Contribution à la classification de données binaires et qualitatives

Franck Marchetti

Laboratoire de Recherche en Informatique de Metz

THESE

présentée à

L'UNIVERSITE DE METZ

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE METZ

(mention sciences)

SPECIALITE INFORMATIQUE

par

Franck MARCHETTI

CONTRIBUTION A LA CLASSIFICATION DE DONNEES BINAIRES ET QUALITATIVES

Soutenue le 15 décembre 1989 devant la commission d'examen

messieurs

G. CELEUX (rapporteur), Chargé de Recherche à l'INRIA

E. DIDAY (rapporteur), Professeur à l'Université de Paris IX

Y. GARDAN (examinateur), Professeur à l'Université de Metz

G. GOVAERT (directeur de thèse), Professeur à l'Université de Metz

G. LE CALVE (examinateur), Professeur à l'Université de Rennes II

Laboratoire de Recherche en Informatique de Metz

THESE

présentée à

L'UNIVERSITE DE METZ

pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITE DE METZ

(mention sciences)

SPECIALITE INFORMATIQUE

par

Franck MARCHETTI

CONTRIBUTION A LA CLASSIFICATION DE DONNEES BINAIRES ET QUALITATIVES

Soutenue le 15 décembre 1989 devant la commission d'examen

messieurs

G. CELEUX (rapporteur), Chargé de Recherche à l'INRIA

E. DIDAY (rapporteur), Professeur à l'Université de Paris IX

Y. GARDAN (examinateur), Professeur à l'Université de Metz

G. GOVAERT (directeur de thèse), Professeur à l'Université de Metz

G. LE CALVE (examinateur), Professeur à l'Université de Rennes II

A mes parents

REMERCIEMENTS

Ce travail a été réalisé sous la direction de Monsieur le Professeur G. GOVAERT. Je tiens à lui exprimer toute ma reconnaissance pour son aide, son dynamisme et sa compétance dont il m'a fait profiter tout au long de cette thèse.

Mes remerciements vont également à Monsieur le professeur G. CELEUX qui s'est intéressé de près à ce travail. Je lui exprime toute ma gratitude ainsi qu'à Monsieur le Professeur G. LE CALVÉ pour avoir accepté de rapporter cette thèse, pour les propositions judicieuses qu'ils m'ont suggérées et pour leur participation à la commission d'examen.

Je remercie également Messieurs les Professeurs E. DIDAY et Y. GARDAN pour avoir accepté de participer au jury.

Enfin, j'adresse une pensée particulière à tous les membres du Département Informatique pour leur gentillesse et leur disponibilité à toute épreuve.

TABLE DES MATIERES

INTRODUCTION

Une des premières étapes préliminaires à une méthode d'analyse des données est de définir et de mettre sous forme de tableau les données à étudier. En règle générale, ce tableau est à deux dimensions : il permet de décrire un ensemble d'individus à l'aide d'un ensemble de variables. Les tableaux auxquels nous nous intéressons ici sont les tableaux dont les éléments ne peuvent prendre que deux valeurs différentes. En général, les valeurs 0 et 1 sont utilisées pour coder ces deux possibilités et, dans ce travail, ces deux réponses sont considérées de façon parfaitement symétrique. On parle alors de tableaux de données binaires. Nous proposons ici des méthodes de classification pour ce type de tableaux et plus particulièrement pour les tableaux suivants :

- **Le tableau de variables binaires** : une variable est dite binaire lorsque son domaine de définition ne contient que deux valeurs distinctes. Celles-ci sont généralement codées par les valeurs 0 et 1. Les variables logiques, les attributs de descriptions, les variables sur signes de présence-absence sont des exemples de variables binaires.

- **Le tableau de codage binaire additif** : il s'agit de la transformation d'un tableau de variables qualitatives ordinales (appelé encore tableau de modalités) par le codage binaire additif (R.R. Sokal et P.H.A. Sneath 1963). Ce codage permet essentiellement de rester cohérent avec la notion d'ordre entre les modalités d'une variable. Ce tableau sera traité de la même façon que le tableau de variables binaires.

- **Le tableau disjonctif complet** : ce tableau résulte de la transformation, par le codage disjonctif complet, d'un tableau de variables qualitatives nominales encore appelé tableau de modalités ou questionnaire multiple (M. Jambu 1972, L. Lebart et al. 1977). Le codage disjonctif complet consiste à transformer chaque modalité d'une variable qualitative nominale en une variable binaire.

De nombreuses méthodes d'analyse de données ont pour support l'espace vectoriel $\mathbf{R}^p$ (F. Caillez et J.-P. Pages 1976). Une nouvelle étape préliminaire consiste alors à plonger les individus et les variables dans cet espace. Cette opération peut se réaliser directement comme pour les tableaux de variables quantitatives, ou indirectement, par l'intermédiaire d'un changement de codage, comme pour les tableaux de variables qualitatives. On dispose ainsi de deux ensembles de points inclus dans $\mathbf{R}^p$ et représentant l'ensemble des individus et l'ensemble des variables. On parle alors de nuage des individus et de nuage des variables. A partir des tableaux envisagés ici, on peut définir des nuages de points particuliers : les composantes des points étant binaires, l'espace considéré est alors un espace du type $\{0,1\}^p$ que l'on note également $\mathbf{B}^p$.

Les méthodes que nous envisageons ici sont des méthodes de type **classification automatique** portant sur l'ensemble des individus (ou celui des variables). Nous nous intéressons en particulier à celles dont la mise en place nécessite la définition d'un critère mesurant la qualité de la partition obtenue. Habituellement, on considère l'ensemble à classifier comme inclus dans l'espace $\mathbf{R}^p$ que l'on munit d'une métrique. Le critère, qui dépend de cette métrique, est une mesure de la qualité de la partition obtenue. Les ensembles à classifier, définis à partir des trois tableaux initiaux, peuvent être traités en suivant une telle approche. Cependant, si l'espace support est $\mathbf{R}^p$, les critères définis dans

ce cadre deviennent difficilement interprétables par rapport aux données initiales. En effet, on ne tient pas compte de la forme et de la structure particulière des éléments à classifier.

L'**idée de base** de ce travail repose sur le principe de **la conservation de la structure initiale** des données. Les méthodes de classification automatique proposées ici vérifient ce principe et fournissent des résultats directement interprétables par rapport aux données initiales, qu'elles soient simplement binaires ou issues d'un codage de modalités. Pour cela, individus et variables sont plongés dans l'espace $\mathbf{B}^p$ que l'on munit d'une métrique simple : la distance entre deux individus est le nombre de composantes différentes entre les deux points représentatifs de ces individus. Sur l'espace binaire, il s'agit exactement de la distance en valeurs absolues ou **distance L_1**. C'est dans ce cadre que nous allons construire des méthodes de classification automatique pour chacun des trois tableaux, méthodes reposant sur l'optimisation d'un critère établi à partir de cette distance.

Nous envisageons ensuite une approche métrique de $\mathbf{B}^p$ (muni de la distance L_1) en reprenant les notions habituelles de l'espace $\mathbf{R}^p$ muni d'une métrique euclidienne (F. Caillez et J.-P. Pages 1976). Nous définissons principalement une **inertie binaire** originale qui nous permet de démontrer des propriétés importantes. Cette inertie binaire permet notamment de donner une interprétation en terme d'inertie aux critères associés aux méthodes de classification sur tableaux de variables binaires et de les situer ainsi dans un contexte plus habituel.

Avant de décrire le contenu précis de ce travail, nous allons indiquer l'importance des données étudiées ici dans le domaine de la statistique. Il existe de nombreuses variables, dites discrètes, ne pouvant prendre par nature qu'un nombre restreint de valeurs. Citons par exemple les variables associées à la situation familiale (célibataire, veuf, divorcé, marié, ...) ou encore à la situation professionnelle d'une personne (cadre, artisan, étudiant, ...). Les variables ainsi définies sont appelées variables qualitatives nominales, les différentes possibilités sont appelées modalités. Il est également possible de construire une variable de ce type à partir d'une variable continue, découpée en différents intervalles, chaque intervalle étant alors associé à une modalité de la variable qualitative ainsi construite (J.Y. Lafaye 1979). Citons par exemple la variable associée à la taille d'une personne dont un découpage peut conduire aux qualitatifs : très petit, petit, moyen, grand très grand. Dans ce cas, les modalités sont ordonnées, on parle de variables qualitatives ordinales. D'autres variables ne peuvent être observées que de façon discrète, mais possèdent une structure latente continue. C'est le cas par exemple du degré de remission ou de progression d'une maladie. Lorsqu'une variable n'a que deux modalités, on parle simplement de variable binaire. Par exemple, un caractère peut être présent ou absent chez un individu. Le caractère est donc une attribution (par exemple avoir des ailes) ou une privation (ne pas en avoir). Les variables binaires peuvent aussi provenir de la transformation de données qualitatives par le codage disjonctif complet (chaque modalité est associée à une variable binaire).

Au cours de ces dernières années, de nombreuses méthodes statistiques ont été développées pour analyser ces données. Ce type d'approche a souvent été envisagé, moins cependant que pour les données quantitatives ou continues. Des modèles tenant compte de la notion de probabilité ont été développés et analysés.

D.R. Cox (1969) propose un modèle particulier aux données binaires appelé modèle logistique linéaire, à partir duquel il développe des méthodes de régression et d'analyse discriminante. Il montre que ces modèles logistiques linéaires jouent un rôle analogue à la théorie des modèles linéaires gaussiens dans l'analyse des données continues.

D'autres approches statistiques ont été effectuées, donnant lieu à des modèles spécifiques et variés, aussi bien pour les données binaires que pour les données qualitatives nominales et ordinales (S.J. Haberman 1978, Y.M.M. Bishop et al. 1980, D.J. Finney 1978, P.L. Plackett 1981, J.A. Anderson 1983 et 1984, W.E. Barlow et P. Feigl 1985).

Citons également l'importance des données binaires dans la pratique de la reconnaissance des formes. Dans ce domaine, nous pouvons citer les travaux de J.D. Tubbs (1989) qui reprend les travaux de E. Diday et J.C. Simon (1980) en s'appuyant d'avantage sur la notion de probabilité attchée aux données.

En ce qui concerne les données qualitatives, on pourra se référer aux travaux de L. Lebart et al. (1977). Il s'attache à la description des dépendances entre plusieurs variables dans une approche métrique des données. Il propose, en ce sens, une technique appelée analyse factorielle des correspondances multiples. On citera également J.P. Benzécri (1977) qui expose diverses interprétations des facteurs fournis par l'analyse précédente.

En classification automatique, l'approche la plus courante des données binaires consiste à définir, tout d'abord, une mesure de dissimilarité qui soit adaptée, puis à appliquer des méthodes classiques sur le tableau de dissimilarités ainsi défini. Dans ce type d'approche, les résulats obtenus dépendent évidemment du choix de l'indice. A ce titre, des études détaillées d'indices de similarité ou de dissimilarité ont été effectuées (P.H.A. Sneath et R.R. Sokal 1963, R.M. Cormack 1971, J.P. Benzécri 1973, I.C. Lerman 1973, M.R. Andeberg 1973, B.S. Duran et P.L Odell 1974, F. Caillez et J.P. Pages 1976, M. Jambu 1978, A.D. Gordon 1981, E. Diday, J. Lemaire, J. Pouget et F. Testu 1982).

Signalons également les travaux de B. Fichet et C. Le Calvé (1984) qui étudient les propriétés et la nature géométrique d'indices de similarités sur données binaires. J.C. Gower et P. Legendre (1986) étudient la capacité de certains indices à produire des matrices de distances métriques et euclidiennes. Ce travail est effectué pour des données de différents types, binaires et qualitatifs, et les auteurs concluent par des recommandations quant au choix d'un indice.

Une autre approche, aussi très fréquente, est la transformation des données discrètes initiales en données continues. On parle souvent de codage. G. Saporta (1968), C. Dupond-Gatelmand (1978), F. Taleng (1980), P. Cazes et al (1977) s'attachent, entre autres, à ce problème.

Quelques méthodes spécifiques de classification automatique pour données binaires ont été développées, aussi bien dans le domaine de la classification hiérarchique que pour la recherche de partitions. Différentes approches ont été envisagées, reposant sur les notions de métrique ou de probabilité, ou encore sur la théorie des mathématiques discrète (graphes, treillis) ou de l'algèbre booléenne. Nous énumérons ci-dessous quelques travaux relatifs à ces méthodes.

Dans le domaine de la classification hiérarchique, M.W. Buser et C. Baroni-Urbani (1982) proposent une méthode qui n'utilise pas la notion de dissimilarité entre objets à classifier, mais un concept d'homogénéité de classes basé sur la notion de probabilité d'erreur.

Toujours pour les données binaires, A. Reinert (1983) propose une méthode de classification descendante hiérarchique. Pour cela, il contruit des dichotomies successives optimisant un critère défini à partir de la métrique du Khi2. Après chaque dichotomie, la qualité des résultats est améliorée par un algorithme d'échange.

Lorsque les données sont qualitatives, différents critères d'agrégation peuvent être construit à partir de la métrique du Khi2, particulièrement adaptée à ce type de données (M. Roux 1985).

Lorsque le problème est posé sous la forme d'une recherche de partition optimisant un critère, une approche consiste à définir celui-ci à partir de modèles statistiques associés à chaque classe. La partition optimale peut alors être obtenue en utilisant les techniques d'estimation du maximum de vraisemblance.

Une autre technique consiste à approximer la partition au moyen d'une méthode de type nuées dynamiques. Celle-ci est utilisée par H.H. Bock (1986) qui propose deux modèles spécifiques pour données qualitatives et construit les algorithmes associés.

Pour les données qualitatives, une autre approche consiste à définir un critère à partir de la métrique du Khi2. Cette approche a été envisagée par H. Ralambondrainy (1988) qui propose un algorithme de type nuées dynamiques (appelé MNDQAL) utilisant un tel critère.

Pour des données binaires, nous pouvons citer les travaux de J.C. Gower (1974). Celui-ci introduit la notion de vecteurs binaires caractérisant les variables dans les classes et la notion de prédictions correctes. Il définit un critère évaluant cette prédiction et représentant simplement le nombre de différences entre les individus et les vecteurs représentatifs des classes. Il ne propose pas d'algorithme spécifique et suggère l'utilisation d'un algorithme de transfert. Par contre, il propose une solution au problème du nombre de classes à rechercher. Cette approche répond au problème de fidélité des résultats par rapport aux données initiales et correspond au cas le plus simple des méthodes envisagées dans ce travail.

Nous pouvons également citer la méthode de classification croisée sur données binaires (G. Govaert 1983). L'auteur propose de rechercher simultanément une partition des individus et des variables et utilise, pour cela, un algorithme appelé CROBIN et optimisant un critère analogue à celui proposé par J.C. Gower (1974). Ce critère représente simplement le nombre de différences entre le tableau de données initial et le tableau associé au couple de partitions recherché.

D'autres méthodes spécifiques aux données binaires et qualitatives ordinales sont décrites par A. Guénoche et B. Monjardet (1987). Les auteurs présentent un ensemble de méthodes relevant de l'algèbre de boole ou des mathématiques dicrètes, qu'il s'agisse de structures algébriques ou combinatoires. Nous pouvons citer des méthodes reposant sur l'algèbre de boole (A. Guénoche 1985, Cl. Flament 1976), celles utilisant la théorie des graphes comme support (B. Monjardet 1980), celles reposant sur la notion de treillis (B. Lefebvre 1977, B. Lefebvre et J. Losfeld 1979) ou encore celles opérant par réordonnancements (J. Bertin 1977, G. Caraux 1984).

Comme nous l'avons indiquée au début de cette introduction, nous envisageons de construire des méthodes de classification respectant la structure initiale des données. Pour chaque type de données, les méthodes proposées reposent sur l'optimisation d'un critère défini à partir de la distance L_1. Cette approche a été envisagée à la suite d'une étude approfondie de l'utilisation de cette distance dans le cas binaire simple et dans le cas de données binaires issues du codage disjonctif complet (de données qualitatives nominales) et du codage binaire additif (de données qualitatives ordinales).

Dans le premier chapitre, après une présentation de l'utilisation de la distance L_1 dans divers domaine, nous rappelons les caractéristiques qui lui sont associées dans le cadre de l'espace $\mathbf{R}^p$. Les notions mises en évidence sont la médiane d'une variable et le centre

médian d'un nuage de points. A ces deux caractéristiques centrales sont associées des mesures de dispersion que nous rappelons. Nous reprenons alors ces notions dans le cadre de l'espace B^p. On aboutit à des résultats très simples et facilement interprétables qui vont trouver une application en matière de classification.

En particulier, il s'avère que la médiane d'une variable binaire n'est autre que l'une des deux valeurs 0 ou 1 utilisées pour coder les deux possibilités. Cette valeur est déterminée par une règle très simple : la médiane est la valeur 1 ou 0 majoritaire de la variable. Il nous est alors possible de représenter le tableau initial par un simple vecteur de B^p. Une conséquence de cette appoche est que la caractéristique de dispersion s'interprète simplement comme le nombre de différences entre les valeurs du tableau initial et les médianes des variables.

Lorsque les données sont issues d'un codage additif binaire, nous retrouvons des résultats analogues. Ainsi, la valeur caractérisant une variable qualitative ordinale, n'est autre que sa médiane. De plus, la règle fournissant la médiane d'une variable binaire s'applique parfaitement à ce type de données. Le tableau initial est alors résumé par un vecteur de modalités, facilement interprétable.

Lorsque les données proviennent du codage disjonctif complet, il nous est nécessaire d'imposer une contrainte lors de la recherche des caractéristiques associées à la distance. La démarche suivie ici diffère de la précédente. Cependant, la contrainte nous permet de caractériser une variable qualitative nominale par l'une de ses modalités. Un tableau de telles variables peut alors se résumer par un vecteur de modalités.

Dans les trois chapitres qui suivent, nous proposons un algorithme de classification pour chaque type de tableaux envisagés ici. A chaque fois, nous définissons des critères spécifiques à partir de la distance L_1 et construisons des méthodes reposant sur un même principe et fournissant des solutions locales au problème de classification posé. En reprenant les principaux résultats de l'étude de la distance en valeurs absolues, il est possible de construire des partitions telles que les classes obtenues soient simplement résumées par des vecteurs de même type que les éléments à classifier : si les données sont binaires, une classe est caractérisée par un vecteur binaire; si les données binaires proviennent du codage d'un tableau de modalités, une classe est caractérisée par un vecteur de modalités. Une autre caractéristique des méthodes développées ici est la facile interprétation des critères associés aux partitions. A chaque fois, ils s'interprètent en terme d'écart ou nombre de différences entre données initiales et résumés des classes. De plus, des études comparatives avec des méthodes existantes et des applications sont proposées. Les méthodes de classification automatique développées dans cette étude sont les suivantes :

MNDBIN pour les tableaux de variables binaires (chapitre 2),

MNDORD pour les tableaux de codage binaire additif (chapitre 3),

MNDDIJ pour les tableaux de codage disjonctif complet (chapitre 4).

Les programmes correspondants ont été écrits et intégrés au logiciel d'analyse des données SICLA (Système Interactif de CLassification Automatique) développé par l'équipe de "classification automatique et reconnaissance des formes" de l'INRIA. Ce système permet la gestion, la description élémentaire et le traitement de données par les méthodes d'analyse multidimentionnelle. Son objectif principal est la diffusion de méthodes issues de recherches récentes, principalement celles concernant la classification automatique et notamment les Nuées Dynamiques.

Dans le chapitre 5, nous nous intéressons plus pariculièrement à l'espace binaire $\mathbf{B}^p$ muni de la distance en valeurs absolues. Bien que cette distance ne soit pas euclidienne, il est possible de démontrer, sous certaines hypothèses, des propiétés analogues à celles vérifiées par la distance euclidienne usuelle sur l'espace vectoriel $\mathbf{R}^p$. En particulier, à partir de la distance L_1, nous définissons une **inertie binaire**, puis nous démontrons que celle-ci vérifie, comme l'inertie usuelle, une relation de type Huyghens. Il est alors possible d'établir une relation de décomposition de l'inertie sous la forme habituelle :

inertie totale = inertie intraclasse + inertie interclasse.

Ce résultat important va nous permettre de replacer la méthode de classification MNDBIN dans un contexte plus habituel : le critère optimisé apparait comme une inertie intraclasse de la partition et il est aussi possible de donner une interprétation de cette méthode en terme d'optimisation d'une inertie interclasse. Un autre point important de ce chapitre est la définition d'une mesure d'information associée à un tableau binaire. Cette mesure est bien sûr définie à partir de l'inertie binaire, et elle est introduite de façon analogue aux mesures d'informations associées aux tableaux de mesures ou de contingence (G. Govaert 1983). Il est alors possible de replacer la méthode de classification croisée CROBIN (G. Govaert 1983) dans le contexte général de la classification croisée : d'une part la méthode recherche un tableau associé à un couple de partition résumant le tableau initial de sorte que la perte d'information due aux regroupements en classes soit minimale; d'autre part nous montrons que les deux algorithmes intermédiaires utilisés par la méthode ne sont autres qu'un seul et même algorithme (une version généralisée de l'algorithme MNDBIN). Toujours à partir de cette inertie binaire, nous pouvons définir des indices d'agrégations et ainsi construire une hiérarchie des individus ou des variables en utilisant l'algorithme habituel de classification ascendante hiérarchique. Nous terminons ce chapitre par un exemple d'application des indices définis ici et ceux habituellements utilisés.

