIEURS SEGMENTS ? /O/N/
←⊞→ O
→ ENTREZ LES NUMEROS DES SEGMENTS CHOISIS ? ←⊞→ 2

EFFECTIFS DU SEGMENT 2 *ET REPARTITION SELON*
LES ETATS DES VARIABLES EXPLICATIVES
LE SEGMENT CONTIENT 6 *UNITES STATISTIQUES*
****

ETATS DE LA VARIABLE EXPLICATIVE X1

	R1	R2	R3	R4	EFFECTIFS
Y =0	25.00	25.00	50.00	0.00	4.00
Y =1	50.00	0.00	0.00	50.00	2.00
TOTA	33.33	16.67	33.33	16.67	6.00

ETATS DE LA VARIABLE EXPLICATIVE X2

	INSA	SATI	BON	EXC	EFFECTIFS
Y =0	25.00	25.00	25.00	25.00	4.00
Y =1	100.00	0.00	0.00	0.00	2.00
TOTA	50.00	16.67	16.67	16.67	6.00

ETATS DE LA VARIABLE EXPLICATIVE X3

	ASSE	*BEAU*	*EFFECTIFS*
Y =0	50.00	50.00	4.00
Y =1	0.00	100.00	2.00
TOTA	33.33	66.67	6.00

ETATS DE LA VARIABLE EXPLICATIVE X4

	X43	*X44*	*EFFECTIFS*
Y =0	50.00	50.00	4.00
Y =1	50.00	50.00	2.00
TOTA	50.00	50.00	6.00

ETATS DE LA VARIABLE EXPLICATIVE X5

	DIV	*SEPA*	*EFFECTIFS*
Y =0	75.00	25.00	4.00
Y =1	50.00	50.00	2.00
TOTA	66.67	33.33	6.00

IMPRESSION DES MEILLEURES SEGMENTATIONS A UN NOEUD ? /O/N/ ←⊞→ ?

L'UTILISATEUR PEUT OBTENIR LES N (NOMBRE CHOISI) MEILLEURES DICHOTOMIES QUI DEFINISSENT UN SEGMENT POUR UN OU PLUSIEURS NOEUDS.
CETTE OPTION PERMET :
- *D'EFECTUER DES ETUDES DE SENSIBILITE SUR LES DICHOTOMIES*
- *DE COMPARER LES DIFFERENCES QUI EXISTENT ENTRE UNE DICHOTOMIE MANUELLE ET AUTOMATIQUE.*

IMPRESSION DES MEILLEURES SEGMENTATIONS A UN NOEUD ? /O/N/ ←⊞→ O
→ ENTREZ LES NUMEROS DES NOEUDS POUR LESQUELS LES MEILLEURES SEGMENTATIONS SONT A CALCULER ? ←⊞→ 0
→ NOMBRE DES MEILLEURES DICHOTOMIES A CALCULER ? ←⊞→ 5

IMPRESSION DES MEILLEURES DICHOTOMIES POUR LE SEGMENT 0
****

VALEUR DU CHI2	*VARIABLE EXPLICATIVE*	*CODES*
3.7781	*X3*	*VILA VA*
3.7781	*X3*	*ASSE BEAU*

VALEUR DU CHI2	VARIABLE EXPLICATIVE	CODES			
2.9670	X1	R1	R2	R3	
2.9670	X1	R4	R5		

VALEUR DU CHI2	VARIABLE EXPLICATIVE	CODES			
1.9005	X1	R1	R2	R3	R4
1.9005	X1	R5			

VALEUR DU CHI2	VARIABLE EXPLICATIVE	CODES		
1.9005	X5	MARI	DIV	SEPA
1.9005	X5	CELI	VEUF	

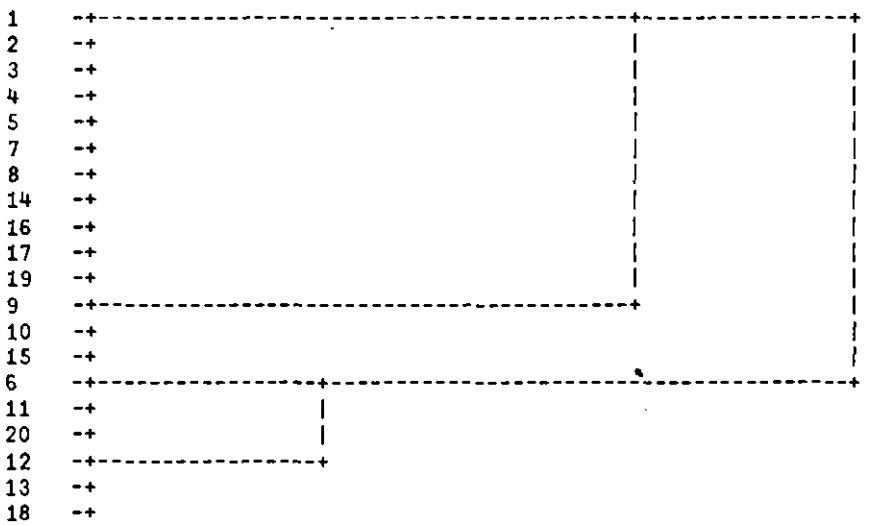
VALEUR DU CHI2	VARIABLE EXPLICATIVE	CODES			
1.8315	X1	R1	R2		
1.8315	X1	R3	R4	R5	

DESSIN DE L'ARBRE ? /O/N/ O

IMPRESSION DE L'ARBRE

****

<-------> *VAUT* 3.3333E0 *INDICE=NOMBRE D'ELEMENTS DE LA CLASSE*



```
SEGMENTATION A UN NOEUD NON ENCORE EXPLORE ? /O/N/ ←⊞→ ?

L'UTILISATEUR PEUT SEGMENTER UN NOEUD NON ENCORE EXPLORE.
VOICI LA LISTE DES NOEUDS NON ENCORE EXPLORES:
11 12 21

SEGMENTATION A UN NOEUD NON ENCORE EXPLORE ? /O/N/ ←⊞→ O
→ NUMERO DU PROCHAIN NOEUD ? ←⊞→ 12
OPTION POUR LA SEGMENTATION ? /AUT/INT/ ←⊞→ INT
         3 NIVEAU
            SEGMENT:12
            VARIABLE: X5          ETATS: CELI DIV  SEPA VEUF
            EFFECTIF:11
            EFFECTIF Y=1 10
            EFFECTIF Y=0 1
            POURC. S.POP:Y=1   90.91
            POURC. S.POP:Y=0    9.09
            POURC. POP:Y=1   50.00
            POURC. POP:Y=0    5.00
            CHI2=4.641

OPTION INTERACTIVE ? /AUT/MAN/MAC/FIN/ ←⊞→ ?

L'UTILISATEUR CHOISIRA POUR LA DICHOTOMIE DU NOEUD IMPRIME CI-DESSUS
UNE DES POSSIBILITES SUIVANTES:
  /AUT/ : RECHERCHE DE LA MEILLEURE DICHOTOMIE CALCULEE AUTOMATIQUEMENT
  /MAN/ : CHOIX IMPERATIF D'UNE VARIABLE EXPLICATIVE, LA RECHERCHE DE LA MEILLEURE
          DICHOTOMIE POUR CETTE VARIABLE EXPLICATIVE EST CALCULEE
          AUTOMATIQUEMENT.
  /MAC/ : CHOIX IMPERATIF D'UNE VARIABLE EXPLICATIVE ET DES ETATS SUR L'UN DES
           SEGMENTS
  /FIN/ : ARRET DE LA SEGMENTATION EN CE NOEUD (POSSIBILITE DE LE REEXPLORER
          PAR LA SUITE)

OPTION INTERACTIVE ? /AUT/MAN/MAC/FIN/ ←⊞→ MAN
→ NOM DE LA VARIABLE EXPLICATIVE ? ←⊞→ X1
            4 NIVEAU
               SEGMENT:121
               VARIABLE: X1          ETATS: R1   R2
               EFFECTIF:2
               EFFECTIF Y=1 1
               EFFECTIF Y=0 1
               POURC. S.POP:Y=1   50.00
               POURC. S.POP:Y=0   50.00
               POURC. POP:Y=1    5.00
               POURC. POP:Y=0    5.00
               CHI2=4.95
```

```
OPTION INTERACTIVE ? /AUT/MAN/MAC/PIN/ <-┘ PIN

LE SEGMENT 122 EST HOMOGENE.

IMPRESSION DES ETAPES DE LA SEGMENTATION ? /O/N/ <-┘ O

REPRESENTATION DE L'ARBRE
**************************

        3 NIVEAU
          SEGMENT:12
          VARIABLE: X5            ETATS: CELI DIV  SEPA VEUP
          EFFECTIF:11
          EFFECTIF Y=1 10
          EFFECTIF Y=0 1
          POURC. S.POP:Y=1  90.91
          POURC. S.POP:Y=0   9.09
          POURC. POP:Y=1  50.00
          POURC. POP:Y=0   5.00
          CHI2=4.641

          4 NIVEAU
            SEGMENT:121
            VARIABLE: X1            ETATS: R1   R2
            EFFECTIF:2
            EFFECTIF Y=1 1
            EFFECTIF Y=0 1
            POURC. S.POP:Y=1  50.00
            POURC. S.POP:Y=0  50.00
            POURC. POP:Y=1   5.00
            POURC. POP:Y=0   5.00
            CHI2=4.95

          4 NIVEAU
            SEGMENT:122
            VARIABLE: X1            ETATS: R3   R4   R5
            EFFECTIF:9
            EFFECTIF Y=1 9
            EFFECTIF Y=0 0
            POURC. S.POP:Y=1 100.00
            POURC. S.POP:Y=0   0.00
            POURC. POP:Y=1  45.00
            POURC. POP:Y=0   0.00
            CHI2=4.95
```

```
LISTE DES ELEMENTS DE CHAQUE SEGMENT ? /O/N/ ←⊞→ O

LISTE DES ELEMENTS DES SEGMENTS
*******************************

        3 NIVEAU
          SEGMENT:12
          VARIABLE:X5          ETATS:CELI DIV  SEPA VEUF
          OBJETS POUR LESQUELS Y=1
          1          3          4          5          7
          8          14         16         17         19

          OBJETS POUR LESQUELS Y=0
          2

          4 NIVEAU
            SEGMENT:121
            VARIABLE:X1          ETATS:R1   R2
            OBJETS POUR LESQUELS Y=1
            3
            OBJETS POUR LESQUELS Y=0
            2

          4 NIVEAU
            SEGMENT:122
            VARIABLE:X1          ETATS:H3   R4   R5
            OBJETS POUR LESQUELS Y=1
            1          4          5          7          8
            14         16         17         19

IMPRESSION DES TABLEAUX DE CONTINGENCE POUR UN OU PLUSIEURS SEGMENTS ? /O/N/
←⊞→ N
IMPRESSION DES MEILLEURES SEGMENTATIONS A UN NOEUD ? /O/N/ ←⊞→ N
DESSIN DE L'ARBRE ? /O/N/ ←⊞→ N
SEGMENTATION A UN NOEUD NON ENCORE EXPLORE ? /O/N/ ←⊞→ N
QUEL BRANCHEMENT ? /EXP/NOM/ETA/OPT/ISE/IEF/IET/IME/ARB/NOE/ ←⊞→ ?

LISTE DES BRANCHEMENTS POSSIBLES DANS LE MODULE DE SEGMENTATION.
1) ON PEUT EXECUTER UNE NOUVELLE SEGMENTATION SUR LA MEME MATRICE DES DONNEES
   AVEC LA POSSIBILITE DE VARIER QUELQUES PARAMETRES

   /EXP/ : CHOIX D'UNE NOUVELLE VARIABLE EXPLIQUEE
   /NOM/ : DEFINIR DES VARIABLES COMME NOMINALES ALORS QU'ELLES ETAIENT
           CONSIDEREES AUPARAVANT COMME ORDINALES (OU INVERSEMENT).
           PARAMETRE NON MODIFIE:
             VARIABLE EXPLIQUEE
```