Dans le chapitre 6, nous commençons par un bref résumé des différentes méthodes d'analyse factorielle utilisées dans le cas binaire. Nous proposons ensuite une méthode d'analyse en composantes principales spécifique aux données binaires et qui soit en accord avec la forme initiale des données. Pour cela, nous reprenons les hypothèses habituelles de l'analyse en composantes principales que nous replaçons dans le contexte de l'espace binaire muni de la distance L_1. En se sens, nous définissons l'inertie d'un nuage de $\mathbf{B}^p$ par rapport à un sous-ensemble de $\mathbf{B}^p$, et cela à partir de la notion d'inertie binaire précédemment définie. Sous certaines contraintes, cette approche rejoint la méthode de classification MNDBIN et, lorsque ces contraintes sont supprimées, on retrouve la méthode d'analyse factorielle booléenne présentée dans le logiciel BMDP par M.R. Mickey, P. Mundel et L. Engelman (1984). En adoptant la représentation matricielle du modèle d'analyse factorielle booléenne, nous reprenons et résumons les méthodes proposées ici dans le cas binaire. Nous terminons encore une fois par des exemples d'application en détaillant à chaque fois les résultats obtenus.

CHAPITRE 1

ETUDE DE LA DISTANCE L_1 POUR DES DONNEES BINAIRES ET QUALITATIVES

1. INTRODUCTION

La distance L_1 n'est pas euclidienne, mais elle est très utilisée dans le domaine de la statistique. En régression, des algorithmes ont été proposés pour l'estimation des paramètres de modèles formulés à partir de cette distance. J.E Gentle, V.A. Sposito et S.C. Narula (1988) se proposent de comparer l'efficacité de certains de ces algorithmes et notamment ceux construits par R.D. Amstrong et D.S. Kung (1978), R.D. Amstrong et al. (1978), P. Bloomfield et W. Steiger (1980), L.A. Josvanger et V.A. Sposito (1983).

L'utilisation de cette distance est souvent associée à la notion de médiane et de centre médian (J.C. Gower 1974b, F.K. Bedall et H. Zimmermann 1979). Le centre médian d'un nuage de points est introduit comme le point minimisant un critère de dispersion défini à partir de la distance L_1. Ce centre peut être déterminé de façon analytique mais ne vérifie pas la propriété d'unicité. R. Kosfeld (1986) reprend la notion de médiane d'une variable réelle et considère deux extensions pour une variable multidimensionnelle à valeurs dans $\mathbf{R}^p$: la première n'est autre que le centre médian; la seconde est appelée "spatial median" et permet de minimiser un critère défini à partir de la distance euclidienne usuelle. La seconde extension est plus difficilement calculable mais vérifie la propriété d'unicité.

Dans ce chapitre, nous nous intéressons aux propriétés et caractéristiques de la distance L_1 sur un espace du type $\{0,1\}^p$ ou $\mathbf{B}^p$. Les trois tableaux de données binaires décrits dans l'introduction (tableau de variables binaires, tableau de codage binaire additif et tableau de codage disjonctif complet) permettent de définir des nuages de points, à composantes dans $\{0,1\}$, représentant les individus. Ces trois nuages se distinguent par la stucture particulière de leurs éléments : soit ils sont à composantes dans $\{0,1\}$ sans contrainte, soit leurs composantes vérifient le codage disjonctif complet ou encore le codage binaire additif. Ces nuages sont donc tous inclus dans un espace du type $\mathbf{B}^p$. Il reste à choisir une métrique sur cet espace. Nous proposons d'utiliser la distance suivante : la distance entre deux points sera le nombre de composantes différentes entre ces deux points. Sur l'espace $\mathbf{B}^p$, il s'agit exactement de la distance en valeurs absolues ou distance L_1.

Dans ce chapitre, nous définissons les caractéristiques de cette distance pour chacun de ces nuages de points. Dans tous les cas, nous essayons d'obtenir des caractérisques facilement interprétables et permettant de décrire simplement les tableaux de données.

Le premier paragraphe traite essentiellement de la distance en valeurs absolues dans le cadre de l'espace vectoriel $\mathbf{R}^p$. En particulier, nous présentons les caractéristiques associées à savoir la notion de médiane d'une variable et la notion de centre médian d'un nuage de points.

Dans le second paragraphe, nous reprenons ces caractéristiques lorsque l'espace est $\mathbf{B}^p$. En fait, les caractéristiques mises en évidence sont la médiane d'une variable binaire et le centre médian d'un nuage de points associé à un tableau de variables binaires. Les résultats obtenus nous semblent intéressants dans la mesure où la nature et la structure des données est respectée.

Nous reprenons ensuite ces résultats lorsque le nuage de points est celui associé à un tableau de codage binaire additif. Les composantes des points vérifiant la contrainte de codage, nous parlons alors de vecteurs binaires de modalités. Les propriétés mises en évidence dans le cadre de l'espace $\mathbf{B}^p$ restent utilisables ici, car elles permettent de rester en accord avec la structure des données. Une autre approche est également envisagée, celle-ci consiste à travailler directement sur le tableau de modalités. A partir des lignes de ce tableau, on définit un nuage de points dont les composantes sont des modalités. Nous parlons alors de vecteurs de modalités. Ce nuage particulier est inclus dans un espace **E**, que nous définissons, et que l'on munit de la distance en valeurs absolues. Nous déterminons également les caractéristiques de la distance sur cet espace **E**. Nous montrons alors l'équivalence avec l'approche précédente.

Une étude semblable est effectuée pour le tableau de codage disjonctif complet et le tableau de variables qualitatives nominales. A partir du tableau de codage, on définit un nuage dont les éléments ont une structure de vecteur binaire de modalités. Pour ces points particuliers, les caractéristiques de la distance en valeurs absolues ne sont plus en accord avec la forme initiale des données. Il est nécessaire d'introduire la contrainte de codage pour obtenir des caractéristiques interprétables. A partir du tableau de modalités on peut définir un nuage de points (ayant une structure de vecteurs de modalités) inclus dans un espace **E** (du même type que celui défini précédemment). Cet espace **E** est muni de la distance égale au nombre de composantes différentes entre les deux points considérés. Dans ce cadre particulier, les caractéristiques mises en évidence sont facilement interprétables. Enfin, une étude comparative entre les deux approches proposées pour ce type de données est effectuée.

2. LA DISTANCE EN VALEURS ABSOLUES

Dans ce paragraphe, nous nous intéressons plus particulièrement à l'aspect descriptif de la distance en valeurs absolues.

Des caracatéristiques de valeurs centrales et de dispersion sont souvent associées aux distances (F. Caillez et J.P. Pages 1976). Pour la métrique euclidienne usuelle sur l'espace $\mathbf{R}^p$, celles-ci correspondent à la moyenne et à l'écart-type d'une variable réelle, au centre de gravité et à l'inertie d'un nuage de points. Pour la distance en valeurs absolues nous pouvons définir des notions analogues.

Dans ce paragraphe, nous précisons ces notions pour une variable quantitative, puis pour un nuage de points pouvant être considéré comme le nuage associé à un tableau croisant individus et variables quantitatives.

2.1 DEFINITION

Notons d la distance en valeurs absolues.

Entre deux points quelconques $x=(x^1,x^2,\ldots,x^p)$ et $y=(y^1,y^2,\ldots,y^p)$ de $\mathbf{R}^p$, elle s'exprime par :

$$d(x,y) = \sum_{j=1}^{p} |x^j - y^j|$$

2.2 CARACTERISTIQUES D'UNE VARIABLE REELLE

Soit I={1,2,...,n} un ensemble de n individus décrit par une variable réelle x. Soit E(x) l'échantillon des valeurs prises par les individus de I. On note :

$$E(x) = \{ (x_i, p_i), i \in I \}$$

où

x_i est la valeur prise par la variable x sur l'individu i,

p_i la pondération associée à l'individu i et $\sum_{i \in I} p_i = 1$.

Le plus souvent, on choisit la pondération 1/n pour tous les individus. Dans cette étude, nous conservons la notation p_i désignant une pondération quelconque.

Les caractéristiques habituelles sont la moyenne et l'écart-type de la variable x.

2.2.1 Caractéristique de valeur centrale : la médiane

Définition

*Toute valeur λ de **R** minimisant :*

$$\sum_{i \in I} p_i / x_i - \lambda /$$

*est une caractéristique de valeur centrale de la variable x. Elle est appelée **médiane** de l'échantillon de la variable x. Nous parlons plus simplement de **médiane de la variable**.*

Remarques

La médiane vérifie la propriété suivante : elle est comprise entre la plus grande et la plus petite valeur de l'échantillon. Soit :

$$\text{Min } \{x_1, x_2, \ldots, x_n\} \leq \lambda \leq \text{Max } \{x_1, x_2, \ldots, x_n\}$$

De plus, la médiane existe toujours, mais ne possède pas la propriété d'unicité. Dans l'exemple ci-dessous, tout point de l'intervalle [2,3] est médiane :

$$E(x) = \{1, 2, 3, 4\}$$

Nous pouvons également démontrer qu'un échantillon admet toujours pour médiane l'une des valeurs qu'il contient. Ainsi, soit l'échantillon admet une médiane unique et appartenant à E(x), soit il admet un intervalle médian dont les bornes sont des valeurs figurant dans E(x). Nous donnons ici deux exemples, en supposant que toutes les mesures soient munies d'une même pondération :

(i) $E(x) = \{1, 2, 3, 4, 5\}$ médiane unique et égale à 3

(ii) $E(x) = \{1, 2, 3, 4\}$ intervalle médian [2,3]

Règle de calcul de la médiane

Lorsque toutes les valeurs de l'échantillon sont munies d'une même pondération (en général 1/n ou plus simplement 1), la médiane peut être déterminée de façon analytique à partir d'une règle très simple. Elle ne dépend que du nombre de valeurs de l'échantillon : si l'échantillon contient un nombre impair de valeurs, la médiane est la valeur centrale de l'échantillon; si l'échantillon contient un nombre pair de valeurs, il admet un intervalle

médian dont les bornes sont les deux valeurs centrales. En résumé, et si on suppose l'échantillon trié de sorte que $x_1 \leq x_2 \leq \ldots \leq x_n$, la règle fournit les résultats suivants :

- si n est impair, une médiane unique x_k où k est égal à la partie entière de $(n/2 + 1)$,
- si n est pair, un intervalle médian $[x_k, x_{k+1}]$ où k est égal à la partie entière de $(n/2)$.

Les exemples (i) et (ii) illustrent cette règle. Nous pouvons ajouter le cas particulier où, par exemple, $E(x)=\{1,2,2,2\}$. La règle fournit un intervalle médian dont les bornes sont toutes deux égales à 2, on choisit alors cette unique valeur pour médiane.

Cependant, cette règle présente des limites. En effet, dans le cas où les mesures sont munies de pondérations différentes, elle ne fournit pas nécessairement la bonne solution. Les trois exemples ci-dessous en sont des illustrations :

- l'ensemble $\{1, 2, 3, 4, 5\}$ muni des pondérations $(\frac{1}{10}, \frac{2}{10}, \frac{1}{10}, \frac{5}{10}, \frac{1}{10})$ admet une médiane unique de valeur 4,
- l'ensemble $\{1, 2, 3, 4, 5\}$ muni des pondérations $(\frac{5}{20}, \frac{6}{20}, \frac{1}{20}, \frac{7}{20}, \frac{1}{20})$ admet une médiane unique de valeur 2,
- l'ensemble $\{1, 2, 3, 4, 5\}$ muni des pondérations $(\frac{1}{10}, \frac{1}{10}, \frac{3}{10}, \frac{4}{10}, \frac{1}{10})$ admet un intervalle médian [3,4].

Enfin, lorsque deux valeurs seulement sont en jeu, la médiane est facilement calculable. En effet, un ensemble de deux réels x_1 et x_2, munis respectivement des pondérations p_1 et p_2, admet pour médiane la valeur x_1 ou x_2 de plus forte pondération associée. Si p_1 et p_2 sont égales, tout réel de l'intervalle $[x_1,x_2]$ est médiane.

2.2.2 Caractéristique de dispersion : l'écart moyen

Nous définissons une mesure de la dispersion des n éléments de E(x) autour d'un réel quelconque. Nous la notons $\mathbf{EC_t}(x)$, égale à la somme pondérée (ou moyenne) des valeurs absolues des écarts des n mesures au réel t. Nous l'appelons écart par rapport à t et son expression est :

$$\mathbf{EC_t}(x) = \sum_{i \in I} p_i |x_i - t|$$

Par définition, cette quantité est minimale pour la médiane de l'échantillon.

Définition

On appelle ***écart moyen*** *de la variable x, l'écart par rapport à sa médiane λ et on le note :*

$$\boldsymbol{EC(x) = EC_{\lambda}(x)}$$

L'écart moyen et l'écart-type d'une variable ont des significations analogues : ce sont des quantités mesurant la dispersion de l'échantillon autour des valeurs centrales, que sont respectivement la médiane et la moyenne de la variable.

2.2.3 Exemple

Soient E(x)={1,2,3,4,5} l'échantillon et {1/10,2/10,1/10,5/10,1/10} les pondérations associées. Les caractéristiques habituelles de la variable ainsi que celles associées à la distance en valeurs absolues sont :

- moyenne = 3.3 et écart type = 1.19
- médiane = 4 et écart moyen = 0.9

2.3 CARACTERISTIQUES D'UN NUAGE DE POINTS

Soit I={1,2,...,n} un ensemble de n individus décrit par un ensemble J={1,2,...,p} de p variables quantitatives. A chaque individu i est associé un point x_i de $\mathbf{R}^p$ défini par :

$$x_i = (x_i^1, x_i^2, \ldots, x_i^p)$$

où x_i^j représente la valeur prise par la variable j sur l'individu i.

On obtient ainsi un nuage de n points pondérés, noté N(I), et défini par :

$$N(I) = \{ (x_i, p_i), i \in I \}$$

où p_i est la pondération associée à l'individu i.

Pour la distance euclidienne, les caractéristiques du nuage sont le centre de gravité et l'inertie (par rapport au centre de gravité). Ici, nous définissons des caractéristiques analogues pour la distance en valeurs absolues.

2.3.1 Caractéristique de valeur centrale du nuage : le centre médian

Définition

Tout élément λ de $\mathbf{R}^p$ minimisant :

$$\sum_{i \in I} p_i \, d(x_i, \lambda)$$

est une caractéristique de valeur centrale du nuage. Un tel point est appelé ***centre médian*** *du nuage.*

De cette définition, se dégage le problème de la recherche du centre médian d'un nuage de $\mathbf{R}^p$. Afin de résoudre celui-ci, écrivons la quantité à minimiser de la façon suivante :

$$\sum_{j \in J} \sum_{i \in I} p_i \, | x_i^j - \lambda^j |$$

Il suffit alors de rechercher, pour tout j, le réel λ^j minimisant :

$$\sum_{i \in I} p_i \, | x_i^j - \lambda^j |$$

La solution est de choisir, pour λ^j, la valeur médiane de l'ensemble $\{(x_i^j, p_i), i \in I\}$.

Finalement, le centre médian est le point dont les composantes sont les médianes des variables. Par conséquent, il existe toujours, mais n'est pas unique.

2.3.2 Caractéristique de dispersion du nuage : l'inertie

Sur l'espace $\mathbf{R}^p$ muni de la distance euclidienne, on dispose de la notion habituelle d'inertie du nuage par rapport à un point t quelconque. La caractéristique de dispersion est alors l'inertie du nuage par rapport à son centre de gravité. Ici, l'espace $\mathbf{R}^p$ est muni de la distance en valeurs absolues. Pour cette distance, nous définissons une inertie particulière.

Définitions

Soit $N(I)=\{(x_i, p_i), i \in I\}$ un nuage de $\mathbf{R}^p$ et soit d la distance en valeurs absolues. Nous définissons l'inertie du nuage par rapport à un point t par :

$$\mathfrak{I}_t(N(I)) = \sum_{i \in I} p_i \, d(x_i, t)$$

*Nous appelons **inertie du nuage** N(I) l'inertie du nuage par rapport à son centre médian et on note :*

$$\mathfrak{I}(N(I)) = \mathfrak{I}_\lambda N(I))$$

où λ est le centre médian du nuage N(I).

Lorsque l'espace $\mathbf{R}^p$ est muni de la distance euclidienne usuelle, l'inertie est minimale pour le centre de gravité du nuage. De plus, il est possible d'exprimer cette inertie en fonction des variances des variables.

Ici, l'espace $\mathbf{R}^p$ est muni de la distance en valeurs absolues. Par définition, l'inertie du nuage est minimale pour le centre médian. D'autre part, on peut exprimer l'inertie en fonction des écarts moyens des variables. En effet, on a :

$$\mathfrak{I}(\mathbf{N}(I)) = \sum_{i \in I} \sum_{j \in J} p_i \, | x_i^j - \lambda^j | = \sum_{j \in J} \mathbf{EC}(j)$$

où

EC(j) exprime l'écart moyen de la variable j

3. DESCRIPTION D'UN TABLEAU BINAIRE

Nous nous intéressons ici aux tableaux croisant individus et variables binaires. Nous supposons, en outre, que les deux modalités d'une variable de ce type sont codées par les valeurs de l'ensemble {0,1}.

Nous donnons, tout d'abord, les notations utilisées pour le tableau de variables binaires et pour le nuage des individus défini à partir de ce tableau. Chaque individu est représenté par un vecteur binaire; le nuage des individus est inclus dans un espace du type $\{0,1\}^p$ que l'on note $\mathbf{B}^p$.

Pour mesurer les proximités entre individus, nous proposons d'utiliser la distance égale au nombre de valeurs qui ne sont pas identiques pour les deux vecteurs binaires correspondants. Sur l'espace $\mathbf{B}^p$, il s'agit exactement la distance en valeurs absolues.

Dans un second paragraphe, nous reprenons les notions de médiane et d'écart moyen pour l'échantillon d'une variable binaire. Nous montrons alors qu'il existe toujours une médiane binaire, facilement déterminée à partir d'une règle que nous préciserons.

Puis, nous nous intéressons aux caractéristiques de description d'un nuage de $\mathbf{B}^p$. Nous montrons que le centre médian du nuage a la particularité d'appartenir à ce même espace. Nous établissons également une relation liant l'inertie du nuage à l'inertie par rapport à un point quelconque de $\mathbf{B}^p$.

Dans le dernier paragraphe, nous montrons que les résultats obtenus sont indépendants des valeurs a et b choisies pour coder les deux possibilités d'une variable binaire.

3.1 NOTATIONS

Le tableau

Soit $\mathbf{X}(I,J)$ le tableau croisant un ensemble $I=\{1,2,\ldots,n\}$ de n individus et un ensemble $J=\{1,2,\ldots,p\}$ de p variables binaires. On note :

$$\mathbf{X}(I,J) = (x_i^j)$$

où x_i^j, élément de la $i^{ème}$ ligne et de la $j^{ème}$ colonne du tableau, est la valeur 0 ou 1 prise par la variable j sur l'individu i.

Le nuage

Chaque individu i est représenté, dans $\mathbf{B}^p$, par le point x_i défini par :

$$x_i = (x_i^1, x_i^2, \ldots , x_i^p)$$

A partir du tableau $\mathbf{X}(I,J)$, on définit ainsi le nuage de n points pondérés $\mathbf{N}(I)$ par :

$$\mathbf{N}(I) = \{ (x_i, p_i), i \in I \}$$

où p_i est la pondération associée à l'individu i. On suppose encore que $\sum_{i \in I} p_i = 1$.

La distance

Si x et y sont deux éléments de $\mathbf{B}^p$, la distance en valeurs absolues d entre ces deux points s'exprime par :

$$d(x,y) = \sum_{j=1}^{p} |x^j - y^j|$$

3.2 CARACTERISTIQUES D'UNE VARIABLE BINAIRE

Soient x une variable binaire et E(x) l'échantillon des valeurs observées sur l'ensemble I des individus. On note :

$$E(x) = \{ (x_i, p_i), i \in I \}$$

où x_i est la valeur 0 ou 1 prise par la variable x sur l'individu i,

p_i la pondération associée à l'individu i.

3.2.1 La médiane

Propriété
Tout échantillon d'une variable binaire, à valeurs dans {0,1}, admet une médiane appartenant à ce même ensemble {0,1}.

preuve
Cherchons λ tel que :

$$\sum_{i \in I} p_i |x_i - \lambda| \quad \text{soit minimum}$$

Cette expression peut être décomposée de la façon suivante :

$$\sum_{i \in I} p_i (1-x_i)\lambda + \sum_{i \in I} p_i x_i (1-\lambda)$$

que nous réécrivons :

$$\lambda P_0 + (1-\lambda) P_1$$

où $P_0 = \sum_{i \in I} p_i (1-x_i)$ la somme des pondérations des valeurs 0,

et $P_1 = \sum_{i \in I} p_i x_i$ la somme des pondérations des valeurs 1.

Par conséquent, le problème revient à rechercher la médiane des deux valeurs 0 et 1, munies respectivement des pondérations P_0 et P_1. La solution est de prendre pour médiane la valeur 0 ou 1, suivant que P_0 est plus grand ou plus petit que P_1. Si les deux pondérations sont différentes, la médiane est unique. Par contre, si P_0 et P_1 sont identiques, la quantité à minimiser est indépendante de la valeur médiane recherchée. Dans ce cas, tout réel de l'intervalle [0,1] est médiane. On dispose ainsi d'une règle permettant de déterminer facilement une médiane binaire présentant l'avantage (que ne possède pas la moyenne dans ce cas précis) d'être du même type que les données

Cas où les individus sont munis de pondérations égales
Lorsque les individus sont munis d'une même pondération, la règle fournit une médiane ayant une interprétation particulièrement simple.