/ETA/ : DONNER DES NOMS AUX ETATS DE UNE OU PLUSIEURS VARIABLES EXPLICATIVES.
PARAMETRES NON MODIFIES:
VARIABLE EXPLIQUEE: Y
DEFINITION DES VARIABLES COMME NOMINALES ET/OU ORDINALES
/OPT/ : CHOIX D'UNE NOUVELLE OPTION POUR LA SEGMENTATION: AUTOMATIQUE OU PAR INTERVENTION MANUELLE
PARAMETRES NON MODIFIES
VARIABLE EXPLIQUEE:Y
DEFINITION DES VARIABLES COMME ORDINALES ET/OU NOMINALES
NOMS DES ETATS POUR CHAQUE VARIABLE EXPLICATIVE

2) *ON PEUT OBTENIR DES RENSEIGNEMENTS CONCERNANT LA SEGMENTATION AU NOEUD* 12

/ISE/ : IMPRESSION DES ETAPES DE LA SEGMENTATION (ARBRE)
/IEF/ : IMPRESSION DES EFFECTIFS DES SOUS-POPULATIONS DE CHAQUE NOEUD.
/IET/ : IMPRESSION DES TABLEAUX DE CONTINGENCE
/IME/ : IMPRESSION DES MEILLEURES SEGMENTATIONS
/ARB/ : IMPRESSION DU DENDOGRAMME.

3) *ON PEUT EXECUTER UNE SEGMENTATION A UN NOEUD NON ENCORE EXPLORE:*

/NOE/ : CHOIX D'UN NOEUD NON ENCORE EXPLORE.

LISTE DES BRANCHEMENTS DANS D'AUTRES MODULES :
/E/P/A/C/U/P/B/H/W/X/D/M/I/V/T/Q/J/S/

→ EXPLICATIONS SUPPLEMENTAIRES /O/N/ ? ←↵→ N S O

QUEL BRANCHEMENT ? /EXP/NOM/ETA/OPT/ISE/IEF/IET/IME/ARB/NOE/ ←↵→ S O
FIN DE L'EXECUTION DU PACKAGE

L'ANALYSE STATISTIQUE SUR DIFFERENTS JEUX DE DONNEES

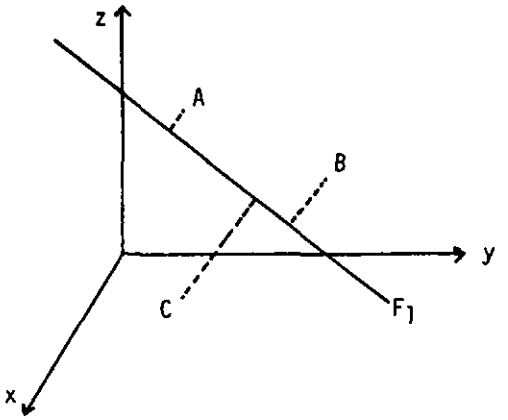
Le chapitre précédent a permis au lecteur de comprendre les principes statistiques des différentes techniques implantées dans CLASFAC et de connaître les options qui ont été retenues. Toutefois nous ne nous sommes pas occupés des liens qui pouvaient exister entre les matrices-input et output des différentes méthodes. Aussi ce chapitre visera-t-il deux buts: le premier est d'apporter une conclusion statistique en montrant dans quel cas les deux grandes familles de l'analyse multivariable, l'analyse factorielle et la classification automatique, peuvent être complémentaires l'une de l'autre, et le deuxième objectif est d'expliquer en détail comment cette complémentarité peut être exploitée techniquement à l'aide des différentes matrices-input et output. A cette fin, le processus de définition (création) des matrices de données sera exposé.
De plus, nous profiterons de ces propos pour expliquer, comment on peut gérer les matrices de données afin de gagner de la place disponible en mémoire centrale.

1. LA COMPLEMENTARITE DES METHODES STATISTIQUES

Nous allons tout d'abord expliquer en quoi et comment la classification automatique et l'analyse factorielle peuvent être complémentaires l'une de l'autre.

1.1. Complémentarité de la classification automatique par rapport à l'analyse factorielle

Lors de l'interprétation des positions des observations (ou des variables) dans l'espace factoriel, nous savons que parfois des points paraissent proches l'un de l'autre, sans qu'ils ne soient pour autant similaires. Cet effet est dû à la projection des points sur des axes, comme l'illustre l'exemple suivant.



Dans l'espace à 3 dimensions la distance entre les points A et B, d(A,B), est plus petite que les distances aux autres points, d(B,C) et d(A,C). Par contre, si nous étudions les projections des points sur la droite F_1, le point B paraîtra plus proche du point C que du point A. Cet effet est dû à la projection des points sur l'axe F_1.

Un moyen simple permet de connaître quelles sont les observations (ou les variables) qui sont véritablement les plus similaires. Nous utiliserons pour ce faire la classification automatique. Par une étude conjointe du plan factoriel et du dendrogramme, l'interprétation de la position des projections des observations (ou des variables) pourra être facilitée. Par exemple, si les projections de deux observations sont proches l'une de l'autre, mais que les résultats d'une classification hiérarchique montrent que la première observation est agrégée tout d'abord avec d'autres observations, puis seulement avec la deuxième observation, l'utilisateur pourra être amené de qualifier leur proximité comme étant accidentelle.

De même, si les projections des observations sur un plan factoriel forment quelques groupes bien distincts, l'utilisateur pourra vérifier ce résultat en effectuant un partitionnement avec le même nombre de groupes.

Comme le montrent ces deux exemples, l'utilisation des méthodes de classification automatique renforce la fiabilité

des résultats de l'analyse factorielle.

1.2. Complémentarité de l'analyse factorielle par rapport à la classification automatique

Les résultats de la classification automatique, à l'exception de la segmentation, fournissent des informations concernant la structure des objets en classes, mais ils ne donnent aucune indication concernant l'influence des variables lors de la création de ces classes.
Un partitionnement restitue les classes les plus homogènes et les plus différenciées les unes des autres. Lors de l'interprétation des résultats, on pourra donc affirmer que les objets qui forment une classe sont les plus apparentés possibles. Par contre on ne peut en aucun cas identifier les variables qui ont permis d'affecter les objets aux classes. Il en va de même des classes qu'on forme quand on coupe un dendrogramme à un certain niveau de son indice.
Par contre, une analyse factorielle permet de déterminer les variables qui influencent la constitution des classes. A la suite d'une classification automatique, on pourra donc exécuter une analyse factorielle et représenter graphiquement les projections des observations et des variables. De plus, il est possible de projeter sur le plan factoriel les centres d'inertie des classes obtenues par un partitionnement et d'étudier ensuite ce plan factoriel.
Au lieu d'effectuer une classification automatique directement sur la matrice des données, on peut commencer par une analyse factorielle et classer ensuite les projections des objets. Cette façon de faire est recommandée chaque fois que la matrice des données contient des variables fortement corrélées.

2. LE PROCESSUS DE DEFINITION DE DONNEES

Pour chaque technique statistique, le chapitre précédent a donné la liste des matrices-input qui peuvent être utilisées et les matrices-output qu'elle génère. Afin de donner une

vue globale et non plus ponctuelle des modules, nous allons résumer le processus de définition des données à l'aide de plusieurs tableaux.
Nous allons tout d'abord donner la liste exhaustive de toutes les matrices de données qui peuvent être traitées ou manipulées par un module de CLASFAC. Le nom mis entre "/" correspond à l'abréviation utilisée dans le dialogue.

/DON/ : matrice des données
/OBS/ : matrice des observations de la matrice des données
/VAR/ : matrice des variables de la matrice des données
/OSU/ : matrice des observations supplémentaires
/VSU/ : matrice des variables supplémentaires
/POB/ : matrice des projections des observations
/PVA/ : matrice des projections des variables
/SOB/ : matrice des projections des observations supplémentaires
/SVA/ : matrice des projections des variables supplémentaires
/IVO/ : matrice des centres d'inertie des observations
/IVV/ : matrice des centres d'inertie des variables
/IFO/ : matrice des centres d'inertie des projections des observations
/IFV/ : matrice des centres d'inertie des projections des variables
/PCO/ : matrice des projections des centres d'inertie des observations
/PCV/ : matrice des projections des centres d'inertie des variables

Remarquons que /OBS/ et /VAR/ sont physiquement identiques à /DON/. Logiquement /OBS/ correspond à la matrice des vecteurs-ligne et /VAR/ à la matrice des vecteurs-colonne de /DON/.

La figure 2.5 illustre le processus de définition des matrices. Seulement les modules statistiques et les modules d'entrée-sortie qui génèrent des matrices-résultat y sont

présentés.
En plus des flux des matrices, ce schéma rend compte de certaines règles d'ordonnancement des appels aux modules.
La légende adoptée pour la taille des matrices est la suivante:

classes : nombre de classes (choisi par l'utilisateur dans le module /P/ ou obtenu par l'algorithme /8/)
var. supp : nombre de variables supplémentaires (choisi par l'utilisateur)
obs. supp : nombre d'observations supplémentaires (choisi par l'utilisateur)
facteurs : nombre de facteurs

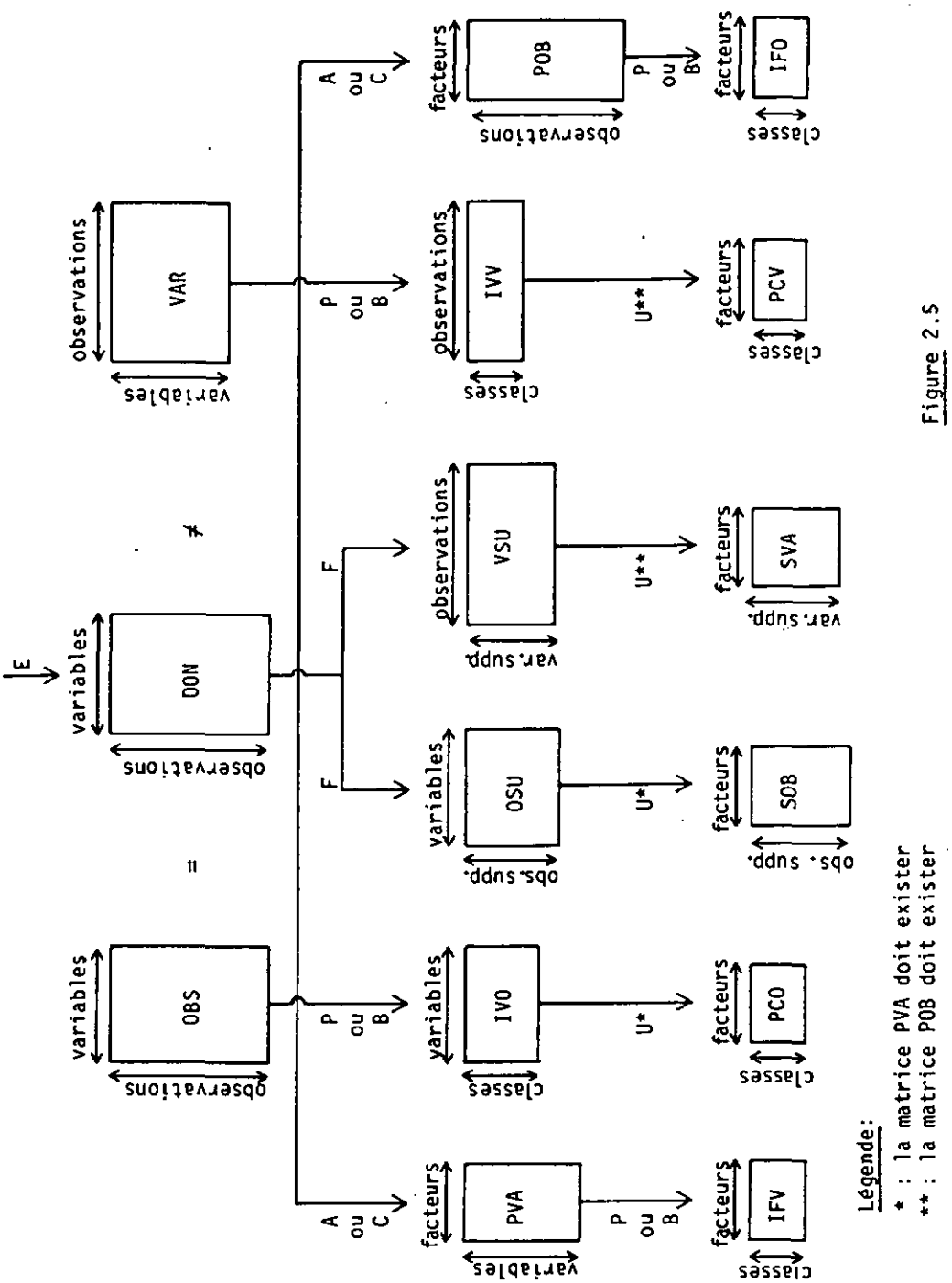

E
DON
observations
variables
≠
=
VAR
observations
variables
OBS
observations
variables
P ou B
A ou C
F
F
PVA
facteurs
variables
P ou B
IVO
variables
classes
U*
PCO
facteurs
classes
IFV
facteurs
classes
OSU
variables
obs. supp.
U*
SOB
facteurs
obs. supp.
VSU
observations
var. supp.
U**
SVA
facteurs
var. supp.
IVV
observations
classes
U**
PCV
facteurs
classes
POB
facteurs
observations
P ou B
IFO
facteurs
classes
Légende:
* : la matrice PVA doit exister
** : la matrice POB doit exister

Figure 2.5

Le tableau 2.6 récapitule les matrices-input qui peuvent être utilisées pour chaque module.

	MODULES ENTREE-SORTIE						ANALYSE STATISTIQUE			
NOMS DES MATRICES	E	F	M	D	I	V	A,C,W	U	P,B	M,X
DON	oui		oui		oui	oui	oui			
OBS				oui					oui	oui
VAR				oui					oui	oui
OSU		oui			oui	oui		oui		
VSU		oui			oui	oui		oui		
POB			oui	oui	oui	oui			oui	oui
PVA			oui	oui	oui	oui			oui	oui
SOB			oui	oui	oui	oui				
SVA			oui	oui	oui	oui				
IVO			oui	oui	oui	oui		oui		oui
IVV			oui	oui	oui	oui		oui		oui
IFO			oui	oui	oui	oui				
IFV			oui	oui	oui	oui				
PCO				oui	oui	oui				
PCV				oui	oui	oui				

Tableau 2.6

Un oui (souligné) correspond soit à la définition soit à une modification (par exemple extension, correction, réduction) de la matrice.

Les tableaux 2.7, 2.8 et 2.9 expliquent le processus de définition de matrice de données lors de l'appel aux différents modules.
La première ligne contient les modules appelés. A chaque module correspondent deux colonnes: la première donne la liste des matrices-input possibles pour le module choisi et la deuxième la matrice-résultat correspondant à la matrice-input sur la même ligne.

PROCESSUS DE CREATION DE LA MATRICE DES DONNEES ET DES MATRICES SUBSIDIAIRES			
E		F	
(1)	(2)	(1)	(2)
a)	DON	DON	OSU
		DON	VSU

Tableau 2.7

Légende:

a) Lors de la création de la matrice des données /DON/ par le module /E/, toutes les matrices sont initialisées, car il serait incohérent de conserver, par exemple, les projections des observations ou des variables obtenues à partir d'une autre matrice de données.