En supposant que :

$$\forall\, i \in I \quad p_i = \rho$$

on a alors :

$$P_0 = \sum_{i \in I} \rho (1-x_i) = \rho\, n_0$$

où $n_0 = \sum_{i \in I} (1-x_i)$ le nombre de valeurs 1 de l'échantillon,

$$P_1 = \sum_{i \in I} \rho\, x_i = \rho\, n_1$$

où $n_1 = \sum_{i \in I} x_i$ le nombre de valeurs 0 de l'échantillon.

La médiane est alors la valeur minimisant :

$$\lambda n_0 + (1-\lambda) n_1$$

Finalement, la médiane apparait comme la valeur 0 ou 1 la plus souvent choisie par les individus. Elle représente la valeur majoritaire de l'échantillon.

Remarque
La moyenne habituelle possède la propriété suivante : on peut remplacer une partie d'un ensemble par sa moyenne, munie de la somme des pondérations, sans changer la moyenne de l'ensemble. La médiane ne vérifie pas cette propriété.

Pour illustrer cette remarque, considérons l'ensemble E composé des 9 valeurs binaires {1,1,1,1,0,0,0,0,0} chacune de pondération 1. Il est alors possible d'appliquer simplement la règle de la majorité pour déterminer la médiane. Soient également les trois parties A, B et C telles que :

$$E = A \cup B \cup C$$

$$A = \{1, 1\} \text{ de médiane } a = 1 \text{ de pondération } 2$$

$$B = \{1, 1, 0\} \text{ de médiane } b = 1 \text{ de pondération } 3$$

$$C = \{0, 0, 0, 0\} \text{ de médiane } c = 0 \text{ de pondération } 4$$

La médiane de E est 0, alors que la médiane des valeurs a, b et c, munies de leur pondération respective, est égale à 1. Il se pose donc un nouveau problème : celui du choix de la pondération à associer à la médiane binaire. Cela sera l'objet du chapitre 5.

3.2.2 L'écart moyen

On peut décomposer l'expression de l'écart moyen d'une variable binaire x de la façon suivante :

$$\mathbf{EC}(x) = \sum_{i \in I} p_i |x_i - \lambda|$$

$$\Leftrightarrow \quad \mathbf{EC}(x) = \lambda P_0 + (1-\lambda) P_1$$

$$\Leftrightarrow \quad \mathbf{EC}(x) = \begin{cases} P_0 \text{ si } \lambda = 1 \\ P_1 \text{ si } \lambda = 0 \end{cases}$$

$$\Leftrightarrow \quad \mathbf{EC}(x) = \sum \{ p_i / i \in I \text{ et } x_i \neq \lambda \}$$

L'écart moyen est donc égal à la somme des pondérations associées aux valeurs différentes de la médiane binaire. Il s'interprète comme une mesure de la présence de la valeur 0 ou 1 de plus faible pondération.

Si tous les individus sont munis d'une pondération égale, la médiane est la valeur 0 ou 1 majoritaire de l'échantillon. L'écart moyen représente alors, à un facteur constant près, le nombre de valeurs minoritaires de l'échantillon. En effet, si on suppose que :

$$\forall i \in I \quad p_i = \rho$$

on a alors :

$$\mathbf{EC}(x) = \sum \{ \rho / i \in I \text{ et } x_i \neq \lambda \}$$

$$\Leftrightarrow \quad \mathbf{EC}(x) = \rho \, \mathrm{Card}\{ i \in I / x_i \neq \lambda \}$$

3.2.3 Exemple d'application

Soit l'echantillon E(x) = {1,0,0,1,0,0,1,0,0,0}. En supposant que chaque mesure est munie d'une pondération 1, on obtient :

- une médiane égale à 0,
- un écart moyen de 3.

3.2.4 Propriété

Nous montrons ici une propriété qui sera utilisée dans la suite de ce travail. Il s'agit d'exprimer l'écart par rapport à un réel quelconque en fonction de l'écart moyen. Cette propriété permettra, notamment, d'établir une relation analogue pour l'inertie.

Propriété

Soient λ et $EC(x)$ la médiane et l'écart-moyen de la variable binaire x. L'écart par rapport à un réel quelconque de l'intervalle [0,1] est lié à l'écart moyen par la relation suivante :

$$\forall t \in [0,1] \quad EC_t(x) = EC(x) + |t - \lambda||P_0 - P_1|$$

Preuve
On a :

$$\mathbf{EC}_t(x) = tP_0 + (1-t)P_1$$

$$\mathbf{EC}(x) = \lambda P_0 + (1-\lambda)P_1$$

En effectuant la différence entre $\mathbf{EC}_t(x)$ et $\mathbf{EC}(x)$, on obtient :

$$\mathbf{EC}_t(x) - \mathbf{EC}(x) = P_0(t-\lambda) + P_1(\lambda-t)$$

$$\Leftrightarrow \quad \mathbf{EC}_t(x) - \mathbf{EC}(x) = (t-\lambda)(P_0-P_1)$$

Cette différence étant toujours de même signe puisque :

- Si $P_0 \geq P_1$ alors $\lambda = 0$ et la différence est positive,
- Si $P_0 \leq P_1$ alors $\lambda = 1$ et la différence est positive,
- Si $P_0 = P_1$ alors pour tout t de [0,1] la différence est nulle,

nous pouvons écrire :

$$\mathbf{EC}_t(x) - \mathbf{EC}(x) = |t - \lambda||P_0 - P_1|$$

ce qui démontre la propriété.

On retrouve bien que la médiane est la valeur d'écart minimum.

3.3 CARACTERISTIQUES D'UN NUAGE DE POINTS

3.3.1 Centre médian du nuage

Propriété
Un nuage de B^p admet toujours un centre médian appartenant au même espace B^p.

Preuve
Cette propriété est une conséquence de la propriété de la médiane binaire. En effet, par définition, le centre médian du nuage **N**(I) est tout point λ de R^p minimisant :

$$\sum_{i\in I} \sum_{j\in J} p_i |x_i^j - \lambda^j|$$

ce qui signifie que, pout tout j, le réel λ^j minimise :

$$\sum_{i\in I} p_i |x_i^j - \lambda^j|$$

Les données étant binaires, λ^j est la médiane binaire de l'ensemble des valeurs prises par la variable j sur l'ensemble des individus. Le centre médian est donc le point dont les composantes sont les valeurs 0 ou 1 médianes des variables. Le nuage **N**(I) admet donc un centre médian de la même forme que les données initiales. Le tableau **X**(I,J) est ainsi résumé par un vecteur binaire facilement interprétable.

Notons que, comme pour la médiane, remplacer une partie d'un ensemble de points par son centre médian peut changer le centre médian de l'ensemble.

3.3.2 Inertie du nuage

Si on note λ le centre médian du nuage **N**(I), l'expession de l'inertie est :

$$\Im(\mathbf{N}(I)) = \sum_{i\in I} \sum_{j\in J} p_i \, | x_i^j - \lambda^j | = \sum_{i\in I} \sum_{j\in J} \{p_i / x_i^j \neq \lambda^j\}$$

Dans cette expression, seules les composantes différentes de celles du centre médian fournissent une contribution non nulle à l'inertie. Si on choisit une même pondération pour tous les points, l'inertie sera égale, à un facteur constant près, au nombre de valeurs minoritaires prises par les variables binaires. En effet, si on suppose que :

$$\forall\, i\in I \quad p_i = \rho$$

on a alors :

$$\Im(\mathbf{N}(I)) = \sum_{i\in I} \sum_{j\in J} \{\rho / x_i^j \neq \lambda^j\} = \rho \sum_{j\in J} \mathrm{Card}\{i\in I / x_i^j \neq \lambda^j\}$$

3.3.3 Exemple d'application

Considérons le tableau croisant 5 individus et 3 variables binaires, représenté ci-après. Pour ce tableau, et en supposant des pondérations toutes égales à 1, nous avons calculé le vecteur binaire le résumant, ainsi que la mesure de dispertion autour de ce vecteur.

Nous obtenons alors les résultats suivants :

1 0 1 1 1 0 0 0 1 1 0 0 1 1 1	centre médian :	(1, 0, 1)
	inertie :	5
tableau de données		

3.3.4 Propriété

Sur l'espace $\mathbf{R}^p$ muni de la distance euclidienne, on dispose d'une relation liant l'inertie d'un nuage par rapport à un point quelconque à l'inertie du nuage par rapport à son centre de gravité : la relation de Huyghens.

Nous allons déterminer une relation analogue sur l'espace $\mathbf{B}^p$ muni de la distance en valeurs absolues.

Propriété

Soient λ le centre médian d'un nuage N(I) de $\mathbf{B}^p$ et t un point quelconque de $\mathbf{B}^p$. L'inertie du nuage par rapport au point t s'exprime, en fonction de l'inertie de N(I), de la façon suivante :

$$\mathfrak{I}_t(N(I)) = \mathfrak{I}(N(I)) + \sum_{j\in J} |t^j - \lambda^j| |P_0^j - P_1^j|$$

$$où\ P_0^j = \sum_{i\in I} p_i(1-x_i^j)\ \ et\ \ P_1^j = \sum_{i\in I} p_i x_i^j$$

La propriété reste valable pour un point t dont toutes les composantes sont dans l'intervalle [0,1].

Preuve

L'inertie par rapport au point t peut s'écrire comme la somme des écarts des variables par rapport aux composantes de t :

$$\mathfrak{I}_t(\mathbf{N}(I)) = \sum_{i\in I} \sum_{j\in J} p_i |x_i^j - t^j|$$

$$\Leftrightarrow \quad \mathfrak{I}_t(\mathbf{N}(I)) = \sum_{j\in J} \mathbf{EC}_{t^j}(j)$$

En reportant la relation exprimant $\mathbf{EC}_{t^j}(j)$ en fonction de $\mathbf{EC}(j)$, on obtient :

$$\mathfrak{I}_t(\mathbf{N}(I)) = \sum_{j\in J} (\mathbf{EC}(j) + |t^j - \lambda^j| |P_0^j - P_1^j|)$$

$$\Leftrightarrow \quad \mathfrak{I}_t(\mathbf{N}(I)) = \mathfrak{I}(\mathbf{N}(I)) + \sum_{j\in J} |t^j - \lambda^j| |P_0^j - P_1^j|$$

La second membre de cette somme étant toujours positif, on retrouve bien une inertie minimale pour le centre médian. Cependant, le second terme de cette expression ne correspond pas à une inertie et nous empêche d'obtenir, sur $\mathbf{B}^p$, une relation ayant le même sens que la relation de Huyghens sur $\mathbf{R}^p$.

3.4 INDEPENDANCE VIS A VIS DU CODAGE

Lorsque les deux modalités d'une variable binaire sont codées par deux valeurs quelconques a et b, on est amené à considérer l'espace $\mathbf{E}=\{a,b\}^p$. Sur cet **E**, la distance égale au nombre de composantes différentes entre deux points n'est plus la distance en valeurs absolues. Cependant, elle ne dépend pas des valeurs effectives de a et b et s'exprime, comme sur l'espace $\mathbf{B}^p$, de la façon suivante :

$$\forall\, x \text{ et } y \in \mathbf{E} \quad d(x,y) = \sum_{j \in J} \delta^j(x,y)$$

$$\text{où } \delta^j(x,y) = \begin{cases} 1 \text{ si } x^j \neq y^j \\ 0 \text{ sinon} \end{cases}$$

Nous montrons ici que cette distance fournit des caractéristiques équivalentes pour toutes valeurs a et b retenues pour le codage. En fait, les résultats obtenus sont indépendants des valeurs effectives de a et b.

3.4.1 Caractéristiques d'une variable binaire

On reprend les mêmes notations pour la variable x, l'ensemble I de n individus et pour l'échantillon $E(x)=\{(x_i, p_i), i \in I\}$. Seul change le codage de la variable binaire qui repose ici sur les deux valeurs a et b.

La valeur centrale de E(x) est tout réel λ minimisant la quantité :

$$\sum_{i \in I} \{p_i / x_i^j \neq \lambda\}$$

Si on note :

$$P_a = \sum_{i \in I} \{p_i / x_i^j = a\}$$

$$P_b = \sum_{i \in I} \{p_i / x_i^j = b\}$$

l'échantillon se ramène à l'ensemble des deux valeurs a et b munies respectivement des pondérations P_a et P_b. La valeur centrale est alors la valeur a ou b de plus forte pondération associée et n'est autre que la médiane de E(x).

La mesure de dispersion autour de la médiane s'écrit :

$$\sum_{i \in I} \{p_i / x_i^j \neq \lambda\} = \begin{cases} P_a \text{ si } \lambda = b \\ P_b \text{ si } \lambda = a \end{cases}$$

Elle ne dépend que des pondérations et non des valeurs effectives de a et b. Elle est égale à l'écart moyen EC(x) (tel qu'il a été défini pour le codage par 0 et 1).

3.4.2 Caractéristiques d'un nuage de points

On reprend la notation **N(I)** pour le nuage défini à partir du tableau de variables binaires **X(I,J)**. Le codage étant réalisé par les valeurs a et b, le nuage est inclus dans l'espace **E**.

Le point central du nuage est tout élément λ de $\mathbf{R}^p$ minimisant la quantité :

$$\sum_{i\in I} p_i\, d(x_i,\lambda) = \sum_{i\in I}\sum_{j\in J} p_i\, \delta^j(x_i,\lambda) = \sum_{i\in I}\sum_{j\in J} \{p_i / x_i^j \neq \lambda^j\}$$

Pour déterminer λ, il suffit de trouver, pour tout j, la valeur λ^j minimisant :

$$\sum_{i\in I} \{p_i / x_i^j \neq \lambda^j\}$$

Ce problème a une solution évidente : on choisit, pour λ^j, la valeur a ou b médiane de la variable j. Le point recherché apparait alors comme le centre médian du nuage N(I). De plus, ce point a la particularité d'appartenir à l'espace **E**.

Si λ est le centre médian du nuage, la caractéristique de dispersion s'exprime par :

$$\sum_{i\in I}\sum_{j\in J} \{p_i / x_i^j \neq \lambda^j\}$$

Encore une fois, cette quantité ne dépend que des pondérations des individus. En fait, cette mesure de dispersion apparait comme l'inertie du nuage $\mathfrak{I}(\mathbf{N}(I))$ (telle qu'elle a été définie pour la distance en valeurs absolues sur l'espace $\mathbf{B}^p$).

3.4.3 Généralisation

Nous allons maintenant considérer une distance plus générale dépendant d'un paramètre e. Pour toute valeur de e, on aboutit toujours au même type d'interprétation des caractéristiques associées à la distance. C'est ce que nous montrons dans la suite.

Considérons donc la famille de distances :

$$\forall\ x \text{ et } y \in \mathbf{E} \quad d_e(x,y) = \sum_{j\in J} \delta_e^j(x,y)$$

où

$$\delta_e^j(x,y) = \begin{cases} e \text{ si } x^j \neq y^j \\ 0 \text{ sinon} \end{cases} \quad \text{et} \quad e \in \mathbf{R}$$

Notons que, dans le cas particulier où e prend la valeur la - bl, d_e n'est autre que la distance en valeurs absolues sur **E**. Lorsque e prend la valeur 1, on retrouve la distance d égale au nombre de composantes différentes entre les deux points considérés.

Les caractéristiques associées à toute distance d_e sont équivalentes à celles associées à la distance d. En effet, un échantillon à valeurs dans {a,b} est caractérisé par sa médiane, un nuage de **E** est caractérisé par son centre médian appartenant également à **E**. Si on note **EC**(j,e) l'écart moyen de la variable j défini à partir de la distance d_e, on a :

$$\mathbf{EC}(j,e) = e \sum_{i\in I} \{ p_i / x_i^j \neq \lambda^j\} = e\, \mathbf{EC}(j)$$

De façon analogue, l'inertie diffère du facteur e de l'inertie définie dans le cas du codage à valeur dans {0,1}.

Remarquons que l'on obtient des caractéristiques analogues (médiane et centre médian) en choisissant pour e une valeur e^j dépendant de la variable j.

3.5 CONCLUSION

Dans ce paragraphe, nous avons mis en évidence quelques propriétés de l'espace $\mathbf{B}^p$ muni de la distance en valeurs absolues. Les caractéristiques associées présentent l'avantage d'être en accort avec la structure d'espace binaire. De plus, elles peuvent être facilement calculées.

Un application intéressante concerne, tout particulièrement, les tableaux de variables binaires : une variable est caractérisée par l'une des deux valeurs 0 ou 1 (la médiane); le tableau est résumé par un vecteur de $\mathbf{B}^p$ (le centre médian). De plus, les mesures de dispersion autour de ces deux caractéristiques s'interprétent simplement en terme de différences entre données initiales et valeurs médianes.

Tous ces résultats trouverons une application lorsque nous aborderons la classification sur tableau de variables binaires.

4. DESCRIPTION D'UN TABLEAU DE CODAGE ADDITIF

Nous nous intéressons maintenant aux tableaux de variables qualitatives ordinales et, plus particulièrement, à la transformation de ces tableaux par le codage binaire additif.

Dans un premier temps, nous précisons les notations utilisées pour le tableau de modalités $\mathbf{Z}(I,Q)$, croisant individus et variables qualitatives ordinales, et pour le nuage associé à ce tableau. A chaque variable correspond un ensemble de modalités qui sont généralement représentées par les valeurs $\{1,2,\ldots,m_q\}$ (où m_q représente le nombre de modalités de la variable q de Q). Nous définissons alors l'espace $\mathbf{E}$ comme le produit de ces ensembles de modalités (la $q^{ème}$ coordonnée de tout point de $\mathbf{E}$ appartient à l'ensemble $\{1,2,\ldots,m_q\}$). Le nuage défini à partir du tableau $\mathbf{Z}(I,Q)$ appartient à cet espace. Nous parlons alors de nuage de vecteurs de modalités.

Ensuite, après avoir décrit le codage binaire additif d'une variable, on donne les notations utilisées pour le tableau de codage $\mathbf{X}(I,J)$ et pour le nuage associé. Cette fois, le nuage est inclus dans l'espace $\mathbf{B}^m = \{0,1\}^m$ (où m est ici le nombre total de modalités des variables de Q). Les points de ce nuage ont une structure particulière : leurs composantes résultent de la transformation, par le codage binaire additif, du point de l'espace $\mathbf{E}$ correspondant. Dans ce cas, nous parlons de vecteurs binaires de modalités et nous appelons $\mathbf{F}$ la restriction de $\mathbf{B}^m$ à ces vecteurs.

Nous proposons alors de munir les espaces $\mathbf{E}$ et $\mathbf{F}$ de la distance en valeurs absolues. Nous montrons alors la propriété suivante : la distance entre deux vecteurs de modalités de $\mathbf{E}$ est égale à la distance entre les deux vecteurs binaires de modalités correspondants de l'espace $\mathbf{F}$.

Nous envisageons ensuite de déterminer les caractéristiques associées à la distance sur ces deux espaces. La caractéristique fondamentale sur l'espace $\mathbf{E}$ est la médiane. Dans le cas le plus général (pondérations quelconques), cette médiane ne répond à aucune règle de recherche simple (paragraphe 2.2.1). Par contre, sur l'espace $\mathbf{F}$, les caractéristiques de la distance sont facilement déterminées : il suffira d'utiliser la règle de calcul de la médiane binaire. Cette règle permet d'obtenir ici, sans y apporter aucune modification, des caractéristiques en accord avec la forme initiale des données et donc interprétables par rapport à celles-ci. En utilisant ce résultat et la propriété sur les distances, nous montrons

qu'il est possible de calculer, pour ces données qualitatives ordinales, les caractéristiques associées à la distance en valeurs absolues sur l'espace **E**.

4.1 NOTATIONS

4.1.1 Le tableau de modalités

Le tableau
Soit **Z**(I,Q) le tableau de modalités croisant un ensemble I={1,2,...,n} de n individus et un ensemble Q={1,2,...,p} de p variables qualitatives ordinales. Soit également J_q={1,2,...,m_q} l'ensemble des modalités de la variable q. On note :

$$\mathbf{Z}(I,Q) = (z_i^q)$$

où $z_i^q \in J_q$ représente la modalité de la variable q choisie par l'individu i.

Le nuage
A partir des n lignes du tableau **Z**(I,Q), on définit le nuage de n points pondérés $\mathbf{N}_Z(I)$ par :

$$\mathbf{N}_z(I) = \{ (z_i, p_i), i \in I \}$$

où $z_i = (z_i^1, z_i^2, \ldots, z_i^p)$ est le vecteur représentant l'individu i,

p_i la pondération associée à l'individu i et $\sum_{i \in I} p_i = 1$.

L'espace et la distance
Les données qualitatives ordinales peuvent être considérées comme des données quantitatives. Le nuage **N**(I) est alors plongé dans l'espace $\mathbf{R}^p$. Si cet espace est muni de la distance euclidienne usuelle, les caractéristiques mises en évidence sont : la moyenne d'une variable, le centre de gravité du nuage. Celles-ci ne sont pas nécessairement de la même forme que les données initiales (la moyenne n'est pas une modalité, le centre de gravité n'est pas un vecteur de modalités).

Nous proposons de plonger le nuage **N**(I) dans l'espace **E** défini comme le produit $J_1 x J_2 x \ldots x J_p$. Nous munissons **E** de la distance en valeurs absolues, notée d_E. Cette distance convient pour les données du type qualitatives ordinales et sa définition est :

$$\forall\, x \text{ et } y \in \mathbf{E} \quad d_E(x,y) = \sum_{q=1}^{p} |x^q - y^q|$$

La médiane est la caractéristique fondamentale associée à cette distance. Comme nous l'avons vu, un ensemble admet pour médiane une valeur figurant dans cet ensemble. C'est essentiellement cette propriété qui a motivé le choix de la distance en valeurs absolues. En effet, les valeurs intervenant ici étant des modalités, la médiane d'une variable qualitative ordinale est alors une modalité.