PROCESSUS DE DEFINITION DES MATRICES DANS LES MODULES D'ANALYSE STATISTIQUE									
A,C		P,B		U		H,X		W	
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
a)DDN	PDB	DBS	IVD	DSU	SDB	DBS		DDN	
	PVA	VAR	IVV	VSU	SVA	VAR			
		PDB	IFD	IVD	PCD	PDB			
		PVA	IFV	IVV	PCV	PVA			
						IVD			
						IVV			

Tableau 2.8

Remarque:

Les modules H, X et W ne génèrent pas de matrice-résultat.

Légende:

a) L'appel à un module d'analyse factorielle provoque l'initialisation de toutes les matrices de projections.

PROCESSUS DE DEFINITION DES DONNEES DANS LES MODULES D'ENTREE-SORTIE											
E		F		M		O		I		V	
(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
[a]*DON	DON			[b]*DON	DON			DON	DON	DON	-
		OSU	OSU			OBS	OBS	OSU	OSU	OSU	-
		VSU	VSU			VAR	VAR	VSU	VSU	VSU	-
				POB	POB	POB	POB	POB	POB	POB	-
				PVA	PVA	PVA	PVA	PVA	PVA	PVA	-
				IVO	IVO	IVO	IVO	IVO	IVO	IVO	-
				IVV	IVV	IVV	IVV	IVV	IVV	IVV	-
				IFO	IFO	IFO	IFO	IFO	IFO	IFO	-
				IFV	IFV	IFV	IFV	IFV	IFV	IFV	-
				SOB	SOB	SOB	SOB	SOB	SOB	SOB	-
				SVA	SVA	SVA	SVA	SVA	SVA	SVA	-
						PCO	PCO	PCO	PCO	PCO	-
						PCV	PCV	PCV	PCV	PCV	-

Tableau 2.9

Remarques:

- Pour les modules /E/ et /F/ un nom de matrice souligné signifie que la matrice a été modifiée physiquement (par exemple lors d'une correction, d'une extension, d'une réduction de la matrice).
- Pour le module /M/, un nom de matrice souligné signifie que la matrice a été modifiée logiquement.
- Pour les modules /D/ et /I/ le même nom de matrice apparaît dans la colonne des matrices-résultat, ce qui signifie que la matrice-input n'a subi aucune modification.
- Le trait dans la colonne (2) du module /V/ signifie que la matrice a été détruite.

Légende:

[a] * L'appel au module /E/ dans le but de créer, de corriger, d'étendre ou de réduire la matrice des données, provoque l'initialisation de toutes les autres matrices à l'exception des matrices des observations et des variables supplémentaires.

[b] * La création d'une sous-matrice par le module /M/, définie à partir de la matrice des données provoque l'initialisation de toutes les autres matrices de données à l'exception des matrices des observations et des variables supplémentaires.

3. LES MODULES LIES A LA GESTION DE L'ESPACE DE TRAVAIL

Rappelons que lors de l'exécution de CLASFAC, l'utilisateur dispose d'un espace de travail qui correspond à une zone de la mémoire centrale dont la taille est fixe. Cette zone de travail ne contient en début de session que les fonctions nécessaires à l'exécution du package. En cours de session, l'utilisateur provoque la création de matrices et remplit peu à peu la zone de travail. Dès que la taille de la

mémoire encore diaponible eat insuffiaante pour contenir la ou lea matrices-réaultat générée par un module, l'intarprète APL, envoie le maaaage)WS FULL et l'exécution du package eat arrêtée.
CLASFAC offre deux modules, /R/ et /V/ facilitant la geation de la zone de travail. Le module /R/ permet de connaître la taille de la mémoire centrale encora disponible et la place occupée par chacune des matrices de donnéea. Le module /V/ permet de détruire une matrice.
L'utilisateur est donc invité à appeler périodiquement le module /R/ et lorsque la place encore disponible devient exiguë, il détruira toutes les matrices dont il n'a pas beaoin en appelant le module /V/.

Exemple

Noua allons profiter de cet exemple pour présenter le module de commentaires /J/. Les modules suivanta sont appelés /J/, /R/, /J/, /V/ et /R/.

```
      GO

→ INITIALISATION DE TOUTES LES VARIABLES ? /O/N/ ←⊞→ N
MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ J

*****************************
* MODULE DE COMMENTAIRES /J/ *
*****************************

→ ENTREZ LE COMMENTAIRE ? ←⊞→
?

LA SYNTAXE DES COMMENTAIRES DEPEND DU MODE DU DIALOGUE:
 1) EN MODE INTERACTIF:
    L'UTILISATEUR PEUT INTRODUIRE PLUSIEURS LIGNES DE COMMENTAIRES.
    POUR SORTIR DU MODULE, IL DOIT TAPER DEUX FOIS '' RETOUR DE CHARIOT ''
     (LIGNE VIDE)

 2) EN MODE PSEUDO-BATCH:
    LA SYNTAXE EST :
        [ TEXTE1 / TEXTE2 / TEXTE3 / ]
        AVEC TEXTE1 TEXTE2   : COMMENTAIRES
                    /        : CE SIGNE INDIQUE LE PASSAGE A LA LIGNE
                               (SUPERPOSITION DES SIGNES APL /-)
    OU  [0] SI L'UTILISATEUR VEUT REPRENDRE L'ANCIEN COMMENTAIRE
```

```
REMARQUE:
LE SIGNE ' EST INTERDIT DANS LES COMMENTAIRES

EXEMPLE
*******
   /R/ : MODULE DE PLACE DISPONIBLE
   /V/ : MODULE D INITIALISATION


MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←┘ R

*********************************
*MODULE DE PLACE DISPONIBLE /R/*
*********************************


LA PLACE RESTANTE EN MEMOIRE CENTRALE EST DE  401648 OCTETS LE 27.12.1982 3H 41

LES JEUX DE DONNEES SUIVANTS SONT DEFINIS:
LES DONNEES OCCUPENT 736 OCTETS
LES OBSERVATIONS SUPPLEMENTAIRES OCCUPENT 172 OCTETS
LES CENTRES D'INERTIE DES OBSERVATIONS OCCUPENT 432 OCTETS
LES PROJECTIONS DES OBSERVATIONS OCCUPENT 548 OCTETS
LES PROJECTIONS DES VARIABLES OCCUPENT 652 OCTETS
LES PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES OCCUPENT 140 OCTETS

MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←┘ J [0]

*******************************
* MODULE DE COMMENTAIRES /J/ *
*******************************

  EXEMPLE
  *******
     /R/ : MODULE DE PLACE DISPONIBLE
     /V/ : MODULE D INITIALISATION

MODULE CHOISI ? /E/F/A/C/U/P/B/H/X/W/D/M/I/V/T/Q/R/J/S/?/ ←┘ V

*******************************
*MODULE D'INITIALISATION /V/ *
*******************************

QUELLE MATRICE ? /DON/OSX/VSX/POB/PVA/SOB/SVA/IVO/IVV/IFO/IFV/PCO/PCV/MIN/ ←┘ ?
```

```
LISTE DE TOUTES LES MATRICES QUI PEUVENT ETRE EFFACEES:
  /DON/ : MATRICE DES DONNEES
  /POB/ : PROJECTIONS DES OBSERVATIONS
  /PVA/ : PROJECTIONS DES VARIABLES
  /IVO/ : CENTRES D'INERTIE DES OBSERVATIONS
  /IVV/ : CENTRES D'INERTIE DES VARIABLES
  /IFO/ : CENTRES D'INERTIE DES COMPOSANTES DES OBSERVATIONS
  /IFV/ : CENTRES D'INERTIE DES COMPOSANTES DES VARIABLES
  /PCO/ : PROJECTIONS DES CENTRES D'INERTIE DES OBSERVATIONS DE LA MATRICE DES DONNEES
  /PCV/ : PROJECTIONS DES CENTRES D'INERTIE DES VARIABLES DE LA MATRICE DES DONNEES
  /SOB/ : PROJECTIONS DES OBSERVATIONS SUPPLEMENTAIRES
  /SVA/ : PROJECTIONS DES VARIABLES SUPPLEMENTAIRES
  /MIN/ : NOMS DES VARIABLES EXPLIQUATIVES

  /OSX/ : OBSERVATIONS SUPPLEMENTAIRES ET PROJECTIONS DES OBSERVATIONS
          SUPPLEMENTAIRES
  /VSX/ : VARIABLES SUPPLEMENTAIRES ET PROJECTIONS DES VARIABLES
          SUPPLEMENTAIRES

QUELLE MATRICE ? /DON/OSX/VSX/POB/PVA/SOB/SVA/IVO/IVV/IFO/IFV/PCO/PCV/MIN/ ←⊞→ IVO *
POB PVA SOB

MATRICE IVO EFFACEE


MATRICE POB EFFACEE


MATRICE PVA EFFACEE


MATRICE SOB EFFACEE

QUEL BRANCHEMENT ? /DON/OSX/VSX/POB/PVA/SOB/SVA/IVO/IVV/IFO/IFV/PCO/PCV/MIN/
←⊞→ R

*********************************
*MODULE DE PLACE DISPONIBLE /R/*
*********************************


LA PLACE RESTANTE EN MEMOIRE CENTRALE EST DE  402172 OCTETS LE 27.12.1982 3H 42

LES JEUX DE DONNEES SUIVANTS SONT DEFINIS:
LES DONNEES OCCUPENT 736 OCTETS
LES OBSERVATIONS SUPPLEMENTAIRES OCCUPENT 172 OCTETS

MODULE CHOISI ? /E/F/A/C/U/P/B/R/X/W/D/M/I/V/T/Q/R/J/S/?/ ←⊞→ S O
FIN DE L'EXECUTION DU PACKAGE
```