4.1.2 Le codage binaire additif

Le codage binaire additif est un codage dit à modalités ordonnées : il permet de rester cohérent avec l'ordre des modalités de la variable.

Soit z une variable qualitative ordinale à modalités dans {1,2,...,m}. Le codage binaire additif de z est une application **c** qui, à toute modalité j de z, associe un élément x de $\{0,1\}^m$. Nous pouvons la définir de la façon suivante :

$$\forall\, j \in \{1, 2, \ldots, m\} \quad \mathbf{c}(j) = x = (x^1, x^2, \ldots, x^m)$$

$$\text{où pout tout k on a } x^k = \begin{cases} 1 \text{ si } k \leq j \\ 0 \text{ si } k > j \end{cases}$$

Ce codage permet de recalculer la modalité codée. Si x est le codage d'une modalité j de z, on a :

$$\sum_{k=1}^{m} x^k = j$$

Considérons, par exemple, une variable à 5 modalités. Celles-ci sont alors codées de la façon suivante :

modalité	codage
1	1 0 0 0 0
2	1 1 0 0 0
3	1 1 1 0 0
4	1 1 1 1 0
5	1 1 1 1 1

4.1.3 Le tableau binaire de codage

Le tableau
Soit **X**(I,J) la transformation, par le codage binaire additif, du tableau de modalités **Z**(I,Q). L'ensemble J={1,2,...,m} (où m est le nombre total de modalités des variables de Q) permet d'indicer les colonnes de **X**(I,J).

Une colonne q de **Z**(I,Q) contient les valeurs prises par la variable q sur les individus de I. Cette colonne peut être transformé par le codage binaire additif. Soit $\mathbf{c}_q$ l'application qui, à toute modalité de la variable q, associe le vecteur de $\{0,1\}^{m_q}$ défini par :

$$\forall\, i \in I \quad \mathbf{c}_q(z_i^q) = (x_i^{q(1)}, x_i^{q(2)}, \ldots, x_i^{q(m_q)})$$

$$\text{où} \quad \forall\, j \in J_q \quad x_i^{q(j)} = \begin{cases} 1 \text{ si } j \leq z_i^q \\ 0 \text{ si } j > z_i^q \end{cases}$$

et où q(j), représentant l'indice de J correspondant à la modalité j de la variable q, est défini par :

$$q(j) = \sum_{k=1}^{q-1} m_k + j$$

Après avoir transformé les p colonnes de **Z**(I,Q), nous obtenons le tableau de codage binaire additif **X**(I,J).

Exemple
Considérons un ensemble de 6 individus identifiés par les chiffres **1** à **6**. Ceux-ci sont décrits par 3 variables qualitatives ordinales, identifiés par les lettres **a**, **b** et **c**. Supposons

que les modalités des variables soient {1,2,3} pour **a**, {1,2,3,4} pour **b** et {1,2,3,4,5} pour **c**. Les valeurs choisies par les individus sont représentées sous la forme d'un tableau indiqué en figure 1. Le tableau de codage binaire additif est représenté en figure 2.

	a	b	c
1	3	2	3
2	2	4	1
3	3	1	5
4	3	3	2
5	3	3	4
6	1	3	3

figure 1
tableau de modalités

1	1	1	1	1	1	0	0	1	1	1	0	0
2	1	1	0	1	1	1	1	1	0	0	0	0
3	1	1	1	1	0	0	0	1	1	1	1	1
4	1	1	1	1	1	1	0	1	1	0	0	0
5	1	1	1	1	1	1	0	1	1	1	1	0
6	1	0	0	1	1	1	0	1	1	1	0	0

figure 2
tableau de codage

Le nuage
A partir du tableau de codage, on définit le nuage **N**(I) par :

$$\mathbf{N}(I) = \{ (x_i, p_i), i \in I \}$$

où $x_i = (x_i^1, x_i^2, \ldots, x_i^m)$ est le vecteur représentant l'individu i,

p_i la pondération associée à l'individu i et $\sum_{i \in I} p_i = 1$.

Chaque point x_i du nuage **N**(I) résulte de la transformation (par le codage binaire additif) du point z_i appartenant au nuage $\mathbf{N_z}(I)$. Ces deux points sont liés par la relation permettant de retrouver z_i à partir de x_i, ce qui s'exprime ici par :

$$\forall\, i \in I, \forall\, q \in Q \quad \sum_{j \in J_q} x_i^{q(j)} = z_i^q$$

L'espace et la distance
Le nuage **N**(I) est inclus dans l'espace $\mathbf{B^m}$ que l'on restreint ici aux seuls vecteurs binaires de modalités. Nous appelons **F** cette restriction de $\mathbf{B^m}$. A tout point de cet espace correspond alors un point de **E** et réciproquement. On munit **F** de la distance en valeurs absolues d définie ici par :

$$\forall\, x \text{ et } y \in \mathbf{F} \quad d(x,y) = \sum_{j \in J} |x^j - y^j| = \sum_{q \in Q} \sum_{j \in J_q} |x^{q(j)} - y^{q(j)}|$$

Sur l'espace **F**, la distance n'a pas une signification très claire. Dans le paragraphe suivant, nous montrons une propriété permettant d'interpréter simplement cette distance.

4.2 PROPRIETE

Les deux approches que nous avons envisagées sont les suivantes :

(i) les données non codées sont plongés dans l'espace **E** muni de la distance d_E,

(ii) les données codées sont plongés dans l'espace **F** muni de la distance d.

La propriété énoncée et démontrée ci-après montre l'équivalence entre ces deux approches.

Propriété
*Soient d la distance en valeurs absolues sur **F** et d_E la même distance sur **E**. Soient z_i et $z_{i'}$ deux points de **E** et soient x_i et $x_{i'}$ leur codage respectif dans **F**. On a alors égalité entre les distances :*

$$d_E(z_i,z_{i'}) = d(x_i,x_{i'})$$

Preuve

La distance entre les deux points x_i et $x_{i'}$ de F s'écrit :

$$d(x_i,x_{i'}) = \sum_{j\in J} |x_i^j - x_{i'}^j| = \sum_{q\in Q}\sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}|$$

Les points x_i et $x_{i'}$ étant les transformations de z_i et $z_{i'}$, on a :

$$\sum_{j\in Jq} x_i^{q(j)} = z_i^q \quad \text{et} \quad \sum_{j\in Jq} x_{i'}^{q(j)} = z_{i'}^q$$

Pour tout q, on peut alors écrire :

$$\sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = \sum_{j\in Jq} x_i^{q(j)}(1-x_{i'}^{q(j)}) + \sum_{j\in Jq} x_{i'}^{q(j)}(1-x_i^{q(j)})$$

$$\Leftrightarrow \quad \sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = \sum_{j\in Jq} x_i^{q(j)} + \sum_{j\in Jq} x_{i'}^{q(j)} - 2\sum_{j\in Jq} x_i^{q(j)}x_{i'}^{q(j)}$$

$$\Leftrightarrow \quad \sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = z_i^q + z_{i'}^q - 2\,\mathrm{Inf}(z_i^q,z_{i'}^q)$$

$$\Leftrightarrow \quad \sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = |z_i^q - z_{i'}^q|$$

D'où l'égalité :

$$d(x_i,x_{i'}) = \sum_{q\in Q}\sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = \sum_{q\in Q} |z_i^q - z_{i'}^q| = d_E(z_i,z_{i'})$$

4.3 CARACTERISTIQUES D'UNE VARIABLE QUALITATIVE ORDINALE

Il s'agit de déterminer les caractéristiques d'une variable qualitative ordinale. Deux études sont envisagées, en considérant d'abord un échantillon de modalités, puis l'échantillon transformé par le codage binaire additif.

4.3.1 Etude de l'échantillon d'une variable

Soit z une variable qualitative à modalités ordonnées dans J={1,2,...,m}. On dispose de l'échantillon E(z) des valeurs prises par les n individus d'un ensemble I. On note toujours :

$$E(z) = \{ (z_i, p_i), i \in I \}$$

où z_i est la modalité choisie par l'individu i de pondération associée p_i.

Par définition, la médiane de E(z) est tout réel A minimisant la quantité :

$$\sum_{i\in I} p_i |z_i - A|$$

Dans le cas quantitatif, nous avons déjà énoncé plusieurs résultats (paragraphe 2.2.1 de ce chapitre). Ceux-ci restent valables lorsque la variable est qualitative ordinale.

En particulier, un échantillon admet pour médiane l'une des valeurs qu'il contient. Ces valeurs étant ici des modalités, la valeur médiane est alors une modalité. D'autre part, lorsque les mesures de l'échantillon sont munies de pondérations égales, on a le résultat suivant : la médiane ou l'intervalle médian sont définis à partir des valeurs "centrales" de l'échantillon. Pour des pondérations quelconques, cette règle ne peut plus être appliquée.

Dans la suite, nous proposons une nouvelle règle fournissant toujours une solution au problème de recherche de la médiane. Elle repose sur le codage binaire additif des modalités et sur la règle de calcul de la médiane binaire.

4.3.2 Etude de l'échantillon transformé par le codage binaire additif

Nous recherchons ici les caractéristiques associées à la distance en valeurs absolues lorsque les données suivent le codage binaire additif.

Considérons toujours l'échantillon E(z) d'une variable qualitative ordinale à m modalités dans J. Le codage binaire additif d'un élément z_i de E(z) est un point x_i de $\{0,1\}^m$ défini par :

$$x_i = (x_i^1, x_i^2, \ldots, x_i^m)$$

$$\text{où} \quad \forall j \in J \quad x_i^j = \begin{cases} 1 \text{ si } j \leq z_i \\ 0 \text{ si } j > z_i \end{cases}$$

Nous obtenons ainsi un ensemble de n points pondérés de $\{0,1\}^m$. Le centre médian de cet ensemble peut être facilement déterminé à partir de la règle de calcul de la médiane binaire. Notons a ce centre médian, la règle fournit alors le résultat suivant :

$$\forall j \in J \quad a^j = \text{médiane } \{ (x_i^j, p_i), i \in I \}$$

Les m composantes binaires du point a, prises séparément, n'ont aucune signification particulière. Cependant, si on s'intéresse à l'élément de $\{0,1\}^m$, une interprétation est possible, comme le montre la propriété énoncée ci-dessous.

Propriété
Soient z une variable qualitative ordinale à modalités dans $J=\{1,2,\ldots,m\}$ et $E(z)=\{(z_i, p_i), i \in I\}$ l'échantillon des valeurs observées sur un ensemble I de n individus. Soit également l'ensemble $\{(x_i, p_i), i \in I\}$ inclus dans $\{0,1\}^m$ tel que chaque x_i soit le codage binaire additif de la modalité z_i. Cet ensemble admet alors un centre médian correspondant au codage binaire additif d'une modalité de la variable. Cette modalité est la valeur médiane de l'échantillon E(z).

Preuve
Considérons le tableau de codage x(I,J) de la variable z (la notation **X** correspond au tableau de codage d'un ensemble de variables). Il s'agit du tableau d'ordre (n,m), où les n lignes sont définies à partir des n vecteurs binaires $\{(x_i, p_i), i \in I\}$.

La médiane d'une colonne j de ce tableau correspond à la $j^{ème}$ composante du centre médian a.

Pour ce tableau particulier, il résulte que :

- Si la médiane d'une colonne j de ce tableau est égale à 1, toute colonne d'indice $j' \leq j$ admet également pour médiane la valeur 1 (la colonne j contient des valeurs 1 qui se retrouvent toutes dans la colonne j', cette dernière contenant éventuellement d'autres valeurs 1).

- Si une colonne j est de médiane 0, toute colonne d'indice $j' \geq j$ est également de médiane 0 (la colonne j contient des valeurs 0 qui se retrouvent toutes dans la colonne j', cette dernière contenant éventuellement d'autres valeurs 0).

Dans ces conditions, le centre médian apparait comme le codage d'une modalité (les composantes de a vérifient le codage binaire additif). Soit A cette modalité que l'on peut déterminer de la façon suivante :

$$A = \sum_{j \in J} a^j$$

Il reste maintenant à préciser la signification de cette modalité.

Par définition, le vecteur a minimise la quantité :

$$\sum_{i \in I} \sum_{j \in J} p_i |x_i^j - a^j|$$

qui peut encore s'écrire :

$$\sum_{i \in I} p_i \sum_{j \in J} |x_i^j - a^j| = \sum_{i \in I} p_i |z_i - A|$$

Cette quantité étant minimale, la valeur A apparait alors comme la médiane de l'echantillon E(z). La propriété est alors démontrée.

Nous disposons donc d'une méthode permettant de calculer, dans tous les cas et pour des pondérations (p_i) quelconques, la médiane d'un échantillon de modalités. Elle fournit également une indication supplémentaire, en ce sens qu'elle permet de trouver soit une médiane unique (cas où toutes les colonnes de **x**(I,J) admettent chacune une médiane binaire unique), soit un intervalle médian (cas où plusieurs colonnes successives du tableau **x**(I,J) admettent un intervalle médian [0,1]). De plus, les bornes de cet intervalle sont bien des modalités apparaissant dans l'échantillon. Des exemples illustratifs sont proposés dans le prochain paragraphe.

Il nous reste à préciser la signification de l'écart moyen pour les données codées. Comme la médiane, l'écart moyen de l'ensemble des valeurs d'une colonne de **x**(I,J) n'a pas de sens particulier. Par contre, l'inertie de l'ensemble $\{(x_i, p_i), i \in I\}$ est une quantité intéressante puisque :

$$\Im(\{(x_i, p_i), i \in I\}) = \sum_{i \in I} \sum_{j \in J} p_i |x_i^j - a^j|$$

$$\Leftrightarrow \quad \Im(\{(x_i, p_i), i \in I\}) = \sum_{i \in I} p_i |z_i - A|$$

$$\Leftrightarrow \quad \Im(\{(x_i, p_i), i \in I\}) = \mathbf{EC}(z)$$

où **EC**(z) est l'écart moyen de la variable z

4.3.3 Exemples illustratifs

Considérons une variable qualitative ordinale à modalités dans {1,2,3,4,5}. Nous proposons d'appliquer les résulats précédents à différents échantillons de cette variable.

Exemple 1
Considérons ici un échantillon dont les valeurs sont munies de pondérations identiques. Ci-dessous, on représente l'échantillon et sa transformation par le codage binaire additif. La règle de calcul de la médiane binaire appliquée aux colonnes du tableau de codage fournit une médiane égale à 2.

échantillon	codage
1	1 0 0 0 0
1	1 0 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
3	1 1 1 0 0
3	1 1 1 0 0
4	1 1 1 1 0
5	1 1 1 1 1
médiane :	
2	1 1 0 0 0

Exemple 2
Considérons un échantillon dont les valeurs sont encore munies de pondérations identiques. L'échantillon et le tableau de codage sont indiqués ci-dessous. La règle ne fournit pas ici une médiane unique mais un intervalle médian [2,4].

échantillon	codage
1	1 0 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
2	1 1 0 0 0
4	1 1 1 1 0
4	1 1 1 1 0
5	1 1 1 1 1
5	1 1 1 1 1
5	1 1 1 1 1
médiane :	
[2, 4]	1 1 * * 0

Le symbole '*' ci-dessus signifie que la médiane en colonne est tout élément de l'intervalle [0,1]. Toute valeur comprise entre 2 (correspondant à une valeur 0 pour le symbole) et 4 (valeur 1 pour le symbole) est médiane de l'échantillon..

Exemple 3
Considérons maintenant un échantillon dont les valeurs sont munies des pondérations indiquées ci-dessous. Les valeurs binaires résultant du codage additif d'une modalité de

l'échantillon sont ainsi toutes munies de la même pondération, celle associée à cette modalité. La règle est alors appliquée à ces valeurs binaires pondérées et fournit une médiane égale à 2.

échantillon	pondérations	codage
1	$\frac{5}{20}$	1 0 0 0 0
2	$\frac{6}{20}$	1 1 0 0 0
3	$\frac{1}{20}$	1 1 1 0 0
4	$\frac{7}{20}$	1 1 1 1 0
5	$\frac{1}{20}$	1 1 1 1 1
médiane :		
2		1 1 0 0 0

4.4 CARACTERISTIQUES D'UN NUAGE DE POINTS

Les caractéristiques associées à la distance en valeurs absolues sont le centre médian d'un nuage et l'inertie. Nous reprenons ici ces notions pour le nuage $\mathbf{N_Z}(I)$ défini à partir du tableau $\mathbf{Z}(I,Q)$ et pour le nuage $\mathbf{N}(I)$ défini à partir du tableau $\mathbf{X}(I,J)$.

Par définition, les composantes du centre médian sont des valeurs médianes. Le nuage $\mathbf{N_Z}(I)$ est inclus dans l'espace **E** et admet alors pour centre médian un vecteur de modalités. Le nuage $\mathbf{N}(I)$ est inclus dans l'espace **F** et, d'après la propriété vue dans le paragraphe précédent, il admet pour centre médian un vecteur binaire de modalités appartenant à l'espace **F**. Dans ce paragraphe, nous énonçons une propriété exprimant le lien entre ces deux centres médians.

4.4.1 Propriété et caractéristiques

Propriété
*Soit a le point de **F** codage du point A de **E**. On a alors l'équivalence suivante :*

a centre médian du nuage N(I)

$\Leftrightarrow$

A centre médian du nuage $N_Z(I)$

Preuve
Si a est le centre médian du nuage $\mathbf{N}(I)$, il minimise la quantité :

$$\sum_{i \in I} p_i \, d(x_i, a)$$

Soit A le codage de a dans l'espace **E**. L'égalité entre la distance en valeurs absolues d sur **F** et d_E sur **E** permet d'écrire :

$$\sum_{i \in I} p_i d(x_i, a) = \sum_{i \in I} p_i d_E(z_i, A)$$

Cette quantité étant minimale, le point A apparait alors comme le centre médian du nuage $N_Z(I)$. Réciproquement, si A est le centre médian de $N_Z(I)$, le codage a de A est centre médian de N(I).

La dispersion d'un nuage autour de son centre médian est mesurée par l'inertie. Ici et comme le montre l'égalité précédente, il y a égalité entre l'inertie du nuage N(I) et celle de $N_Z(I)$. Les deux approches envisagées aboutissent ainsi à des résulats identiques.

4.4.2 Exemple d'application

Soit le tableau croisant 6 individus et 3 variables qualitatives ordinales (figure 1). Ces variables sont identifiées par les lettres **a**, **b** et **c** et possèdent respectivement 3, 4 et 5 modalités. Le tableau de codage est représenté en figure 2. Supposons également que les pondérations soient toutes égales à 1.

Pour chaque tableau, nous avons construit le nuage associé, recherché le centre médian et calculé l'inertie (celle-ci est à chaque fois égale à 13). Les deux centres obtenus, indiqués sous chaque tableau ci-dessous, sont des vecteurs résumant les données initiales.

	a	**b**	**c**
1	3	2	3
2	2	4	1
3	3	1	5
4	3	3	2
5	3	3	4
6	1	3	3
résumé :			
	3	3	3
inertie :			
	3	4	6 = 13

figure 1
tableau de modalités

1	1	1	1	1	1	0	0	1	1	1	0	0	
2	1	1	0	1	1	1	1	1	0	0	0	0	
3	1	1	1	1	0	0	0	1	1	1	1	1	
4	1	1	1	1	1	1	0	1	1	0	0	0	
5	1	1	1	1	1	1	0	1	1	1	1	0	
6	1	0	0	1	1	1	0	1	1	1	0	0	
	1	1	1	1	1	1	0	1	1	1	0	0	
	0	1	2	0	1	2	1	0	1	2	2	1	= 13

figure 2
tableau de codage

4.5 CONCLUSION

Nous avons mis en évidence l'équivalence entre les deux approches :

- tableau de modalités **Z**(I,Q) et espace **E** muni de la distance en valeurs absolues,

- tableau de codage binaire additif **X**(I,J) et espace **F** muni de la distance en valeurs absolues.

Dans les deux cas, le tableau initial peut être résumé par un vecteur de modalités facilement interprétable. De plus, ce vecteur est directement fourni par de règle de calcul de la médiane binaire.

Tous les résultats vus dans ce paragraphe trouveront une application en matière de classification pour ce type de données.

5. DESCRIPTION D'UN TABLEAU DISJONCTIF COMPLET

Dans cette partie, nous nous intéressons aux tableaux de variables qualitatives nominales et, plus particulièrement, à la transformation de ces tableaux par le codage disjonctif complet. Les tableaux de codage disjonctif complet peuvent être obtenus à partir des tableaux de modalités en transformant les modalités en variables binaires. Nous conservons toujours les valeurs 0 et 1 pour le codage de sorte que : la valeur 1 signifie que la modalité a été choisie, la valeur 0 codant alors l'autre possibilité. On peut encore considérer que le tableau de codage est celui croisant l'ensemble des individus et l'ensemble des indicatrices de toutes les modalités.

Nous précisons d'abord les notations utilisées pour le tableau $\mathbf{Z}(I,Q)$, croisant un ensemble I d'individus et un ensemble Q de variables qualitatives ordinales, et pour le nuage $\mathbf{N_Z}(I)$ associé. Ce nuage est inclus dans un espace **E** identique à celui défini pour des données qualitatives ordinales : il s'agit toujours de l'espace des vecteurs de modalités.

Ensuite, après avoir décrit le codage disjonctif complet, nous donnons les notations utilisées pour le tableau binaire $\mathbf{X}(I,J)$ et pour le nuage $\mathbf{N}(I)$ associé. Ce nuage est inclus dans un espace du type $\mathbf{B^m}$, où m est le nombre total de modalités. Nous nous intéressons ici aux seuls points dont les composantes vérifient la contrainte de codage disjonctif complet. Nous appelons **F** cette restriction de $\mathbf{B^m}$. Les éléments de **F** sont encore appelés vecteurs binaires de modalités (ils ont une structure différente de ceux définis dans le cas de données qualitatives ordinales, mais ils correspondent toujours à des codages de modalités).

Nous munissons ensuite les espaces **E** et **F** de la distance égale au nombre de composantes différentes entre deux points. Nous démontrons alors une relation liant la distance sur **E** et celle sur **F**. Sur l'espace **E**, la distance permet de définir des caractéristiques en accord avec la forme initiale des données. En particulier, on montre que la valeur centrale de l'échantillon d'une variable est l'une de ses modalités. Sur l'espace **F**, la distance devient la distance en valeurs absolues.

Contrairement au codage binaire additif, la règle de calcul de la médiane binaire ne fournit pas ici des résultats compatibles avec le codage disjonctif complet. Pour cela, il est nécessaire d'introduire la contrainte de codage (le domaine de recherche est limité à l'espace **F**). Sous ces conditions, cette approche aboutit à des caractéristiques équivalentes à celles obtenues sur l'espace **E**.

5.1 NOTATIONS

5.1.1 Tableau de modalités

Le tableau
Soit $\mathbf{Z}(I,Q)$ le tableau de modalités croisant un ensemble $I=\{1,2,\ldots,n\}$ de n individus et un ensemble $Q=\{1,2,\ldots,p\}$ de p variables qualitatives nominales. On note :

$$\mathbf{Z}(I,Q) = (z_i^q)$$

où

z_i^q est la modalité de la variable q choisie par l'individu i,

$z_i^q \in J_q=\{1, 2, \ldots , m_q\}$ l'ensemble des modalités de la variable q.

Le nuage
A partir du tableau **Z**(I,Q), on définit le nuage de n points pondérés $\mathbf{N_Z}(I)$ par :

$$\mathbf{N_Z}(I) = \{ (z_i, p_i), i \in I \}$$

où

$z_i = (z_i^1, z_i^2, \ldots, z_i^p)$ est le vecteur représentant l'individu i,

p_i la pondération associée à l'individu i et $\sum_{i \in I} p_i = 1$.

L'espace et la distance
Appelons **E** l'espace défini comme le produit $J_1 x J_2 x \ldots x J_p$. Le nuage $\mathbf{N_Z}(I)$ est alors inclus dans **E**. Cet espace est analogue à celui défini dans le cas ordinal où il était muni de la distance en valeurs absolues. Cependant, pour des données qualitatives nominales, cette distance ne permet pas d'évaluer correctement les proximités entre individus. En effet, les modalités d'une variable n'étant pas ordonnées, cette distance n'a plus aucun sens ici.

Nous proposons de munir **E** d'une nouvelle distance d_E convenant mieux pour ce type de données. Il s'agit de la distance égale au nombre de composantes différentes entre deux points :

$$\forall x \text{ et } y \in \mathbf{E} \quad d_E(x,y) = \sum_{q \in Q} \delta^q(x,y)$$

$$\text{où} \quad \delta^q(x,y) = \begin{cases} 1 \text{ si } x^q \neq y^q \\ 0 \text{ sinon} \end{cases}$$

Cette distance présente l'avantage d'être facilement interprétable. En outre, comme nous le montrons dans les prochains paragraphes, les caractéristiques associées sont en accord avec la forme initiale des données. En particulier, la distance permet de caractériser une variable par une modalité et un nuage de points par un vecteur de modalités. Dans la suite, nous précisons la signification de ces caractéristiques.

5.1.2 Le codage disjonctif complet

Soit z une variable qualitative à modalités dans J={1,2,...,m}. Le codage disjontif complet de la variable z est obtenu en associant à chacune des modalités une variable binaire. Parmi ces m variables binaires, une seule peut prendre la valeur 1 (celle choisie par un individu), les autres prenant alors la valeur 0.

On peut définir ce codage comme une application **c** qui, à toute modalité j de J, associe un élément x de $\{0,1\}^m$. La définition de **c** est alors la suivante :

$$\forall j \in J \quad \mathbf{c}(j) = x = (x^1, x^2, \ldots, x^m)$$

$$\text{où} \quad \forall k=1,\ldots,m \quad x^k = \begin{cases} 1 \text{ si } k = j \\ 0 \text{ si } k \neq j \end{cases}$$

La contrainte s'exprime alors par :

$$\sum_{j \in J} x^j = 1$$

Considérons, par exemple, une variable à 5 modalités. Celles-ci sont alors codées de la façon suivante :

modalité	codage
1	1 0 0 0 0
2	0 1 0 0 0
3	0 0 1 0 0
4	0 0 0 1 0
5	0 0 0 0 1

5.1.3 Le tableau de codage

Le tableau
Soit **X**(I,J) la transformation, par le codage binaire additif, du tableau de modalités **Z**(I,Q). L'ensemble J={1,2,...,m} (où m est le nombre total de modalités des variables de Q) permet d'indicer les colonnes de **X**(I,J).

Une colonne q de **Z**(I,Q) contient les valeurs prises par la variable q sur l'ensemble I des individus. Chaque élément de cette colonne peut être transformé par le codage. Soit $\mathbf{c}_q$ qui, à toute modalité de la variable q, associe un vecteur de $\{0,1\}^{m_q}$ de sorte que :

$$\forall\, i \in I \quad \mathbf{c_q}(z_i^q) = (x_i^{q(1)}, x_i^{q(2)}, \ldots, x_i^{q(m_q)})$$

$$\text{où } \forall\, j \in J_q \quad x_i^{q(j)} = \begin{cases} 1 \text{ si } j = z_i^q \\ 0 \text{ si } j \neq z_i^q \end{cases}$$

et où q(j), l'indice de J correspondant à la modalité j de la variable q, est défini par :

$$q(j) = \sum_{k=1}^{q-1} m_k + j$$

Après avoir effectué les transformations des p colonnes de **Z**(I,Q), on obtient le tableau de codage disjonctif complet **X**(I,J).

On peut également considérer que J est l'ensemble des m variables binaires correspondant aux m modalités des p variables de Q. Le tableau **X**(I,J) est alors construit de la façon suivante :

$$x_i^j = \begin{cases} 1 \text{ si l'individu i à choisit la modalité j} \\ 0 \text{ sinon} \end{cases}$$

Une conséquence de ce type de codage est que la marge en ligne du tableau binaire est constante :

$$\forall\, q \in Q \quad \sum_{j \in Jq} x_i^{q(j)} = 1 \quad \text{et} \quad \forall_i i \in I \quad \sum_{j \in J} x_i^j = p$$

Exemple
Considérons un ensemble de 6 individus identifiés par les chiffres **1 à 6**. Ceux-ci sont décrits par 3 variables qualitatives nominales identifiées par les lettres **a**, **b** et **c**. Supposons que les modalités des variables soient {1,2,3} pour **a**, {1,2,3,4} pour **b** et {1,2,3,4,5} pour **c**. Les valeurs observées sur les individus sont représentées sous la forme d'un tableau indiqué en figure 1. Le tableau de codage disjonctif complet est indiqué dans la figure 2. Nous avons choisi d'identifier une variable binaire (associée à

une modalité) par l'identificateur de la variable suivi, en indice, de la modalité (par exemple, la variable binaire asssociée à la modalité 1 de **a** est identifiée par $\mathbf{a_1}$).

	a	b	c
1	3	2	3
2	2	4	1
3	3	1	5
4	3	3	2
5	3	3	4
6	1	3	3

figure 1
tableau de modalités

	a_1	a_2	a_3	b_1	b_2	b_3	b_4	c_1	c_2	c_3	c_4	c_5
1	0	0	1	0	1	0	0	0	0	1	0	0
2	0	1	0	0	0	0	1	1	0	0	0	0
3	0	0	1	1	0	0	0	0	0	0	0	1
4	0	0	1	0	0	1	0	0	1	0	0	0
5	0	0	1	0	0	1	0	0	0	0	1	0
6	1	0	0	0	0	1	0	0	0	1	0	0

figure 2
tableau de codage

Le nuage
A partir du tableau **X**(I,J), on définit le nuage **N**(I) par :

$$\mathbf{N}(I) = \{ (x_i, p_i), i \in I \}$$

où $x_i = (x_i^1, x_i^2, \ldots, x_i^m)$ est le vecteur représentant l'individu i,

p_i la pondération associée à l'individu i et $\sum_{i \in I} p_i = 1$.

Chaque point x_i du nuage **N**(I) résulte de la transformation, par le codage disjonctif complet, du point z_i appartenant au nuage $\mathbf{N_z}(I)$. Le lien entre ces deux points est alors le suivant :

$$\forall\, i \in I,\ \forall\, q \in Q,\ \forall\, j \in J_q \quad x_i^{q(j)} = \begin{cases} 1 \text{ si } z_i^q = j \\ 0 \text{ sinon} \end{cases}$$

L'espace et la distance
Le nuage **N**(I) est inclus dans l'espace $\mathbf{B^m}$. Plus précisément, le nuage est inclus dans la restriction de cet espace aux seuls points dont les composantes vérifient la contrainte de codage disjonctif complet. Soit **F** cet espace contenant des vecteurs binaires de modalités. Ainsi, à tout vecteur de modalités de **E** va correspondre un unique vecteur binaire de modalités de **F** et réciproquement.

Nous proposons de munir **F** de la distance d égale au nombre de composantes différentes entre les deux points considérés. Sur cet espace, il s'agit de la distance en valeurs absolues :

$$\forall\, x \text{ et } y \in \mathbf{F} \quad d(x,y) = \sum_{j \in J} |x^j - y^j| = \sum_{q \in Q} \sum_{j \in J_q} |x^{q(j)} - y^{q(j)}|$$

Il nous reste à préciser la signification de cette distance.

5.2 PROPRIETE

Les deux approches que nous envisageons ici sont :

(i) les données non codées sont plongées dans l'espace **E** muni de la distance d_E,

(ii) les données codées sont plongées dans l'espace **F** muni de la distance d.

La propriété ci-dessous montre l'équivalence entre ces deux approches.

Propriété
*Soient z_i et $z_{i'}$ deux points de **E** et leur correspondant respectif x_i et $x_{i'}$ dans **F**. On a alors l'égalité suivante :*

$$d_E(z_i,z_{i'}) = 2\ d(x_i,x_{i'})$$

Preuve
Dans **F**, on a :

$$d(x_i,x_{i'}) = \sum_{j\in J} |x_i^j - x_{i'}^j| = \sum_{q\in Q}\sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}|$$

Puisque :

$$\forall\, q\in Q \quad \sum_{j\in Jq} |x_i^{q(j)} - x_{i'}^{q(j)}| = \begin{cases} 2 \text{ si } z_i^q \neq z_{i'}^q \\ 0 \text{ sinon} \end{cases} = 2\,\delta^q(z_i,z_{i'})$$

on en déduit :

$$d(x_i,x_{i'}) = 2 \sum_{q\in Q} \delta^q(z_i,z_{i'}) = 2\ d_E(z_i,z_{i'})$$

5.3 CARACTERISTIQUES D'UNE VARIABLE QUALITATIVE NOMINALE

Nous nous intéressons ici aux caractéristiques d'une variable qualitative nominale. Deux études sont envisagées, en considérant d'abord un échantillon de modalités, puis l'échantillon transformé par le codage disjonctif complet.

5.3.1 Etude de l'échantillon d'une variable

Soient z une variable qualitative nominale à modalités dans J={1,2,...,m} et E(z) l'échantillon des valeurs observées sur un ensemble I de n individus. On note :

$$E(z) = \{\ (z_i, p_i), i \in I\ \}$$

où z_i est la modalité choisie par l'individu i de pondération associée p_i.

A chaque modalité j de J, nous pouvons associer la pondération p^j définie par :

$$p^j = \sum_{i\in I} \{\ p_i / z_i=j\ \}$$

A partir de E(z), nous pouvons construire un nouvel échantillon composé des m modalités de z de la façon suivante :

$$\{\ (j, p^j), j \in J\ \}$$

Si une modalité n'est jamais observée, la pondération associée sera nulle.

Propriété
La caractéristique de valeur centrale associée à la distance d_E est la modalité de plus forte pondération associée. Si tous les individus sont munis de pondérations égales, la caractéristique est la modalité majoritaire relative de l'échantillon.

Preuve
La valeur centrale de l'échantillon est tout nombre A minimisant la quantité :

$$\sum_{i\in I} \{p_i / z_i \neq A\} = \sum_{j\in J} \{p^j / j \neq A\}$$

La solution évidente est de choisir pour A la modalité j de z de plus forte pondération associée.

La dispersion de l'échantillon E(z) autour de cette modalité s'exprime par :

$$\sum_{i\in I} \{p_i / z_i \neq A\}$$

où A est la modalité caractérisant l'échantillon. Il s'agit de la somme des pondérations des modalités différentes de A. On dispose ainsi de deux caractéristiques facilement interprétables par rapport aux données initiales.

Cas où les individus sont munis de pondération égales
Supposons que tous les individus soient munis d'une même pondération :

$$\forall\, i\in I \quad p_i = \rho$$

Dans ce cas, la pondération d'une modalité j de z est égale, à un facteur constant près, au nombre d'individus ayant choisi la modalité j :

$$\forall\, j\in J \quad p^j = \sum_{i\in I} \{\rho / z_i = j\} = \rho\, n^j$$

où n^j est le nombre d'individus ayant choisi j.

La caractéristique de valeur centrale est donc la modalité la plus souvent présente dans E(z). Il s'agit de la modalité majoritaire relative de l'échantillon. La mesure de dispersion associée à E(z) est alors égale au nombre d'individus ayant choisi une modalité différente de la modalité majoritaire relative.

5.3.2 Etude de l'échantillon transformé par le codage disjonctif complet

Nous recherchons ici les caractéristiques de l'échantillon E(z) transformé par le codage disjonctif complet.

Le codage d'une modalité z_i de E(z) est un élément x_i de $\{0,1\}^m$ défini par :

$$x_i = (x_i^1, x_i^2, \ldots, x_i^m)$$

$$\text{où} \quad \forall\, j\in J \quad x_i^j = \begin{cases} 1 \text{ si } j = z_i \\ 0 \text{ si } j \neq z_i \end{cases}$$

A l'échantillon E(z) correspond donc un ensemble de n points de $\{0,1\}^m$. On peut facilement calculer le centre médian de ces n points en utilisant la règle de calcul de la médiane binaire. Le résultat est alors un point de $\{0,1\}^m$ qui n'est pas nécessairement le codage d'une modalité.

Soient z une variable qualitative nominale à 5 modalités et E(z)={1,2,2,4,5,3,2,2,5,1} l'échantillon des observations. Le codage de E(z) est indiqué ci-dessous (où la notation

z_1 représente la modalité 1 de la variable z, la notation z_2 la modalité 2, etc...). Si toutes les pondérations sont égales, la règle appliquée aux colonnes du tableau de codage fournit le vecteur (0,0,0,0,0) qui ne correspond à aucune modalité.

	z_1	z_2	z_3	z_4	
	1	0	0	0	0
	0	1	0	0	0
	0	1	0	0	0
	0	0	0	1	0
	0	0	0	0	1
	0	0	1	0	0
	0	1	0	0	0
	0	1	0	0	0
	0	0	0	0	1
	1	0	0	0	0
médianes des colonnes :					
	0	0	0	0	0

Afin d'obtenir un résultat du même type que les éléments de l'échantillon, il est nécessaire d'imposer la contrainte de codage. Le problème à résoudre est alors de trouver le point a de $\{0,1\}^m$ minimisant :

$$\sum_{i\in I} p_i \sum_{j\in J} |x_i^j - a^j| \quad \text{avec} \quad \sum_{j\in J} a^j = 1$$

Si on note A la modalité codée par le point a, la quantité à minimiser peut s'exprimer par :

$$2\sum_{i\in I} \{p_i / z_i \neq A\}$$

Le problème est alors identique à celui posé dans le cas de l'échantillon non codé. La solution est de choisir pour a le codage de la modalité A de plus forte pondération associée (si tous les individus sont munis de pondérations égales, A est la modalité majoritaire relative de l'échantillon). La mesure de dispersion de l'échantillon autour de cette valeur caractéristique est alors :

$$\sum_{i\in I} \{p_i / z_i \neq A\} = \frac{1}{2} \sum_{i\in I} \sum_{j\in J} p_i |x_i^j - a^j|$$

Reprenons l'échantillon E(z)={1,2,2,4,5,3,2,2,5,1}. Pour des pondérations toutes égales à 1, on obtient les résultats suivants :

- la modalité 2 pour valeur centrale,
- une mesure de dispersion égale à 12.

La mesure 12 indique que 6 valeurs de E(z) sont différentes de la modalité 2.

5.4 CARACTERISTIQUES D'UN NUAGE DE POINTS

Nous étudions ici les caractéristiques du nuage $\mathbf{N_Z(I)}$, inclus dans l'espace **E** muni de la distance d_E, et celles du nuage **N(I)**, inclus dans l'espace **F** muni de la distance en valeurs absolues d.

5.4.1 Cas du nuage associé au tableau de modalités

Considérons le nuage $\mathbf{N_Z(I)}$ défini à partir du tableau de modalités $\mathbf{Z(I,Q)}$.

Recherchons tout d'abord le point A minimisant :

$$\sum_{i\in I} p_i\, d_E(z_i,A) = \sum_{i\in I}\sum_{q\in Q} p_i\, \delta^q(z_i^q,A^q)$$

Pour cela, il suffit de déterminer, pour tout q de Q, la composante A^q minimisant :

$$\sum_{i\in I} p_i\, \delta^q(z_i^q,A^q) = \sum_{i\in I} \{p_i / z_i^q \neq A^q\}$$

On retrouve se retrouve dans les mêmes conditions que prédemment. La solution est donc de choisir pour A^q la modalité de q de plus forte pondération associée. Dans le cas où les individus sont munis de pondérations égales, les composantes du point A sont les modalités majoritaires relatives des variables de Q.

La caractéristique recherchée est donc un point appartenant à l'espace **E** contenant également le nuage. La seconde caractéristique est la mesure de dispersion du nuage autour de son point central A. Sa définition est :

$$\sum_{i\in I} p_i\, d_E(z_i,A) = \sum_{i\in I} \{p_i / z_i^q \neq A^q\}$$

Il s'agit d'une mesure de la présence des valeurs différentes des modalités caractéristiques des variables (elles ont une contribution non nulle). Pour des pondérations p_i toutes égales, elle représente (à un facteur constant près) le nombre de valeurs du tableau qui sont différentes des modalités majoritaires relatives des variables.

5.4.2 Cas du nuage associé au tableau de codage

Reprenons le nuage $\mathbf{N(I)}$ défini à partir du tableau disjonctif complet $\mathbf{X(I,J)}$. Les caractéristiques de ce nuage, pour la distance en valeurs absolues d, sont le centre médian et l'inertie. Le nuage étant inclus dans l'espace $\mathbf{B^m}$, il admet pour centre médian un point de ce même espace, obtenu en appliquant la règle de calcul de la médiane binaire. Pour les mêmes raisons que précédemment, ce point n'appartient pas nécessairement à l'espace **F** des vecteurs binaires de modalités. Pour rester en accord avec les données initiales, nous imposons à ce point d'appartenir à l'espace **F**.

On cherche donc l'élément a de **F** minimisant :

$$\sum_{i\in I} p_i\, d(x_i,a)$$

Au point a correspond le point A de **E**. La quantité à minimiser devient :

$$2\sum_{i\in I} p_i\, d_E(z_i,A)$$

Par définition, celle-ci est minimale pour A, vecteur de modalités caractérisant le nuage $\mathbf{N_Z(I)}$. Le point a apparait alors comme le codage de A.

Finalement, que les données initiales soient plongées dans l'espace **E** muni de la distance d_E ou, après transformation, dans l'espace **F** muni de la distance d, nous aboutissons à des caractéristiques identiques (au facteur 2 près pour la mesure de dispersion).

5.4.3 Exemple d'application

Reprenons le tableau déjà présenté (paragraphe 5.1.3) et croisant 6 individus et 3 variables qualitatives ordinales. Ces variables, identifiées par les lettres **a**, **b** et **c**, possèdent respectivement 3, 4 et 5 modalités. Supposons également que toutes les pondérations soient égales à 1. Dans la figure ci-dessous, nous indiquons le tableau de modalités pour lequel nous allons rechercher les caractéristiques. Pour le tableau de codage disjonctif complet, nous aurions évidement obtenu des résultats identiques.

Le vecteur de modalités résumant le tableau est indiqué sous celui-ci, de manière à mettre en évidence la modalité caractéristique de chaque variable. La mesure de dispersion, égale à 9, montre que le résumé est en désaccord sur 9 valeurs avec le tableau initial.

	a	**b**	**c**
1	3	2	3
2	2	4	1
3	3	1	5
4	3	3	2
5	3	3	4
6	1	3	3
résumé :			
	3	3	3
dispersion :			
	2	3	4 = 9

5.5 CONCLUSION

Contrairement au cas du tableau de codage binaire additif, nous avons eu recours à une contrainte afin d'obtenir des résultats interprétables par rapport aux données initiales. Les caractéristiques de description du tableau de codage disjonctif complet sont équivalentes à celles que l'on obtient en travaillant directement sur le tableau de modalités. Dans les deux cas, une variable est caractérisée par une de modalité et le tableau de codage ou de modalités est résumé par un vecteur de modalités facilement interprétable. A chaque fois, une règle simple permet de déterminer ces caractéristiques. Lorsque les individus sont munis de pondérations identiques, la règle devient la règle de la majorité relative.