QUELQUES ASPECTS DE LA REALISATION INFORMATIQUE

CLASFAC a été écrit en APL et sa réalisation a été influencée en partie par les propriétés et les possibilités de ce langage.
Nous allons seulement présenter les aspects les plus importants qui rendent compte de la conception informatique de CLASFAC et de la façon de résoudre un problème au moyen du langage APL.
Tout au long de cette présentation, nous essayerons de mettre l'accent sur les problèmes que nous avons dû résoudre pour premièrement développer un package destiné à un utilisateur dépourvu de connaissances informatiques et, deuxièmement, pour gérer au mieux la place mémoire disponible dans la zone de travail.

1. LA STRUCTURE DU PACKAGE

CLASFAC est conçu de manière modulaire. Un module est une fonction APL réalisant une opération ou un algorithme statistique sur un (ou plusieurs) jeux de données.
Chaque module est écrit selon le principe descendant, c'est-à-dire, qu'il se divise hiérarchiquement en sous-fonctions. On accède à un module par le module de guidage qui correspond à la fonction GO. Dans le dialogue l'appel à un module correspond à une question de type 1 (réponse réduite à une lettre).
Tout le dialogue se situe au niveau des modules à l'exclusion donc des sous-fonctions.
Cette structure a été retenue pour diverses raisons. Nous allons en citer quelques-unes.

- La souplesse du package lors de l'exécution: il est possible d'appeler en tout temps à la suite d'une question de type 1 n'importe quel module, car le dialogue se situe uniquement au niveau de ceux-ci.

- La maintenance du package est facilitée. Une modification du package n'aura d'incidence que sur un seul module. De même il est facile d'insérer un nouveau module, car seule le module de guidage et quelques modules utilitaires, tels le module MENU dynamique et TRACE, devront être complétés.

- Mémoire centrale versus temps de calcul
  Si d'ici une dizaine d'années, la place de la mémoire centrale disponible deviendra illimitée et le temps de calcul nécessaire à l'exécution d'opérations de plus en plus court, le grand problème auquel nous avons toujours été confronté aura disparu. Il s'agit du dilemne bien connu: "occupation de mémoire centrale versus temps de calcul" ou autrement dit: tout moyen qui permet d'augmenter la place de mémoire centrale disponible dans la workspace a pour effet d'augmenter le temps de calcul (car des opérations supplémentaires devront être exécutées). La version segmentée est une illustration de ce principe. Sa réalisation a été grandement facilitée par la structure modulaire de CLASFAC. Elle est basée sur le principe suivant: le résident permanent est constitué des fonctions appelées par le module de guidage et certaines autres fonctions utilitaires appelées fréquemment par tous les modules. Ce résident permanent est conservé dans la workspace. Les autres fonctions sont mémoriaées sur disque. L'ensemble des fonctions nécessaire à l'exécution d'un module sont réunies en un segment. Lors de l'appel à un module le segment qui lui correspond est chargé en zone de travail. Dès que le module appelé est "déactivé" on libère de nouveau la place rendue disponible pour y charger un autre segment.
  Comme il n'existe aucune fonction standard APL ou commande système permettant de charger un segment pendant l'exécution du package, nous avons développé nous-mêmes les fonctions qui génèrent, gèrent et chargent les différents segments.

2. ASPECIS DE LA REALISATION INFORMATIQUE BASES SUR LA PROPRIETE D'ALLOCATION DYNAMIQUE EN APL

Le langage APL offre la possibilité de l'allocation dynamique de la mémoire. Ce concept permet d'économiser de la place mémoire dans la workspace par le fait que les variables ne deviennent effectives qu'au moment où on les affecte. Ainsi, aucune place mémoire ne doit être réservée au début des fonctions APL en vue d'une affectation possible. Cette propriété permet d'utiliser le minimum de place dans la workspace, car seule la place effectivement utilisée par les variables définies pour une application particulière est retenue. De plus, tous les problèmes de sous-dimensionnement ou de sur-dimensionnement, que l'on rencontre par exemple en FORTRAN disparaissent.
L'allocation dynamique permet donc au programme de s'adapter à n'importe quel volume de données et la matrice des données peut donc avoir théoriquement une dimension quelconque. Cependant CLASFAC doit être exécuté dans une workspace qui possède une taille mémoire fixée à l'avance.

Illustration de la version segmentée:

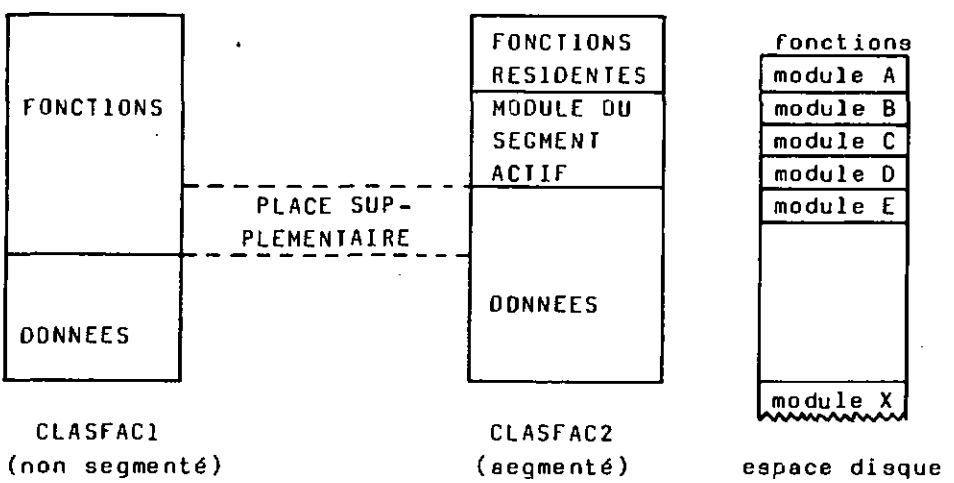


Figure 3.1

Dans une telle workspace active la place en mémoire disponible pour les données est égale à la place restante quand on a soustrait à la taille de la workspace la place utilisée par les fonctions. La version segmentée permet à l'utilisateur de disposer de plus de place pour les données et tout se passe pour lui comme s'il travaillait dans une workspace plus grande. La figure 3.1 illustre le mécanisme de segmentation du programme.

2.1. Le traitement des différents jeux de données

Nous allons expliquer de quelle façon nous avons utilisé l'allocation dynamique pour traiter plusieurs jeux de données.
Comme nous l'avons déjà expliqué ci-dessus, une variable ne devient effective qu'au moment où elle est affectée. L'utilisation (dans une expression) par exemple d'une variable non affectée provoque l'envoi du message d'erreur, ")VALUE ERROR", par l'interprète et l'interruption de l'exécution de CLASFAC. Un tel arrêt ne peut être toléré dans un package destiné à des utilisateurs dépourvus de connaissances informatiques. Dès lors, la première opération qui doit être exécutée est l'initialisation de toutes les variables qui forment les différents jeux de données. En APL on réalise cette opération en définissant des variables vides. Celles-ci occupent une place de mémoire minimum. Ainsi, un tableau de données sera une matrice vide formée de 0 ligne et de 0 colonne.
Des valeurs ne sont effectivement affectées aux variables que lors d'un appel au module.
Une variable ou un jeu de données peut donc avoir trois états possibles: elle n'est pas définie, elle est définie mais vide, elle possède une valeur et est donc affectivement définie.
Le langage distingue le premier cas des deux autres (émission du message ")VALUE ERROR" et arrêt), alors que du point de vue logique il s'agit de différencier les deux premiers cas, qui constituent en fait deux réalisations possibles d'une même situation et le dernier.