Lorsque nous aborderons la partie classification (sur tableau de codage disjonctif complet), tous ces résultats trouveront une application.

CHAPITRE 2

CLASSIFICATION SUR TABLEAU DE VARIABLES BINAIRES

1. INTRODUCTION

La méthode de classification des Nuées Dynamiques (E. Diday et al. 1980) repose essentiellement sur l'utilisation de la notion de noyau associé à chaque classe. La nature de ce noyau peut être très variée. Dans le cas le plus simple et si les variables sont quantitatives, le noyau est un élément de l'espace $\mathbf{R}^p$ contenant l'ensemble à classifier. Cependant, pour des variables binaires, cette approche ne tient pas compte de la forme particulière des données. Dans ce cas, ce type de méthode fournit des noyaux ayant une structure différente des données initiales.

Nous proposons ici une méthode de classification automatique sur tableau de variables binaires. Celle-ci est construite de manière à répondre à l'idée d'intégrité des résultats par rapport aux données initiales. Le principe des méthodes de type Nuées Dynamiques permet de répondre à cette idée. En effet, nous avons la possibilité d'ajouter facilement des contraintes aux noyaux pour lui imposer d'avoir la même structure que les données initiales. Les données étant binaires, chaque élément de l'ensemble à classifier peut être représenté par un vecteur de l'espace $\mathbf{B}^p=\{0,1\}^p$. Il est naturel de vouloir représenter chaque classe par un élément de cet ensemble. Il reste alors à choisir une métrique sur cet espace $\mathbf{B}^p$. Nous proposons d'utiliser comme distance entre deux individus le nombre de valeurs différentes pour les deux vecteurs binaires correspondants. Sur l'espace $\mathbf{B}^p$, il s'agit exactement de la distance en valeurs absolues dont nous avons étudié, dans le chapitre précédent, les principales caractéristiques associées. L'algorithme devient alors particulièrement simple et le critère qui en résulte facile à interpréter.

J.C. Gower (1974) envisageait également de représenter chaque classe par un vecteur binaire. Il ne proposait qu'un algorithme d'échange minimisant un critère analogue à celui que nous utilisons, mais introduit à partir de la notion de prédiction. Le travail réalisé ici est une extension des travaux de J.C. Gower (1974).

Dans le premier paragraphe, nous rappelons rapidement le principe des méthodes des Nuées Dynamiques. Puis, après avoir posé le problème de la classification d'un tableau de variables binaires, nous décrivons l'algorithme construit qui répond à nos exigences. Cet algorithme est appelé MNDBIN : Méthodes des Nuées Dynamiques sur tableau de variables BINaires. Nous indiquons ensuite la signification du critère associé à la partition et celle des noyaux fournis par l'algorithme. Nous définissons également des indices permettant une analyse plus fine des résultats. Enfin, nous indiquons le lien existant entre la méthode proposée ici et la notion de modèle statistique. G. Govaert (1988) montre comment l'algorithme de classification MNDBIN est lié à un modèle précis de mélange de distributions de Bernouilli. Cette approche permet de justifier, a posteriori, l'utilisation pour les données binaires, d'une part de la distance en valeurs absolues, d'autre part de noyaux binaires.

La distance en valeurs absolues est une distance de la famille de Minkowski. Nous étendons alors la méthode des Nuées Dynamiques (sous contrainte de noyaux binaires) à

cette famille de distance. Nous nous intéressons, plus particulièrement, à la méthode utilisant la distance euclidienne usuelle et les centres de gravité comme noyaux. Il s'agit de la Méthode des Nuées Dynamiques sur tableau de variables QuANtitatives appelée MNDQAN (G. Celeux et al 1989) et proposée dans le logiciel d'analyse de données SICLA. Dans cette approche, les données binaires sont traitées comme des données quantitatives et l'ensemble à classifier est considéré comme inclus dans l'espace $\mathbf{R}^p$. Nous comparons ensuite cette méthode avec l'algorithme MNDBIN.

Dans une dernière partie, nous présentons le programme réalisant la classification sur tableau de variables binaires. Ce programme appelé MNDBIN a été intégré au logiciel SICLA. Une application sur un exemple concret est ensuite proposée. Après avoir décrit le tableau binaire étudié, on donnera et interprétera les résultats obtenus. Enfin, on applique la méthode à des tableaux créés par un programme générant des données binaires à partir de lois de Bernoulli.

2. LA MÉTHODE DES NUÉES DYNAMIQUES

Nous rappelons rapidement le principe de la Méthode des Nuées Dynamiques (E. Diday et al. 1980). Considérons un ensemble I de n individus représentés par un ensemble Ω de n points inclus dans un espace **E** (par exemple $\mathbf{R}^p$). On définit un ensemble de noyaux **L**, une distance **D** entre les éléments de **E** et les noyaux de **L**. Le critère **W** de la classification est alors le suivant :

$$\mathbf{W(P,L)} = \sum_{k=1}^{K} \sum_{i \in P_k} \mathbf{D}(x_i, a_k)$$

où

$\mathbf{P} = (P_1, P_2, \ldots, P_K)$ est une partition de l'ensemble I,

$\mathbf{L} = (a_1, a_2, \ldots, a_K)$ est l'ensemble des noyaux des classes de P.

L'algorithme construit itérativement une suite $P^0, L^0, P^1, L^1, \ldots, P^n, L^n$ de partitions et de noyaux en minimisant à chaque étape le critère. Cette construction repose sur la définition des deux fonctions suivantes :

- la fonction d'affectation (notée **f**) : elle consiste à affecter chaque individu à l'une des classes de la partition de manière à optimiser, à chaque fois, le critère **W(f(L),L)**. Bien sûr, elle repose sur le choix de la distance **D**.

- la fonction de représentation (notée **g**) : elle consiste à déterminer les noyaux de la partition de manière à optimiser, à chaque fois, le critère **W(P,g(P))**.

On obtient ainsi à la convergence une partition avec comme résumé de chaque classe le noyau associé.

3. APPLICATION AU TABLEAU DE VARIABLES BINAIRES

Pour utiliser l'algorithme de type Nuées Dynamiques dans le cas de données binaires, on peut soit considérer que l'ensemble des données appartient à l'espace $\mathbf{R}^p$ muni de la distance euclidienne et appliquer la méthode des nuées dynamiques en prenant comme noyaux des éléments de $\mathbf{R}^p$ (méthode des centres mobiles), soit se placer directement dans l'ensemble Ω (l'ensemble des points représentatif des individus) muni d'une

distance quelconque et utiliser la méthode des nuées dynamiques sur tableau de distance (cela revient à imposer aux noyaux d'appartenir à l'ensemble Ω).

Les deux situations précédentes présentent des inconvénients. Dans le premier cas, le résumé de chaque classe et le critère sont difficilement interprétables par rapport aux données initiales. En effet, on a pas tenu compte de la forme particulière des données. Dans le second cas, l'appartenance des noyaux à l'ensemble à classifier peut sembler cette fois trop restrictive. De plus un tel algorithme nécessite la construction puis l'utilisation du tableau des distances sur Ω, ce qui va conduire à une certaine perte d'efficacité en place mémoire et en temps.

La méthode que nous proposons ici va se situer entre ces deux approches. Pour cela, nous utilisons la possibilité d'imposer des contraintes aux noyaux.

3.1 NOTATIONS

Soit $\mathbf{X}(I,J)$ le tableau croisant un ensemble $I=\{1,2,\ldots,n\}$ de n individus et un ensemble $J=\{1,2,\ldots,p\}$ de p variables binaires. On note :

$$\mathbf{X}(I,J) = (x_i^j)$$

où x_i^j représentant la valeur 0 ou 1 prise par l'individu i sur la variable j.

A partir de ce tableau, nous définissons le nuage $\mathbf{N}(I)$ par :

$$\mathbf{N}(I) = \{ x_i, i \in I \}$$

$$\text{où} \quad x_i = (x_i^1, x_i^2, \ldots, x_i^p)$$

Nous n'avons supposé aucune pondération sur les individus. La pondération la plus souvent choisie est 1/n. Le fait d'utiliser celle-ci ne change pas le sens des résultats que nous obtenons sous nos conditions.

Chaque individu i est représenté par un vecteur x_i. L'ensemble des individus est alors représenté par le nuage $\mathbf{N}(I)$ dans l'espace $\mathbf{B}^p$.

Nous munissons cet espace de la distance en valeurs absolues d définie par :

$$\forall\, x \text{ et } y \in \mathbf{B}^p \quad d(x,y) = \sum_{j=1}^{p} |x^j - y^j|$$

où les x^j et les y^j sont les coordonnées des points x et y.

3.2 LE PROBLEME

Il s'agit de déterminer une partition de l'ensemble I en K classes, K étant fixé a priori.

L'algorithme que nous allons écrire doit respecter un principe d'homogénéité : les données à classer et les noyaux doivent être de même nature. L'ensemble à classifier étant représenté par le nuage $\mathbf{N}(I)$ dans $\mathbf{B}^p$, nous imposons aux noyaux de la méthode d'appartenir à ce même espace.

Le problème à résoudre est alors le suivant :

trouver le couple (P,L) où

$P = (P_1, P_2, \ldots, P_K)$ *est une partition de l'ensemble I,*

$L = (a_1, a_2, \ldots, a_K)$ *est l'ensemble des noyaux appartenant à* B^p,

tel que le critère

$$W(P,L) = \sum_{k=1}^{K} \sum_{i \in Pk} d(x_i,a_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$

soit minimum.

3.3 L'ALGORITHME

L'algorithme se construit de la façon habituelle : il s'agit de déterminer la fonction d'affectation **f** telle que **W**(**f**(L),L) soit minimum et la fonction de représentation **g** telle que **W**(P,**g**(P)) soit minimum.

La fonction d'affectation f
Etant donné un ensemble L de K noyaux, un individu i (représenté par x_i) est affecté à la classe P_k dont il est le plus proche du noyau (a_k) au sens de la distance en valeurs absolues.

La fonction de représentation g
Il s'agit de déterminer l'ensemble $L=(a_1,a_2,\ldots,a_K)$ des noyaux des classes de la partition P tel que le critère **W**(P,L) soit minimum. Pour cela, il suffit de rechercher, pour chaque classe P_k, le noyau a_k de B^p minimisant la quantité :

$$\sum_{i \in P_k} d(x_i,a_k)$$

D'après la définition de l'inertie sur l'espace B^p muni de la distance en valeurs absolues (chapitre 1), cette quantité représente exactement l'inertie de l'ensemble $\{x_i, i \in P_k\}$ par rapport au point a_k (on parle plus simplement d'inertie de la classe P_k par rapport au point a_k). Cette inertie est minimale pour a_k centre médian de l'ensemble $\{x_i, i \in P_k\}$ (nous parlons alors de centre médian de la classe P_k).

Dans le chapitre 1, nous avons démontré qu'un nuage de points de B^p admet toujours un centre médian appartenant au même espace B^p. Ce résulat s'applique ici aux noyaux des classes de la partition. Les composantes d'un centre médian a_k ont aussi une signification particulière : ce sont les médianes binaires des p variables dans la classe P_k. De plus, celles-ci sont facilement déterminées à partir de la règle de la majorité (nous n'avons supposé aucune pondération ici). Un noyau a_k se construit donc de la façon

$$a_k = (a_k^1, a_k^2, \ldots, a_k^p) \in B^p$$

où

a_k^j = médiane binaire de l'ensemble $\{ x_i^j, i \in P_k \}$

$\Leftrightarrow$ a_k^j = valeur 0 ou 1 majoritaire de l'ensemble $\{ x_i^j, i \in P_k \}$

Ces noyaux appartiennent à l'espace $\mathbf{B}^p$ et la contrainte de noyaux binaires est donc implicitement respectée. Une restriction cependant : dans le cas où une variable prend, dans une classe, autant de fois la valeur 1 que la valeur 0, la médiane est toute valeur de l'intervalle [0,1]. On peut alors convenir de choisir arbitrairement l'une des deux valeurs 0 ou 1 pour médiane (nous avons choisi la valeur 0). Ce choix n'a aucune influence sur la valeur effective du critère **W**(P,L).

3.4 EXPRESSION DU CRITERE À LA CONVERGENCE

A la convergence, le noyau étant fonction de la partition, on peut exprimer le critère uniquement par rapport à la partition, de sorte que :

$$\mathbf{W}(P) = \mathbf{W}(P,\mathbf{g}(P))$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K} \sum_{j \in J} A_k^j$$

où A_k^j est le nombre d'éléments minoritaires dans la classe P_k pour la variable j.

Le tableau initial est résumé par un ensemble de K noyaux binaires. Le critère **W**(P) représente donc le nombre de fois où la situation obtenue s'écarte de la situation "idéale", celle-ci correspondant au cas où les individus sont identiques aux noyaux des classes auxquelles ils appartiennent.

J.C. Gower (1974) identifie les noyaux à des prédicteurs de classes et définit le critère **W**(P) comme le nombre de prédictions incorrectes. Si on note **B**(P) le nombre de prédictions correctes, on a évidemment :

$$n \times p = \mathbf{W}(P) + \mathbf{B}(P)$$

et, comme le montre cette relation, la méthode revient à maximiser le nombre de prédictions correctes.

Dans le chapitre 1, nous avons défini les notions d'écart moyen et d'inertie associées à la distance en valeurs absolues sur l'espace $\mathbf{B}^p$.

Nous pouvons alors définir l'écart de la variable j dans la classe P_k par :

$$\mathbf{EC}(k,j) = \sum_{i \in P_k} |x_i^j - a_k^j| = A_k^j$$

L'expression de l'inertie d'une classe P_k devient alors :

$$\Im(P_k) = \sum_{i \in P_k} d(x_i,a_k) = \sum_{j \in J} \mathbf{EC}(k,j)$$

Nous obtenons finalement une expression de **W**(P) qui rappelle l'expression habituelle de l'inertie intraclasse :

$$\mathbf{W}(P) = \sum_{k=1}^{K} \sum_{j \in J} \mathbf{EC}(k,j) = \sum_{k=1}^{K} \Im(P_k)$$

3.5 AUTRES EXPRESSIONS DU CRITERE ET PROBLEMES ÉQUIVALENTS

Première approche

Il est possible de démontrer une relation liant l'inertie du nuage **N**(I) par rapport à son centre médian noté a et le critère **W**(P,L). Pour cela, décomposons l'expression de l'inertie du nuage comme suit :

$$\Im(N(I)) = \sum_{i \in I} d(x_i, a)$$

$$\Leftrightarrow \quad \Im(N(I)) = \sum_{i \in I} \sum_{j \in J} |x_i^j - a^j|$$

$$\Leftrightarrow \quad \Im(N(I)) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a^j|$$

$$\Leftrightarrow \quad \Im(N(I)) = \sum_{k=1}^{K} \sum_{j \in J} \mathbf{EC}_a^j(k,j)$$

Dans le chapitre 1, nous avons établi une relation liant l'écart moyen d'une variable à l'écart par rapport à un point quelconque de l'intervalle [0,1]. Cela se traduit ici par :

$$\forall\, k=1,\ldots,K,\ \forall\, j \in J \quad \mathbf{EC}_a^j(k,j) = \mathbf{EC}(k,j) + |n_k^j(1) - n_k^j(0)||a^j - a_k^j|$$

où :

- $n_k^j(1) = \sum_{i \in P_k} x_i^j$ le nombre de fois où j prend la valeur 1 dans P_k,
- $n_k^j(0) = \sum_{i \in P_k} (1-x_i^j)$ le nombre de fois où j prend la valeurs 0 dans P_k.

En reportant cette égalité dans l'expression de l'inertie du nuage **N**(I), nous obtenons :

$$\Im(N(I)) = \sum_{k=1}^{K} \sum_{j \in J} \mathbf{EC}(k,j) + \sum_{k=1}^{K} \sum_{j \in J} |a^j - a_k^j||n_k^j(1) - n_k^j(0)|$$

$$\Leftrightarrow \quad \Im(N(I)) = \mathbf{W}(P,L) + \sum_{k=1}^{K} \sum_{j=1}^{p} |a^j - a_k^j||n_k^j(1) - n_k^j(0)|$$

En notant **T**(P,L) le second terme de cette expression, on obtient :

$$\Im(N(I)) = \mathbf{W}(P,L) + \mathbf{T}(P,L)$$

C'est une relation proche de la relation habituelle de décomposition de l'inertie. Elle n'en a cependant pas les propriétés puisque le terme **T**(P,L) n'est pas une inertie. De plus, le point a n'est pas le centre médian de l'ensemble formé des K noyaux des classes. Le problème de la minimisation du critère **W**(P,L) est équivalent au problème de la maximisation de **T**(P,L). Il reste à préciser la signification de ce nouveau critère.

Notons que seules les valeurs médianes locales (celles des variables dans les classes) différentes des valeurs médianes globales (calculées sur l'ensemble I des individus) ont une contribution non nulle au critère **T**(P,L). Pour une variable j et une classe k, cette contribution est égale à la différence entre le nombre de valeurs 0 et le nombre de valeurs 1 prises par j dans la classe P_k. Cette différence est une mesure de la présence de la valeur majoritaire de la variable dans la classe. Dans ces conditions, maximiser

T(P,L) revient à rechercher une partition en K classes regroupant le mieux possible les valeurs minoritaires du tableau initial (c'est à dire les valeurs différentes des médianes globales correspondantes).

Seconde approche
Il est également possible de décomposer le critère **W**(P,L) de la façon suivante :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} x_i^j(1-a_k^j) + (1-x_i^j)a_k^j$$

$$\Leftrightarrow \quad \mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} x_i^j - \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} x_i^j a_k^j + \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (1-x_i^j)a_k^j$$

$$\Leftrightarrow \quad \sum_{i \in I} \sum_{j \in J} x_i^j = \mathbf{W}(P,L) + \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} x_i^j a_k^j - \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (1-x_i^j)a_k^j$$

En posant :

$$N_1 = \sum_{i \in I} \sum_{j \in J} x_i^j$$

le nombre de valeurs 1 du tableau initial **X**(I,J)

$$E_1(1) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} x_i^j a_k^j$$

le nombre de valeurs 1 communes au tableau initial
et au tableau correspondant à la situation "idéale"

$$E_1(0) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (1-x_i^j)a_k^j$$

le nombre de valeurs 0 du tableau initial
représentées par la valeur 1 dans le tableau "idéal"

nous obtenons :

$$N_1 = \mathbf{W}(P,L) + (E_1(1) - E_1(0))$$

Nous avons déja montré que le problème de minimisation de **W**(P,L) revient à maximiser le nombre de prédictions 0 ou 1 correctes. Nous avons également vu que cela revient à rechercher une partition regroupant le mieux possible les valeurs minoritaires initiales. D'après la relation ci-dessus, ces problèmes sont équivalents au problème d'optimisation de la différence entre le nombre de prédictions 1 correctes et le nombre de prédictions 1 incorrectes. Nous obtenons un résultat analogue en considérant les valeurs 0 du tableau initial et les composantes nulles des noyaux.

3.6 INDICES D'AIDE À L'INTERPRÉTATION

Nous définissons ici des indices permettant d'une part de juger de la qualité de la partition, d'autre part de déterminer les variables caractérisant le mieux les classes obtenues. Il s'agit de fournir à l'utilisateur des indices facilement interprétables. La valeur du critère à la convergence constitue un premier indicateur : il représente le nombre de fois où la situation obtenue s'écarte de la situation "idéale". Cependant, celui-ci doit être observé en regard de la taille du tableau.

Nous définissons un indice plus explicite, égal au pourcentage de données initiales identiques aux noyaux correspondants, par :

$$100x \frac{np - W(P)}{np}$$

$$\text{où} \quad W(P) = \sum_{k=1}^{K} \sum_{j \in J} A_k^j$$

et A_k^j = le nombre d'éléments minoritaires dans la classe P_k pour la variable j.

D'autre part, nous proposons un indice permettant d'évaluer l'homogénéité d'une classe. Il représente le pourcentage d'accords entre individus et noyau d'une classe. Pour une classe P_k de cardinal n_k, il est défini par :

$$100x \frac{n_k p - \sum_{j \in J} A_k^j}{n_k p}$$

Les noyaux sont des vecteurs binaires permettant de distinguer rapidement les particularités de chacune des classes de la partition. En complément, nous proposons pour chaque couple (classe, variable) un indice évaluant l'homogénéité de la variable dans la classe. Il s'agit simplement du pourcentage d'individus de la classe ayant choisi la valeur "idéale" fournie par le noyau. Pour une classe P_k et une variable j, l'indice est défini par :

$$100x \frac{n_k - A_k^j}{n_k}$$

3.7 EXEMPLE SIMPLE D'APPLICATION

Soit un ensemble de 10 micro-ordinateurs, identifiés par les nombres de **1** à **10**, caractérisé par un ensemble de 10 propriétés, identifiées par les lettres **a** à **j**. Le tableau binaire croisant ces deux ensembles (figure 1) est construit à partir du codage par les valeurs 0 et 1, un 1 indiquant que la propriété est vérifiée et un 0 qu'elle ne l'est pas.

On applique l'algorithme MNDBIN en demandant 3 classes, on obtient la partition suivante :

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (\{\mathbf{1}, \mathbf{4}, \mathbf{8}\}, \{\mathbf{2}, \mathbf{5}, \mathbf{6}, \mathbf{10}\}, \{\mathbf{3}, \mathbf{7}, \mathbf{9}\})$$

Celle-ci est représentée (figure 2) en réordonnant simplement les lignes de manière à respecter les classes obtenues.

Nous précisons également tous les indices d'aide à l'interprétation et, dans les figures de la page suivante, on représente :

-figure 3 : les effectifs des classes et le critère d'homogénéité associé à chacune d'elle,

-figure 4 : le tableau des valeurs idéales ou noyaux,

-figure 5 : le tableau d'homogénéité par classe et par variable.

La valeur du critère étant de 15, cela montre que sur 100 données initiales, 15 sont différentes de la valeur idéale correspondante : 85% des données initiales sont donc correctement représentées par les noyaux.

	a	b	c	d	e	f	g	h	i	j
1	1	0	1	0	1	0	0	1	0	1
2	0	1	0	1	0	1	1	0	1	0
3	1	0	0	0	0	0	0	1	1	0
4	1	0	1	0	0	0	0	1	0	0
5	0	1	0	1	1	1	1	0	1	0
6	0	1	0	0	1	1	1	0	1	0
7	0	1	0	0	0	0	0	1	0	1
8	1	0	1	0	1	1	0	1	1	1
9	1	0	0	1	0	0	0	0	0	1
10	0	1	0	1	0	0	1	0	0	0

figure 1
tableau initial

	a	b	c	d	e	f	g	h	i	j
1	1	0	1	0	1	0	0	1	0	1
4	1	0	1	0	0	0	0	1	0	0
8	1	0	1	0	1	1	0	1	1	1
2	0	1	0	1	0	1	1	0	1	0
5	0	1	0	1	1	1	1	0	1	0
6	0	1	0	0	1	1	1	0	1	0
10	0	1	0	1	0	0	1	0	0	0
3	1	0	0	0	0	0	0	1	1	0
7	0	1	0	0	0	0	0	1	0	1
9	1	0	0	1	0	0	0	0	0	1

figure 2
tableau réordonné

	effectifs	**homogénéité**
A	3	87
B	4	88
C	3	80

figure 3
descriptif des classes

	a	b	c	d	e	f	g	h	i	j
A	1	0	1	0	1	0	0	1	0	1
B	0	1	0	1	0	1	1	0	1	0
C	1	0	0	0	0	0	0	1	0	1

figure 4
les noyaux

	a	b	c	d	e	f	g	h	i	j
A	100	100	100	100	67	67	100	100	67	67
B	100	100	100	75	50	75	100	100	75	100
C	67	67	100	67	100	100	100	67	67	67

figure 5
homogénéité par classe et par variable

3.8 REMARQUES

La méthode MNDBIN résume les données initiales par K vecteurs binaires facilement interprétables. La qualité du résultat, qui est fournie par la valeur du critère à la convergence représente simplement le nombre de différences entre données initiales et

valeurs idéales fournies par les noyaux. Les indices d'aides à l'interprétation, que nous avons décrits précédemment, permettent d'effectuer assez rapidement une analyse plus fine des résultats. Enfin, comme nous l'avons démontré dans le chapitre précédent, tous ces résultats sont indépendants du codage binaire retenu.

Les inconvénients sont ceux de toute méthode de type nuées dynamiques, à savoir le problème du choix des éléments de départ et du nombre de classes.

3.9 CLASSIFICATION DE DONNÉES BINAIRES ET MODELE

Les liens existant entre les méthodes de classification automatique et les modèles de statistique inférentielle ont surtout été étudiés lorsque les données sont quantitatives. Le critère d'inertie intraclasse est alors associé à un mélange gaussien (A. Scott et M. Symons 1971, A. Schroeder 1976, G. Celeux 1988).

Lorsque les données sont binaires, G. Govaert (1988) montre comment l'identification d'un mélange de distributions de Bernoulli, avec le même paramètre pour toutes les variables et toutes les classes, correspond au critère de classification utilisé par l'algorithme MNDBIN. Ce lien permet alors de justifier, a posteriori, l'utilisation pour les données binaires, d'une part de la distance L_1, d'autre part de noyaux binaires. En outre, comme il sera illustré dans le paragraphe consacré à l'application de la méthode MNDBIN, ce lien permet d'expliquer les bons résultats que nous avions obtenus sur des données simulées qui suivaient justement ce modèle.

L'auteur propose également d'étendre ce modèle à des paramètres dépendant à la fois des variables et des classes. De plus, cette généralisation permet de retrouver le modèle des classes latentes dans le cas le plus général traité par G. Celeux (1988). Enfin, l'extension du modèle permet de proposer de nouveaux algorithmes utilisant des distances adaptatives de type L_1.

A. Mkhadri (1989) propose différents algorithmes adaptatifs (différents systèmes de pondérations pour les variables sont proposés) et montre, sous certaines hypothèses, ls lien existant entre l'approche géométrique qu'il propose et l'approche probabiliste exposée précédemment.

4. EXTENSION À LA FAMILLE DE DISTANCES DE MINKOWSKI

Nous proposons ici d'utiliser les distances de la famille de Minkowski dans le cadre des méthodes de type Nuées Dynamiques. Lorsque les données sont binaires, M.R. Anderberg (1973) étudie la forme particulière de cette famille de distances.

4.1 LA FAMILLE DE DISTANCES DE MINKOWSKI

Soit **E** un espace de dimension p (par exemple $\mathbf{R}^p$) et d_r la distance de Minkowski de paramètre r. Soient x et y deux points de l'espace **E**, la distance d_r entre ces deux points s'exprime par :

$$d_r(x,y) = \left(\sum_{j=1}^{p} |x^j - y^j|^r \right)^{1/r}$$

Pour la valeur r égale à 1, on retrouve la distance en valeurs absolues, pour la valeur r égale à 2, la distance euclidienne usuelle.

Lorsque l'espace considéré est l'espace $\mathbf{B}^p$, la distance d_r peut s'exprimer en fonction de la distance en valeurs absolues, notée d, de la façon suivante :

$$\forall\, x \text{ et } y \in \mathbf{B}^p \quad d_r(x,y) = \Big(\sum_{j=1}^{p} |x^j - y^j| \Big)^{1/r} = \big(d(x,y)\big)^{1/r}$$

4.2 LES PROBLEMES

Soit l'un ensemble I={1,2,...,n} de n individus décrit par l'ensemble J={1,2,...,p} de p variables binaires. Chaque individu i est représenté, dans l'espace $\mathbf{B}^p$, par un point x_i. L'ensemble des individus est représenté par le nuage N(I).

Le problème est toujours de déterminer une partition de I en K classes, K étant fixé a priori.

Ici, nous munissons l'espace $\mathbf{B}^p$ de la distance d_r. Nous envisageons encore d'utiliser des méthodes du type Nuées Dynamiques, permettant d'imposer la contrainte de noyaux binaires. Deux études sont réalisées ici, dépendantes de l'expression retenue pour le critère à optimiser.

4.3 PREMIERE ÉTUDE

La première famille de problèmes est la suivante :

> *trouver le couple (P,L), où $P=(P_1,P_2,\ldots,P_K)$ est une partition de l'ensemble I et $L=(a_1,a_2,\ldots,a_K)$ l'ensemble des noyaux binaires, tel que le critère :*
>
> $$W(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} d_r^r(x_i,a_k)$$
>
> *soit minumum.*

Notons que, lorsque r prend la valeur 2, on retrouve la notion d'inertie.

Le critère W(P,L) ne dépend pas de la distance d_r choisie. On a en effet :

$$W(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|^r = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k)$$

où d est la distance en valeurs absolues sur $\mathbf{B}^p$. Nous retrouvons le critère optimisé par l'algorithme MNDBIN. Celui-ci fournit donc une solution au problème posé.

Suppression de la contrainte
Si on supprime la contrainte de noyaux binaires, la méthode MNDBIN ne convient plus dès que r est supérieur ou égal à 2.

Quand r prend la valeur 2, la distance considérée est la distance euclidienne usuelle. La solution est alors de prendre, pour noyaux, les centres de gravité des classes. La méthode correspondante est la méthode des centres mobiles.

Enfin, pour les valeurs de r supérieures ou égales à 3, les noyaux deviennent difficilement calculables.

4.4 SECONDE ÉTUDE

La seconde famille de problèmes est la suivante :

> *trouver le couple (P,L), où $P=(P_1,P_2,\ldots,P_K)$ est une partition de l'ensemble I et $L=(a_1,a_2,\ldots,a_K)$ l'ensemble des noyaux binaires, tel que le critère :*
>
> $$W_2(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} d_r(x_i,a_k)$$
>
> *soit minimum.*

Nous nous intéressons, tout d'abord, aux différentes fonctions d'affectation (f) et de représentation (g) associées aux différentes valeurs du paramètre r. Lorsque l'espace de définition est $\mathbf{B}^p$, les distances d_r induisent toutes un même préordre entre vecteurs binaires (représentant les éléments à classifier) et noyaux binaires : pour toute valeur de r, les fonctions d'affectations correspondantes sont équivalentes.

La fonction de représentation (g) va, elle, dépendre du choix du paramètre r. On a alors :

- pour r égal à 1, il s'agit de l'algorithme MNDBIN (les noyaux sont les centres médians des classes).
- pour r supérieur ou égal à 2, le noyau d'une classe n'est pas toujours son centre médian appartenant à $\mathbf{B}^p$ et devient alors difficilement calculable.

Suppression de la contrainte
Pour des noyaux quelconques, la méthode MNDBIN convient pour r égal à 1 (elle fournit encore des noyaux binaires).

Pour r égal à 2, les noyaux ne peuvent être déterminés de façon analytique et nécessitent la mise en place d'un algorithme de recherche. Contrairement au centre médian, le noyau d'une classe est ici unique. J.C. Gower (1974b), F.K. Bedall et H. Zimmermann (1979) ont construit des algorithmes de recherche de ce type de noyau souvent appelé "spatial median" ou médiane spatiale (B.M. Brown 1983, R. Kosfeld 1986). R. Kosfeld (1986) propose également une méthode qu'il compare aux deux algorithmes précédents. Il traite également de l'utilisation du centre médian et de la médiane spatiale en tant que noyau ou caractéristique de valeur centrale d'un nuage de points de $\mathbf{R}^p$.

Pour des valeurs de r supérieures ou égales à 3, les noyaux des classes deviennent difficilement calculables.

5. ETUDE COMPARATIVE

Nous avons mis en évidence deux méthodes de classification de type Nuées Dynamiques. Elles se différentient par le choix de l'espace et de la distance mais également par la nature des noyaux représentatifs des classes.

La méthode MNDBIN utilise la distance en valeurs absolues sur l'espace $\mathbf{B}^p$. Les noyaux fournis appartiennent à $\mathbf{B}^p$. La méthode MNDQAN (G. Celeux et al. 1989) utilise, elle, la distance euclidienne usuelle. Dans ce cas, les données binaires sont plongées dans l'espace $\mathbf{R}^p$. L'ensemble à classifier et les noyaux fournis par la méthode (les centres de gravité des classes) appartiennent à cet espace.

Nous comparons ici les critères associés à ces deux méthodes. Un exemple simple d'application est ensuite proposé.

5.1 ETUDE COMPARATIVE DES CRITERES

Soit P une partition en K classes de l'ensemble des individus.

Le critère associé à la méthode MNDQAN s'exprime par :

$$\mathbf{W_g}(P,L_g) = \sum_{k=1}^{K} \sum_{i \in P_k} d_e^2(x_i,g_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (x_i^j - g_k^j)^2$$

où

d_e représente la distance euclidienne,

$L_g = (g_1, g_2, \ldots, g_K)$ est l'ensemble les centres de gravité des classes.

Le centre de gravité g_k d'une classe P_k de cardinal n_k est défini par :

$$\forall\, j \in J \quad g_k^j = \frac{1}{n_k} \sum_{i \in P_k} x_i^j$$

La composante j du point g_k est égale à la part de valeurs 1 prises par la variable j dans la classe P_k. Remarquons que si cette part est supérieure à 1/2, la médiane est alors égale à 1, sinon elle est égale à 0.

Le critère associé à la méthode MNDBIN s'exprime par :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$

où

$L = (a_1, a_2, \ldots, a_K)$ sont les centres médians des classes,

$\forall\, j \in J \quad a_k^j =$ médiane $\{ x_i^j, i \in P_k \}$.

Plaçons nous dans l'espace $\mathbf{R}^p$ muni de la distance euclidienne d_e. Nous disposons alors de la relation de Huyghens. Pour toute valeur de k, celle-ci nous permet de décomposer l'inertie de la classe par rapport au centre médian a_k de la façon suivante :

$$\sum_{i \in P_k} d_e^2(x_i,a_k) = \sum_{i \in P_k} d_e^2(x_i,g_k) + n_k d_e^2(g_k,a_k)$$

$$\Leftrightarrow \quad \sum_{i \in P_k} \sum_{j \in J} (x_i^j - a_k^j)^2 = \sum_{i \in P_k} \sum_{j \in J} (x_i^j - g_k^j)^2 + n_k\, d_e^2(g_k,a_k)$$

$$\Leftrightarrow \quad \sum_{i \in P_k} \sum_{j \in J} | x_i^j - a_k^j | = \sum_{i \in P_k} \sum_{j \in J} (x_i^j - g_k^j)^2 + n_k\, d_e^2(g_k,a_k)$$

d'où la relation :

$$\mathbf{W}(P,L) = \mathbf{W_g}(P,L_g) + \sum_{k=1}^{K} n_k\, d_e^2(g_k,a_k)$$

Cette relation montre la différence entre le problème d'optimisation de $\mathbf{W}(P,L)$ et celui de l'optimisation de $\mathbf{W_g}(P,L_g)$.

5.2 EXEMPLE COMPARATIF

Considérons un ensemble de 10 individus, identifiés par les nombres **1** à **10**. Ceux-ci sont décrits par 3 variables binaires, identifiées par les lettres **a**, **b** et **c**. Les données sont représentées en figure 1 sous la forme d'un tableau binaire (le codage est toujours réalisé par les valeurs 0 et 1).

Nous appliquons les méthodes MNDBIN et MNDQAN en demandant 2 classes et en partant d'une même partition initiale, par exemple ({**1**,**2**,**3**,**4**,**5**},{**6**,**7**,**8**,**9**,**10**}).

Les centres médians des classes initiales sont a_1=(0,1,0) et a_2=(1,0,1), les centres de gravités g_1=(1/5,3/5,2/5) et g_2=(1,2/5,3/5). Lors de la première étape d'affectation, l'individu numero **1** est affecté à la première classe pour la méthode MNDBIN, à la seconde pour la méthode MNDQAN). Cela montre l'influence de la variable **a**, ayant ici la plus forte proportion de valeurs 1. Finalement, les partitions obtenues à la convergence sont :

(**A**, **B**) = ({**1**, **3**, **5**, **6**, **8**}, {**2**, **4**, **7**, **9**, **10**}) pour l'algorithme MNDBIN,

(**C**, **D**) = ({**2**, **3**, **4**, **5**}, {**1**, **6**, **7**, **8**, **9**, **10**}) pour l'algorithme MNDQAN.

Nous représentons (figures 2 et 3) les deux partitions en réordonnant simplement les lignes de manière à respecter les classes obtenues. Enfin, les noyaux correspondant à ces partitions sont indiqués dans les figures 4 et 5.

	a	**b**	**c**
1	1	1	0
2	0	0	1
3	0	1	0
4	0	0	1
5	0	1	0
6	1	1	0
7	1	0	1
8	1	1	0
9	1	0	1
10	1	0	1

figure 1
tableau initial

	a	**b**	**c**
1	1	1	0
3	0	1	0
5	0	1	0
6	1	1	0
8	1	1	0
2	0	0	1
4	0	0	1
7	1	0	1
9	1	0	1
10	1	0	1

figure 2
après MNDBIN

	a	**b**	**c**
2	0	0	1
3	0	1	0
4	0	0	1
5	0	1	0
1	1	1	0
6	1	1	0
7	1	0	1
8	1	1	0
9	1	0	1
10	1	0	1

figure 3
après MNDQAN

	a	**b**	**c**
A	1	1	0
B	1	0	1

figure 4
centres médians

	a	**b**	**c**
C	0	$\frac{1}{2}$	$\frac{1}{2}$
D	1	$\frac{1}{2}$	$\frac{1}{2}$

figure 5
centres de gravité

Notons qu'après plusieurs essais effectués sur ce tableau élémentaire (à partir de points initiaux différents), les meilleures solutions obtenues par les deux méthodes sont identiques et correspondent à celle représentée par la figure 2. Sur des exemples plus conséquents (paragraphe suivant), les méthodes fournissent des partitions différentes.

6. PROGRAMME ET APPLICATIONS

6.1 PRÉSENTATION DU PROGRAMME

L'un des buts du logiciel SICLA est de diffuser de nouvelles méthodes d'analyse de données. Il contient une bibliothèque de procédures destinée à faciliter la programmation de ces nouvelles méthodes.

C'est dans ce cadre que le programme MNDBIN a été écrit, devenant ainsi une commande du logiciel SICLA. Lors de l'exécution, l'utilisateur doit répondre à un certain nombre de questions permettant de d'initialiser les paramètres de la méthode. Après avoir choisi la structure de données à analyser, l'utilisateur peut :

- selectionner tout ou partie de l'ensemble des variables et des individus,
- choisir le point de départ initial,
- définir le nombre d'essais à effectuer,
- choisir le nombre de classes de la partition,
- chosir le type de sortie des résulats.

6.2 APPLICATIONS

Il s'agit d'étudier un ensemble de plaques-boucles de ceintures démasquinées du Nord-Est de la France, s'échelonnant entre la fin du VI^{ème} et le début du VIII^{ème} siècle (H. Leredde et P. Perin 1980). L'ensemble est consitué de 59 plaques-boucles sur lesquelles ont été observées la présence ou l'absence d'une sélection de 26 critères techniques de fabrication, de forme et de décor. Ces critères sont indiqués dans le listing figurant en page 57. Le problème posé est de structurer ces données pour faire apparaitre des liens entre individus et plaques et, aussi, une évolution dans les techniques de fabrications.

Apppelons MERO ce jeu de données. Plusieurs applications des algorithmes MNDBIN et MNDQAN ont été effectuées. Deux classes sont demandées et seules les meilleures solutions sont retenues. Dans les deux cas, on aboutit à une même partition montrant une nette différence entre deux grands groupes de plaques-boucles. Le listing résultat de l'application MNDQAN figure à la page 59. Nous nous intéressons ici uniquement aux résultats fournis par MNDBIN et illustrés dans les pages 57 (partition obtenue et tableau des valeurs idéales) et 58 (tableau des homogénéités et tableau initial réordonné).

Le critère égal à **224** traduit le fait que **84.5%** des données initiales sont égales aux valeurs idéales fournies par les noyaux. Ces derniers et le tableau des homogénéités font ressortir un ensemble de variables (homogènes et ayant un noyau différent dans les deux classes) fortement discriminantes qui permettent de caractériser deux groupes **A** (classe 1) et **B** (classe 2) de plaques-boucles. Nous énumerons ici les caractères les plus explicites de **A** et **B** :

Groupe A : damasquinure par incrustation dominate (identifiée par **C24**),
damasquinure monochrome argent (**C22**),
fixation par bossette de bronze (**C15**).

Groupe B : damasquinure par placage prédominant (**C25**),
damasquinure bichrome (**C23**),
fixation par clous de fer (**C16**).
fond plaqué d'argent (**C37**),

Nous pousuivons l'étude des données MERO en recherchant une partition en 7 classes. Cette fois les deux méthodes diffèrent quelque peu. Cependant, elles respectent la répartition en deux groupes, en ce sens que toute classe obtenue est incluse soit dans le groupe **A**, soit dans le groupe **B**. Les résultats figurent dans les pages 60, 61 (pour MNDBIN) et 62 (pour MNDQAN).

Le critère associé à MNDBIN égal à **82** indique que **94.6%** des données initiales sont correctement représentées par les noyaux. Nous présentons ci-dessous les trois classes les plus importantes du goupe **A** (les classes 5 et 7 incluses dans **A** ont des effectifs faibles et nous ne les détaillons pas ici) et les deux classes du groupe **B**. Les caractères spécifiques et discriminant les classes d'un même groupe sont les suivants :

Groupe A_1 : classe numero 3, 18 éléments,
bordures de frises géométriques ou de hachures (**C39**),
incrustations avec bande pointillée (**C33**).

Groupe A_2 : classe numero 1, 8 éléments,
plaque de forme ronde (**C01**),
ardillon à plateau rond (**C19**),
trame géométrique exclusive (**C28**).

Groupe A_3 : classe numero 6, 7 éléments,
entrelacs de rubans (**C30**),
incrustations avec ruban plein (**C35**).

Groupe B_1 : classe numéro 2, 10 éléments,
entrelacs animaliers (**C31**),
motifs animaliers (**C41**).

Groupe B_2 : classe numéro 4, 10 éléments,
arabesques (**C32**),
incrustations filiformes (**C36**).

La partition obtenue par MNDQAN permet d'expliquer **79.5%** de l'inertie initiale. On retrouve, à une permutation près, les groupes B_1 (classe 3), B_2 (classe 1), A_2 (classe 2) et A_3 (classe 5). Par contre, le groupe A_1 est pratiquement décomposé en deux classes (numéros 4 et 6), celles-ci étant parfaitement discréminées par la variable **C17** : "plaque dorsale carrée" (d'après les centres de gravités de ces deux classes indiqués en page 62). Remarquons que **61%** des individus du groupe A_1 possédait ce caractère. La méthode MNDQAN a donc divisé ce groupe suivant la dichotomie induite par la variable **C17**.

Globalement, nous aboutissons à des résultats très voisins de ceux obtenus par H. Leredde et P. Perin (1980). Pour nos applications, des interprétations analogues à celles formulées par les auteurs sont possibles. Citons-les : "les applications mettent en évidence une évolution technique, s'acconpagnant d'une évolution chronologique tout à fait nette. En effet, les techniques caractérisant le groupe A_2 apparaissent à la fin du VI^{ème} siécle et jusque vers 600. Celles-ci évoluent entre 600 et 650, comme le montrent les caractéristiques des plaques-boucles inclusent dans les autres classes du groupe **A**. Enfin, les techniques utilisisées pour les plaques-boucles du groupe **B** datent de la deuxième moitié du VII^{ème} et se prolongent jusqu'au début du VIII^{ème} siècle".