Pour réaliser ce but, nous avons associé à chaque jeu de données une fonction booléenne qui vaut 1 respectivement 0 selon que le jeu de données est ou n'est pas effectivement défini. De plus cette fonction vérifie si le jeu de données est compatible avec la matrice /DON/, car au moins une de leurs dimensions est commune, comme nous l'avons déjà vu (cf. figure 2.5).
La principe de définition des différents jeux de données forme la pierre angulaire de 2 modules utilitaires: le module /Q/ (MENU dynamique) et le module /R/ (taille de la mémoire encore disponible).
Rappelons que le module de MENU dynamique permet d'obtenir la liste des modules et des matrices-input que l'utilisateur peut appeler et traiter, compte tenu des opérations et des analyses déjà effectuées.
Le module MENU a été réalisé de la façon suivante. Ce module connaît la liste de tous les modules implantés dans CLASFAC et, pour chaque module, toutes les matrices-input possibles. De plus on lui a fourni, pour chaque module, la liste des matrices de données nécessaires à son exécution (règles de cohérence statistique).
Le module MENU procède alors de la façon suivante: il parcourt la liste des modules de CLASFAC et teste pour chacun d'eux s'il vérifie les règles de cohérence statistique; dans le cas où elles sont respectées; il imprime le nom du module et parmi l'ensemble des matrices-input qui peuvent être traitées, celles qui sont définies; dans le cas contraire, l'analyse du module est abandonnée.

Le module /R/ permet de connaître la taille de la mémoire encore disponible et la place occupée par chaque jeu de données. Il agit de la même façon que le module MENU. Il parcourt la liste de tous les jeux de données qui peuvent être définis dans CLASFAC et lorsqu'il rencontre un jeu de données qui est défini, il calcule la place occupée par celui-ci et l'imprime.

En tout temps, un jeu de données peut être réinitialisé à vide, afin de récupérer de la place. Cette opération est

effactuée par le module /V/.

2.2. Le traltement des réponses aux questions

CLASFAC se déroule eous le forme d'un dialogue entre l'utiliaateur et l'ordinateur. Nous allons expliquer les aspects d'échange d'informationa en deux étapes. D'abord, nous développerons le principe du traitement des réponses entrées par l'utilisateur, puis nous aborderons quelques problèmes soulevés par la validation des réponses.

2.2.1. Le principe du trsitement des réponses

En plus des explications que noue allons donner, le lecteur peut se référer aux organigrammea des figures 3.2 et 3.3 qui donnent une vue aynthétique du processus du traitement des réponsea.
Lea réponsee entrées à la euite d'une question de type 1, réponses données soit en mode intersctif soit en mode paeudo-bstch, sont représentées dans le package sous la forme d'une liste, dans laquelle chaque élément correspond à une réponse complète à une queation. Le nom de cette liste s'ppelle PREP. Deux éléments de la liete sont séparés par un espace.
Nous allons expliquer le processus du traitement dea réponses en le décomposant en 4 phaes.

Première phsse : Entrée d'une liate de réponaes à la suite d'une question de type 1

A la suite d'une queation de type 1, la liste PREP a la forme suivante:

PREP = (REP0, REP1, REP2, ..., REPn)

Le premier élément de la liate PREP représente la réponae à la qusstion coursnte de type 1 et les sutres, lea réponses anticipées qui peuvent être relstives à une queation de type 1 ou 2.

Deuxième phase : Extraction de la réponse à la question courante

La réponse à la question courante est extraite de la liste PREP et une nouvelle sous-liste REP contenant la réponse à la question courante est créée.

PREP := {REP0, REP1, REP2, ..., REPn}

REP := {REP0}

PREP := {REP1, REP2, ..., REPn}

Si l'on n'a pas utilisé le mode pseudo-batch, la liste PREP initiale ne contient que la réponse REP0 et à la suite de l'extraction la liste PREP est donc vide.

PREP = ∅

REP := {REP0}

Troisième phase : Analyse syntaxique et sémantique de la réponse

A ce stade, la liste REP est analysée, afin de vérifier si son contenu est syntaxiquement et sémantiquement correct. Si ce test de validation est négatif, la liste des réponses anticipées PREP doit être détruite et la question reposée.

Quatrième phase : Epuisement de la liste PREP

A l'aide de la réponse correcte REP le package est exécuté jusqu'à la prochaine question. Si la liste PREP n'est pas vide, l'extraction du premier élément de PREP est alors effectué (deuxième phase) et la réponse est validée (troisième phase). Si la liste PREP est vide, on pose la prochaine question.

A la suite d'une question de type 2 (le mode pseudo-batch est alors exclu), PREP est vide et seul REP contient la réponse courante qui doit être analysée syntaxiquement et sémantiquement.

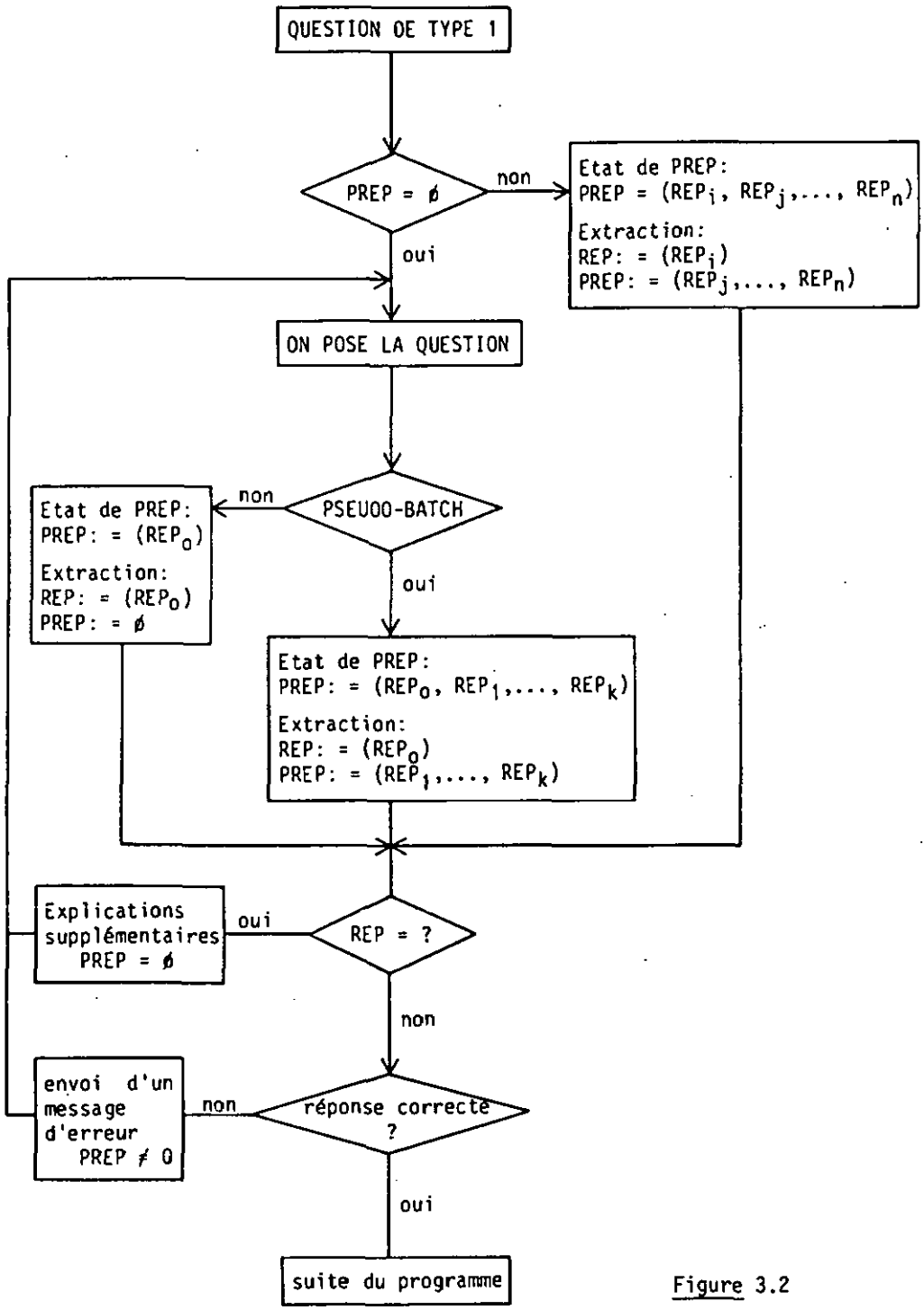

QUESTION DE TYPE 1
PREP = ∅
non
Etat de PREP:
PREP = (REP$_i$, REP$_j$,..., REP$_n$)
Extraction:
REP: = (REP$_i$)
PREP: = (REP$_j$,..., REP$_n$)
oui
ON POSE LA QUESTION
PSEUDO-BATCH
non
Etat de PREP:
PREP: = (REP$_0$)
Extraction:
REP: = (REP$_0$)
PREP: = ∅
oui
Etat de PREP:
PREP: = (REP$_0$, REP$_1$,..., REP$_k$)
Extraction:
REP: = (REP$_0$)
PREP: = (REP$_1$,..., REP$_k$)
REP = ?
oui
Explications supplémentaires
PREP = ∅
non
réponse correcte ?
non
envoi d'un message d'erreur
PREP ≠ 0
oui
suite du programme

Figure 3.2

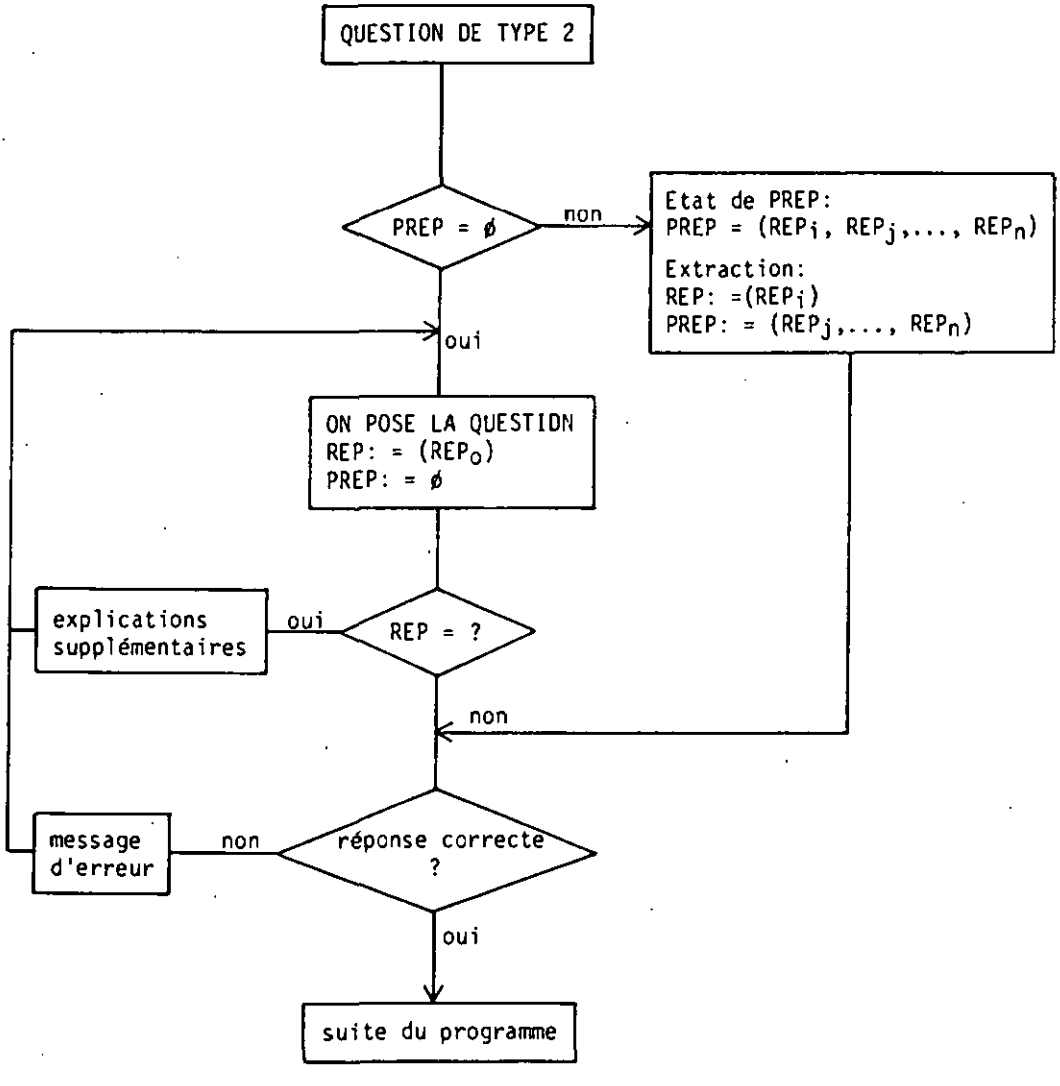

QUESTION DE TYPE 2
PREP = ∅
non
Etat de PREP:
PREP = (REPi, REPj,..., REPn)
Extraction:
REP: =(REPi)
PREP: = (REPj,..., REPn)
oui
ON POSE LA QUESTION
REP: = (REPo)
PREP: = ∅
explications supplémentaires
oui
REP = ?
non
message d'erreur
non
réponse correcte ?
oui
suite du programme

Figure 3.3

2.2.2. La validation des réponses

Pour chaque réponse relative à une question, nous avons vu qu'un contrôle de validité de la réponse est effectué. La validité d'une réponse se fait en deux phases. Lors de la première on contrôle la syntaxe et lors de la deuxième la sémantique. Nous n'allons pas entrer dans les détails de ces deux contrôles, puisque nous en avons déjà parlé dans la première partie. Par contre, il nous paraît important de souligner quels sont les points délicats d'une telle démarche. Le but d'une analyse syntaxique et sémantique est de détecter une erreur d'entrée avant que le paramètre ne soit transmis pour être utilisé par le package. Remarquons que si un paramètre erroné est transmis au package, celui-ci perd totalement le contrôle de la situation, car lors de l'utilisation du paramètre, l'interprète va détecter une erreur, envoyer un message et arrêter l'exécution. C'est pourquoi, lors du développement d'un package destiné à des utilisateurs non informaticiens, il faut toujours essayer de prévoir pour chaque question l'ensemble des réponses erronées qui peuvent être entrées.
La recherche de cet ensemble d'erreurs possibles est une opération délicate, car le programmeur doit imaginer les réponses les plus inattendues. Nous allons, par un exemple, essayer d'analyser les erreurs possibles à la question:

"-ENTREZ LE NOM DE LA VARIABLE QUI CONTIENT LA MATRICE DES DONNEES ET QUI EST STOCKE DANS LA WORKSPACE :"

Pour cette simple question, nous pouvons déjà tester quatre erreurs possibles:

- l'utilisateur a entré plus d'un nom
- le nom entré ne correspond à aucune variable définie dans la workspace
- la variable n'est pas un tableau
- la variable n'est pas de type numérique

En général, il n'est pas possible au programmeur de prévoir toutes les erreurs d'entrée, et il est dès lors recommandé de demander à un ou plusieurs utilisateurs d'expérimenter le package. L'ensemble de leurs remarques et erreurs commises devront être alors prises en considération pour affiner le

package. Signalons pour terminer que si CLASFAC est un package volumineux, cela n'est pas dû uniquement aux algorithmes statistiques et aux opérations de manipulation de données, mais également aux nombreux garde-fous qu'il a fallu prévoir afin de traiter l'ensemble des erreurs d'entrée possibles.

2.3. Le module TRACE

L'information nécessaire au module TRACE permettant de restituer l'historique d'une session de travail, est mémorisée dans une liste. Cette dernière s'appelle PILE. Au début du package cette variable est initialisée comme liste vide:

PILE = ∅

Puis lors de l'appel à un module, la lettre qui caractérise le module et le nom de la matrice-input utilisé par le module sont concaténés à cette liste.

PILE := PILE + {module, matrice}

A la sortie du module, un caractère blanc est concaténé à la variable:

PILE := PILE + {' '}

L'appel consécutif à plusieurs modules a pour effet d'augmenter la taille de la liste. Le module TRACE restitue l'historique de la session par un simple décodage de la liste PILE.

3. LA REALISATION INFORMATIQUE BASEE SUR LES PROPRIETES D'UN LANGAGE INTERPRETATIF

Un langage interprétatif diffère d'un langage compilé par le fait que les instructions ne sont compilées qu'au moment de

leur exécution.
Comparativement à un langage compilé, un langage interprétatif se révèle lent et gourmand en temps de calcul, cependant il permet d'économiser un temps précieux lors de la mise aux points des programmes (debugging), car on peut interrompre toute fonction sur une instruction quelconque, puis on peut demander l'affichage des valeurs prises par certaines variables, leur donner de nouvelles valeurs et relancer l'exécution.
En plus de cette facilité de mise au point, APL permet de "calculer" le nom d'une fonction à appeler. Nous allons expliciter cette possibilité dans le paragraphe suivant et décrire l'usage que nous en avons fait.

3.1. La fonction HELP

Plusieurs possibilités existent pour gérer l'ensemble des explications supplémentaires relatives aux différentes questions. Il est possible par exemple de créer une matrice de caractères. Lors de la requête à une explication supplémentaire, une ou plusieurs lignes de la matrice seraient imprimées. Cette façon de faire soulève un problème de maintenance, car l'adjonction ou l'effacement d'une ligne exige une restructuration de la matrice si l'on veut minimiser la place occupée en mémoire.
La solution que nous avons retenue est basée sur le fait que la source APL est interprété. A chaque explication supplémentaire correspond une fonction APL dont le nom commence par les lettres EXP. Lorsqu'une question est posée à l'utilisateur, le package appelle une fonction qui possède deux paramètres implicites (paramètres d'entrée). Le premier correspond au texte de la question et le deuxième aux lettres qui suivent EXP et qui spécifient ainsi la fonction contenant l'explication supplémentaire relative à la question. Par exemple, pour une question du module d'analyse factorielle en composantes principales, ces deux paramètres sont: "FORME DE LA MATRICE DES DONNEES /BRU/CEN/CRE/ ? :" et "A1". Si l'utilisateur répond par un point d'interrogation, le deuxième paramètre est concaténé à EXP, ce qui donne dans

notre exemple EXPA1. Par l'opérateur APL, noté ⍎ , cette chaîne de caractères va être exécutée et provoquera ainsi l'exécution de la fonction, EXPA1.

3.2. Les fonctions qui gèrent les diverses matrices de données

Nous allons montrer, à quel point il est facile de gérer les différents jeux de données en recourant au même principe que celui utilisé pour réaliser la fonction HELP.
Dans le dialogue, les noms des différents jeux de données sont abrégés à l'aide de 3 lettres (exemples DON, POB, ...). Lorsque l'utilisateur répond à la question "QUELLE DONNEE ? /OBS/VAR/POB/PVA/ :", et si la réponse est correcte, la liste REP contient les 3 lettres qui désignent un jeu de données susceptible d'être utilisé par le module.
Il existe autant de fonctions commençant par les lettres EXIST, AFF, et INI qu'il y a de jeux de données susceptibles d'être utilisés dans le package. Ces fonctions permettent de tester si un jeu de données est défini, d'affecter le jeu de données à des variables de travail et d'initialiser un jeu de données.
Si par exemple, la réponse entrée à la suite de la question: "QUELLE DONNEE ?" est OBS, la fonction qui teste l'existence du jeu de données défini est la fonction EXISTOBS et celle qui affecte les variables à des variables de travail est AFFOBS. Si on décide d'initialiser le jeu de données POB, la fonction INIPOB est exécutée.

3.3. Le branchement en cas de réponses erronées

Si la réponse à une question de type 1 est erronée, un message d'erreur doit être envoyé et il faut se rebrancher sur la question qui doit être reposée. Il est donc théoriquement nécessaire d'écrire deux instructions à suite de chaque question de type 1.
Comme il existe dans CLASFAC plus de 120 questions de type 1, ceci augmenterait considérablement la taille du source des fonctions.

Ces redites ont été évitées par l'écriture d'une seule fonction à laquelle on transmet l'adresse du branchement.

3.4. Le problème soulevé par l'utilisation d'opérateurs très puissants

Lors de la présentation du langage APL, nous avons évoqué que celui-ci possède des opérateurs très puissants. Un d'entre eux permet d'effectuer le produit matriciel de deux matrices A et B. Cette instruction s'écrit A+.xB. Le produit matriciel est utilisé dans CLASFAC entre autres pour calculer la matrice des variances-covariances associée à une matrice de données ainsi que la matrice des distances entre ses objets. Si A et B sont des matrices de grandes tailles, il est clair que les résultats intermédiaires nécessaires à l'exécution de l'instruction A+.xB sont également des matrices de grandes tailles. Lors du calcul d'un produit matriciel, il est donc possible que la place mémoire disponible dans la workspace devienne insuffisante et l'exécution du package est alors interrompue par le "fatidique" message d'erreur, ")WS FULL".
Afin de conserver le contrôle sur la place mémoire disponible dans la workspace, nous avons été obligés dans de nombreux cas de recourir à une programmation moins "élégante" et de ne pas utiliser toute la puissance des opérateurs APL.

L'ANALYSE FACTORIELLE

1. BENZECRI, J.P. et Collaborateurs, L'Analyse des Données, Vol. 2, Dunod 1973.

2. BENZECRI, J.P. et Collaborateurs, Pratique de L'Analyse des Données, Vol. 3, Dunod, 1981.

3. LEBART, L., FENELON, J.P., Statistique et Informatique Appliquées, Dunod, 1975.

4. LEBART, L., MORINEAU, A. et TABARD, N., Techniques de la Deacription Statistique, Dunod, 1977.

5. STROHMEIER, A. Analyse du Questionnaire Migrations dans Le Centre-Jura, Cahiers de Méthodes quantitatives, Université de Neuchâtel, 1975.

6. STROHMEIER, A. et MEMMINGER, L., L'Analyse Factorielle, Publication interne, 1977.

7. STROHMEIER, A., L'Analyse Factorielle des Correspondances. Les champs d'application de la méthode et la codification des données, Cahiers de Méthodes Quantitatives, Université de Neuchâtel, 1975.

8. STROHMEIER, A., L'Analyse Factorielle, Aspects Mathématiques, Cahiers de Méthodes Quantitatives, Univeraité de Neuchâtel, 1977.

LA CLASSIFICATION AUTOMATIQUE

9. ANDERBERG, M.R., Cluster Analysis for Applications, Academic Press, 1973.

10. BENZECRI, J.P. et Collaborateurs, L'Analyse des Données, Vol I : la Taxinomie, 1, Dunod, 1973.

11. BOCK, H.H., Automatiache Klaaaifikation, Vandenhoeck & Ruprecht Göttingen, 1974.

41. BOSS, J.f., La Segmentation du Marché: le point de départ et des techniques, Revue Française du Marketing, Troisième Trimestre 1973, Cahier 4B.

12. CHANDON, J.L. et PINSON, S., Analyae typologique, Maaaon, 1981.

13. DURAN, B.S. et ODELL, P.L., Cluater Analyaia, Springer-Verlag, 1974.

14. EVERITT, B., Cluater Analyaia, Heinemann Ltd, 1974.

15. GRAF-JACOTTET, M., Claaaification Automatique, Aapecta Mathématiquea, Cahiera de Méthodea Quantitativea, Univerait de Neuchâtel, 1979.

16. HARTIGAN, J.A., Clustering Algorithma, John Wiley & Sona, 1975.

17. HUGUES, H., GRIFFON, B. et BOUVEYRON, C., Segmentation et Typologie, Bordaa, 1969.

18. JAMBU, M. et LEBEAUX, M.O., Claaaification Automatique pour l'Analyae daa Donnéea, Vol 1 et 2, Dunod, 1978.

19. JARDINE, N. et SIBSON, R., Mathematical Taxonomy, John Wiley & Sona, 1971.

20. De RHAM, C., La Claaaification Hiérarchique Aacendante et Deacendante de Granda Tableaux de Donnéea, Thèae Univeraité da Neuchâtel, 1980.

21. SNEATH, P.H.A. et SOKAL, R.R., Numerical Taxonomy, Freeman and Company, 1973.

22. SPAETH, H., Cluater-Analyae-Algorithmen, R. Oldenbourg

Verlag, 1975.

23. STEINHAUSEN, D. et LANGER, K., Clusteranalyse, Walter de Gruyter, 1977.

24. STROHMEIER, A., Classification Automatique, les Bases, Cahiers de Méthodes Quantitatives, Université de Neuchâtel, 1978.

25. WILLIAMS, W.T. et LANCE, G.N., Hierarchical Classificatory Methods, John Wiley & Sons, 1977.

L'ANALYSE FACTORIELLE ET LA CLASSIFICATION AUTOMATIQUE

26. BERTIER, P. et BOUROCHE, J.M., Analyse des Données Multidimensionnelles, Presses Universitaires de France, 1975.

27. CAILLIEZ, F. et PAGES, J.P., Introduction à l'Analyse des Données, Smash, 1976.

28. CHARDON, P.A., Méthodes pratiques de dépouillement de questionnaires, Thèse Université de Neuchâtel, 1981.

29. EVRARD, Y. et LE MAIRE, P., Information et Décision en Marketing, Dalloz, 1976.

30. GUIGOU, J.L., Méthodes Multidimensionnelles, Dunod, 1977.

31. HENRY-LABORDERE, A., Analyse de Données, Masson, 1977.

32. LEBART, L., MORINEAU, A. et FENELON J.P., Traitement des Données statistiques, Dunod, 1979.

33. VOLLE, M., Analyse des Données, Economica, 1978.

APL ET L'ANALYSE DES DONNEES

34. MAILLES J.P., MAILLES, D. et BONNEFOUS, S., Analyse des Données et APL, Commissariat à l'énergie atomique, 1976.

35. STROHMEIER. A., VOIDE., A.C., CLASFAC-APL: an Interactive Package for Multivariate Data Analysis, Compstat 1980, Physica-Verlag, Wien, p. 228-233.

LES MANUELS APL DE REFERENCE

36. GILMAN, L., et ROSE, A.J., APL an Interactive Approach, John Wiley and Sons, 1976.

37. LEGRAND, B., Apprendre et Appliquer le Langage APL, Masson, 1979.

38. POLIVKA, R.P. et PAKIN, S., APL the Language and its Usage, Prentice-Hall, 1975.

39. POMMIER, S., Introduction à APL, Dunod, 1978.

40. ROBINET, B., Le Langage APL, Editions Technip, 1971.

A

B

C

G

H

I

J

K

REMERCIEMENTS

Je remercie sincèrement M. le Professeur A. Strohmeier directeur de cette thèse pour ses conseils avisés, ses encouragements et sa grande disponibilité.

J'exprime ma reconnaisance à Messieurs les Professeurs P. Banderet et J. Kohlas pour les discussions instructives qu'ils m'ont fait bénéficier.

Mes remerciements vont également au Dr P.-A. Chardon pour ses nombreux conseils.

J'adresse mes remerciements encore à tous ceux qui ont contribués à ce travail, particulièrement :

- au Dr F. Grize qui m'a mis à ma disposition des programmes de traitement de textes.
- à M. P. Scherrer et M. D. Schulthess par leur lecture critique de cette thèse.
- Mmes M. Badan et S. Jobin pour leur aide dans divers travaux de bureau.
- à mon mari pour sa précieuse collaboration et sa patience sans limite.

Enfin je remercie IBM Suisse qui m'a permis de réaliser ce travail en me mettant à disposition un terminal et du temps de calcul; sans cet appui ce travail n'aurait jamais vu le jour.