Remarque
Si on applique MNDBIN en demandant 5 classes, on retrouve le groupe A_1 qui se voit rajouter 3 éléments de la classe 5 (de la partition en 7 classes), le groupe A_2 un élément de cette même classe et le groupe A_3 la classe 7. Les classes B_1 et B_2 se retrouve à une permutation près qui n'influe pas sur le critère égal à **100** de la partition. Pour MNDQAN, on obtient à une permutation près les mêmes groupes que MNDBIN. Ces partitions seront utilisées dans le chapitre 5 où on traite aussi de la classification ascendante hiérarchique pour données binaires.

```
COMMANDE: MNDBIN <> nuees dynamiques sur variables binaires


      variables selectionnees :
 C01  plaque de forme ronde
 C14  plaque de grande taille
 C15  fixation par bossettes de bronze
 C16  fixation par clous de fer
 C17  plaque dorsale carree
 C19  ardillon a plateau rond
 C22  damasquinure monochrome argent
 C23  damasquinure bichrome
 C24  damasquinure par inscrustation dominante
 C25  damasquinure par placage predominant
 C26  decor sur toute la surface de la plaque
 C28  trame geometrique exclusive
 C29  tresse
 C30  entrelacs de rubans
 C31  entrelacs animaliers
 C32  arabesques
 C33  incrustations avec bande pointillee
 C34  incrustations avec bande hachure ou nid
 C35  incrustations avec ruban plein
 C36  incrustations filiformes
 C37  fond plaque d'argent
 C38  fond hachure ou a trame geometrique
 C39  bordures de frises geometriques ou de hachures
 C40  bordures offrant la repetition du motif
 C41  motifs animaliers
 C42  plaques a bords mouvementes


      valeur du critere obtenu  :  224


      partition obtenue :

      classe  1 :    39 elements

      03   04   05   07   09   11   14   15   17   20   21   22   23   24   25
      26   27   28   30   31   32   33   34   35   36   46   47   48   49   50
      51   52   53   54   55   56   57   58   59

      classe  2 :    20 elements

      01   02   06   08   10   12   13   16   18   19   29   37   38   39   40
      41   42   43   44   45


     tableau des valeurs ideales

     CCCCCCCCCCCCCCCCCCCCCCCCC
     01111122222223333333333444
     14567923456890123456789012

  1   1   1 1             11
  2     1   1 1         11    1
```

```
COMMANDE: MNDBIN <> nuees dynamiques sur variables binaires
```

homogeneite par classe

	C 0 1	C 1 4	C 1 5	C 1 6	C 1 7	C 1 9	C 2 2	C 2 3	C 2 4	C 2 5	C 2 6	C 2 8	C 2 9	C 3 0	C 3 1	C 3 2
1	79	92	89	89	61	82	97	100	100	100	79	76	51	76	97	100
2	100	70	100	95	80	100	100	100	100	100	70	100	95	100	55	50

	C 3 3	C 3 4	C 3 5	C 3 6	C 3 7	C 3 8	C 3 9	C 4 0	C 4 1	C 4 2
1	53	87	82	97	97	66	58	100	92	87
2	100	80	100	90	100	100	90	60	55	75

tableau initial reordonne

```
   ccccccccccccccccccccccccc
   01111122222223333333333444
   14567923456890123456789012

03   1   1 1   1   1     1
04    11 1 1   1   1    11
05   1   1 1  1         1
07   1 1 1 1 1  1   1   1  11
09   1 1 1 1   1   1    11 1
11   1 1 1 1 1  1    1  1
14   1   1 1       1    11
15 1 1   1 1  1
17   1   1 1 1  1    1  1
20    1  1 1   1      1  1 1
21  1 1  1 1   1   1    11
22    11 1 1   1   1    11
23 1 1   1 1  1
24   1 1 1 1   1   1    11
25   1 1 1 1   1   1    11
26  11   1 1   1   1    11
27   1   1 1 1  1    1  1
28  11   1 1   1  11   11
30   1   1 1   1   1    11
31   1   1 1   1   1    11  1
32   1 1 1 1   1   1    11
33   1 1 1 1   1   1    11
34   1 1 1 1    1   1  11
35   1   1 1   1    1  1 1
36   1   1 1   1   1    11
46   1 1 1 1   1   1    11
47 1 1  11 1  1
48   1   1 1 1  1    1      1
49   1 1 1 1   1   1    11
50   1 1 1 1 1   1  1   1  11
51   1     1 1  1    1  1
52   1 1 1 1    1   1   11
53   1 111 1   1   1    11
54 1 1  11 1  1
55 1 1  11 1  1
56 1 1  11 1  1
57 1 1  11 1  1
58 1 1  11 1  1
59   1   1 1 1  1    1  1
-----------------------------
01  1 1   1 1     1   11  1 1
02    11  1 1    1  1 11  111
06    11  1 1    1    11  111
08        1 1    1  1  1   11
10  1 1   1 1     1   11  1
12    1   1 1    1    11  111
13    11  1 1    1    11  11
16  1 1   1 1     1   11  1 1
18    1   1 1    1    11  111
19    1   1 1    1  1 11 1 1
29    11  1 1    1  1  1 1 1
37  1 1   1 11    1   11
38  1 1   1 1     1   11    1
39  1 1   1 11    1   11    1
40    1   1 11    1   11    1
41    1   1 11    1   11    1
42    1   1 11    1   11    1
43    1   1 11    1   11    1
44    1   1 1    1    11   11
45    1   1 1  1      11    1
```

COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables quantitatives

pourcentage d inertie expliquee en 2 classes : 45.46

partition obtenue :

classe numero 1 (effectif= 39)

03	04	05	07	09	11	14	15	17	20	21	22	23	24	25
26	27	28	30	31	32	33	34	35	36	46	47	48	49	50
51	52	53	54	55	56	57	58	59						

classe numero 2 (effectif= 20)

01	02	06	08	10	12	13	16	18	19	29	37	38	39	40
41	42	43	44	45										

coordonnees des centres de gravite des classes de la partition :

variable (effectif)	population (59)	classe 1 (39)	classe 2 (20)
C01	.136	.205	.000
C14	.153	.769E-01	.300
C15	.593	.897	.000
C16	.390	.103	.950
C17	.322	.385	.200
C19	.119	.179	.000
C22	.644	.974	.000
C23	.339	.000	1.00
C24	.661	1.00	.000
C25	.339	.000	1.00
C26	.237	.205	.300
C28	.153	.231	.000
C29	.339	.487	.500E-01
C30	.153	.231	.000
C31	.169	.256E-01	.450
C32	.169	.000	.500
C33	.305	.462	.000
C34	.153	.128	.200
C35	.119	.179	.000
C36	.322	.256E-01	.900
C37	.356	.256E-01	1.00
C38	.441	.667	.000
C39	.424	.590	.100
C40	.136	.000	.400
C41	.203	.769E-01	.450
C42	.339	.128	.750

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires

      valeur du critere obtenu  :  82

      partition obtenue :

      classe  1 :     8 elements
      15   23   47   54   55   56   57   58

      classe  2 :    10 elements
      02   06   08   12   13   18   19   29   44   45

      classe  3 :    18 elements
      04   09   14   21   22   24   25   26   28   30   31   32   33   36   46
      49   52   53

      classe  4 :    10 elements
      01   10   16   37   38   39   40   41   42   43

      classe  5 :     4 elements
      03   05   20   35

      classe  6 :     7 elements
      11   17   27   34   48   51   59

      classe  7 :     2 elements
      07   50


   tableau des valeurs ideales

   CCCCCCCCCCCCCCCCCCCCCCCCC
   01111122222223333333333444
   14567923456890123456789012

1  1 1  11 1  1
2     1   1 1    1    11   11
3   1 1 1 1   1   1    11
4  1 1   1 11    1   11    1
5   1   1 1   1         1
6   1   1 1 1  1    1  1
7   1 1 1 1 1       1  1  11


      homogeneite par classe

        C    C    C    C    C    C    C    C    C    C    C    C    C    C    C    C    C
        0    1    1    1    1    1    2    2    2    2    2    2    2    3    3    3    3
        1    4    5    6    7    9    2    3    4    5    6    8    9    0    1    2

1     100  100  100  100  100   75  100  100  100  100  100  100  100  100  100  100
2     100  100  100   90   60  100  100  100  100  100  100  100   90  100   90  100
3     100   83   83   83   61   94  100  100  100  100  100  100   88   94  100  100
4     100   60  100  100  100  100  100  100  100  100   60  100  100  100  100  100
5     100  100   75   75  100  100  100  100  100  100  100   75   75  100  100  100
6     100  100  100  100   71  100   85  100  100  100   85  100  100  100  100  100
7     100  100  100  100  100  100  100  100  100  100  100  100  100   50   50  100

        C    C    C    C    C    C    C    C    C    C
        3    3    3    3    3    3    3    4    4    4
        3    4    5    6    7    8    9    0    1    2

1     100  100  100  100  100  100  100  100  100  100
2     100   60  100   80  100  100   80   50   90   70
3      94   88  100  100  100  100  100  100   94   94
4     100  100  100  100  100  100  100   70  100   80
5      75   75  100   75   75  100  100  100  100   75
6     100  100  100  100  100   85   85  100  100   85
7     100  100  100  100  100  100  100  100  100  100
```

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires


    tableau initial reordonne

    CCCCCCCCCCCCCCCCCCCCCCCCCC
    01111122222223333333333444
    14567923456890123456789012
 ------------------------------
 15 1 1    1 1   1
 23 1 1    1 1   1
 47 1 1   11 1   1
 54 1 1   11 1   1
 55 1 1   11 1   1
 56 1 1   11 1   1
 57 1 1   11 1   1
 58 1 1   11 1   1
 ------------------------------
 02     11   1 1      1  1 11  111
 06     11   1 1      1     11  111
 08          1 1      1  1   1   11
 12     1    1 1      1     11  111
 13     11   1 1      1     11  11
 18     1    1 1      1     11  111
 19     1    1 1      1  1 11 1 1
 29     11   1 1      1  1  1 1 1
 44     1    1 1      1     11   11
 45     1    1 1   1        11    1
 ------------------------------------
 04     11 1 1    1    1     11
 09    1 1 1 1    1    1     11 1
 14    1    1 1         1     11
 21   1 1   1 1    1    1     11
 22     11 1 1    1    1     11
 24    1 1 1 1    1    1     11
 25    1 1 1 1    1    1     11
 26   11    1 1    1    1     11
 28   11    1 1    1   11     11
 30    1    1 1    1    1     11
 31    1    1 1    1    1     11  1
 32    1 1 1 1    1    1     11
 33    1 1 1 1    1    1     11
 36    1    1 1    1    1     11
 46    1 1 1 1    1    1     11
 49    1 1 1 1    1    1     11
 52    1 1 1 1      1    1   11
 53    1 111 1    1    1     11
 ------------------------------
 01   1 1    1 1       1    11   1 1
 10   1 1    1 1       1    11   1
 16   1 1    1 1       1    11   1 1
 37   1 1    1 11      1    11
 38   1 1    1 1       1    11     1
 39   1 1    1 11      1    11     1
 40     1    1 11      1    11     1
 41     1    1 11      1    11     1
 42     1    1 11      1    11     1
 43     1    1 11      1    11     1
 ------------------------------
 03    1    1 1    1    1       1
 05    1    1 1   1             1
 20     1   1 1    1        1   1   1
 35    1    1 1    1     1  1 1
 ------------------------------
 11    1 1 1 1 1  1     1   1
 17    1    1 1 1  1     1   1
 27    1    1 1 1  1     1   1
 34    1 1 1 1     1     1   11
 48    1    1 1 1  1     1       1
 51    1      1 1  1     1   1
 59    1    1 1 1  1     1   1
 ------------------------------
 07    1 1 1 1 1  1    1    1   11
 50    1 1 1 1 1    1  1    1   11
 ------------------------------
```

COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables quantitatives

pourcentage d inertie expliquee en 7 classes : 79.53

partition obtenue :

classe numero 1 (effectif= 11)

01 10 16 37 38 39 40 41 42 43 45

classe numero 2 (effectif= 9)

05 15 23 47 54 55 56 57 58

classe numero 3 (effectif= 9)

02 06 08 12 13 18 19 29 44

classe numero 4 (effectif= 10)

04 09 22 24 25 32 33 46 49 53

classe numero 5 (effectif= 6)

11 17 27 48 51 59

classe numero 6 (effectif= 10)

03 14 20 21 26 28 30 31 35 36

classe numero 7 (effectif= 4)

07 34 50 52

coordonnees des centres de gravite des classes de la partition :

variable (effectif)	population (59)	classe 1 (11)	classe 2 (9)	classe 3 (9)	classe 4 (10)	classe 5 (6)	classe 6 (10)	classe 7 (4)
C01	.136	.000	.889	.000	.000	.000	.000	.000
C14	.153	.545	.000	.000	.000	.000	.300	.000
C15	.593	.000	1.00	.000	.800	1.00	.800	1.00
C16	.390	1.00	.000	.889	.200	.000	.200	.000
C17	.322	.000	.000	.444	1.00	.167	.000	1.00
C19	.119	.000	.667	.000	.100	.000	.000	.000
C22	.644	.000	1.00	.000	1.00	.833	1.00	1.00
C23	.339	1.00	.000	1.00	.000	.000	.000	.000
C24	.661	.000	1.00	.000	1.00	1.00	1.00	1.00
C25	.339	1.00	.000	1.00	.000	.000	.000	.000
C26	.237	.545	.000	.000	.000	1.00	.000	.500
C28	.153	.000	1.00	.000	.000	.000	.000	.000
C29	.339	.909E-01	.000	.000	1.00	.000	.900	.000
C30	.153	.000	.000	.000	.000	1.00	.000	.750
C31	.169	.000	.000	1.00	.000	.000	.000	.250
C32	.169	.909	.000	.000	.000	.000	.000	.000
C33	.305	.000	.000	.000	1.00	.000	.800	.000
C34	.153	.000	.000	.444	.000	.000	.200	.750
C35	.119	.000	.000	.000	.000	1.00	.000	.250
C36	.322	1.00	.000	.778	.000	.000	.100	.000
C37	.356	1.00	.000	1.00	.000	.000	.100	.000
C38	.441	.000	.000	.000	1.00	.833	.700	1.00
C39	.424	.000	.111	.222	1.00	.000	1.00	.500
C40	.136	.273	.000	.556	.000	.000	.000	.000
C41	.203	.000	.000	1.00	.100	.000	.000	.500
C42	.339	.818	.000	.667	.000	.167	.200	.500

6.3 APPLICATIONS À DES TABLEAUX CONSTRUITS SUIVANT UN MODELE

G. Govaert (1988) a montré comment l'identification d'un mélange de distributions de Bernouilli avec le même paramètre pour toutes les classes et toutes les variables correspond au critère de la méthode MNDBIN.

Nous proposons ici d'appliquer la méthode MNDBIN mais aussi MNDQAN sur des tableaux de données binaires simulées à partir de différentes lois de Bernoulli. Le propramme HASBIN a été construit pour générer de tels tableaux.

Ce programme permet de choisir le nombre d'individus, de variables, de classes, les effectifs de chaque classe ainsi que le paramètre de la loi. Ce dernier correspond alors à la probabilité d'avoir la valeur idéale définie par le noyau.

Un premier essai est effectué à partir de la loi de Bernoulli de paramètre 0.9. Une partition de 7 classes contenant chacune 15 individus décrits par 20 variables binaires est générée. Après des permutations aléatoires (effectuées par HASBIN), on dispose d'un tableau croisant 100 individus (identifiés par les nombres **1** à **105**) et 20 variables binaires (identidiées par **VA1** à **VA20**).

Les noyaux et les classes sont indiqués en page 64. Pour ce tableau, les deux méthodes MNDBIN (page 65) et MNDQAN (page 66) retrouvent exactement la partition simulée.

Un nouvel essai est ensuite effectué pour un paramètre égal à 0.8 (page 67). La partition fournie par MNDBIN (page 68) diffère de 9 individus par rapport à la partition simulée. Les classes (1,6) et (2,4) comptent respectivement une et deux différences avec les classes (5,4) et (6,7) créées par HASBIN.

La partition obtenue par MNDQAN (page 69) compte beaucoup plus de différences, seules les deux classes (2,5) sont voisines des classes simulées (5,4).

```
COMMANDE : HASBIN <> creation de tableaux binaires

        VALEUR DES PARAMETRES

    nombre d'individus                          105
    nombre de variables                          20
    nombre de classes                             7
    parametre de la loi binomiale              0.90

        EFFECTIF DES CLASSES

        classe   1  :  15
        classe   2  :  15
        classe   3  :  15
        classe   4  :  15
        classe   5  :  15
        classe   6  :  15
        classe   7  :  15

        TABLEAU DES VALEURS IDEALES

      v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v
      0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
      1  2  3  4  5  6  7  8  9  1  1  1  1  1  1  1  1  1  1  1  2
                                0  1  2  3  4  5  6  7  8  9  0
1     0  0  1  1  0  0  0  0  1  1  0  1  0  1  0  1  0  1  1  0
2     0  1  0  0  1  0  0  0  1  0  1  0  0  0  1  0  1  1  0  1
3     1  1  1  0  0  0  1  1  1  0  0  0  1  0  0  0  0  1  0  0
4     1  0  0  1  1  1  0  0  0  1  0  0  0  0  0  0  0  1  1  0
5     1  0  0  0  0  1  1  0  0  0  1  0  0  0  1  1  0  0  0  1
6     1  1  0  0  0  0  0  0  0  0  0  1  0  1  0  0  1  1  0  1
7     1  0  1  0  0  0  0  0  1  1  1  0  1  1  0  0  1  0  1  0

        CLASSES DE LA PARTITION SIMULEE

        classe numero :  1
           3    4    5    6    8    9   10   11   13   36   44   49   54   63   75

        classe numero :  2
          14   19   20   23   24   25   27   29   37   40   57   81   90   96  100

        classe numero :  3
           2    7   22   31   32   34   35   38   41   51   52   67   69   86  102
        classe numero :  4
          12   15   30   33   46   47   48   50   55   56   59   61   68   72  101
        classe numero :  5
          16   26   39   42   43   62   64   65   66   70   71   73   74   78  103
        classe numero :  6
           1   17   53   60   76   77   79   80   83   85   87   88   89  104  105
        classe numero :  7
          18   21   28   45   58   82   84   91   92   93   94   95   97   98   99
```

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires


         valeur du critere obtenu  :  214


         partition obtenue :

         classe  1 :    15 elements
         2    7   22   31   32   34   35   38   41   51   52   67   69   86  102

         classe  2 :    15 elements
        12   15   30   33   46   47   48   50   55   56   59   61   68   72  101

        classe  3 :    15 elements
        18   21   28   45   58   82   84   91   92   93   94   95   97   98   99

        classe  4 :    15 elements
        14   19   20   23   24   25   27   29   37   40   57   81   90   96  100

        classe  5 :    15 elements
        16   26   39   42   43   62   64   65   66   70   71   73   74   78  103

        classe  6 :    15 elements
         3    4    5    6    8    9   10   11   13   36   44   49   54   63   75

        classe  7 :    15 elements
         1   17   53   60   76   77   79   80   83   85   87   88   89  104  105



         tableau des valeurs ideales

         VVVVVVVVVVVVVVVVVVV
         AAAAAAAAAAAAAAAAAAA
                  11111111112
         12345678901234567890

       1 111   111   1     1
       2 1  111   1        11
       3 1 1      111 11  1 1
       4  1  1   1 1   1 11 1
       5 1     11    1   11   1
       6    11     11 1 1 1 11
       7 11           1 1  11 1
```

COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables quantitatives

pourcentage d inertie expliquee en 7 classes : 63.92

partition obtenue :

classe numero 1 (effectif= 15)
12 15 30 33 46 47 48 50 55 56 59 61 68 72 101

classe numero 2 (effectif= 15)
1 17 53 60 76 77 79 80 83 85 87 88 89 104 105

classe numero 3 (effectif= 15)
3 4 5 6 8 9 10 11 13 36 44 49 54 63 75

classe numero 4 (effectif= 15)
14 19 20 23 24 25 27 29 37 40 57 81 90 96 100

classe numero 5 (effectif= 15)
2 7 22 31 32 34 35 38 41 51 52 67 69 86 102

classe numero 6 (effectif= 15)
16 26 39 42 43 62 64 65 66 70 71 73 74 78 103

classe numero 7 (effectif= 15)
18 21 28 45 58 82 84 91 92 93 94 95 97 98 99

coordonnees des centres de gravite des variables pour chaque classe de la partition :

variable (effectif)	population (105)	classe 1 (15)	classe 2 (15)	classe 3 (15)	classe 4 (15)	classe 5 (15)	classe 6 (15)	classe 7 (15)
VA 1	.695	1.00	.933	.200	.133	.733	.933	.933
VA 2	.429	.133	.933	.667E-01	.933	.867	.000	.667E-01
VA 3	.429	.133	.667E-01	.867	.200	.667	.133	.933
VA 4	.333	.867	.667E-01	.867	.333	.000	.133	.667E-01
VA 5	.362	.933	.000	.267	1.00	.200	.667E-01	.667E-01
VA 6	.390	.933	.200	.133	.333	.000	1.00	.133
VA 7	.324	.000	.667E-01	.667E-01	.200	.933	.933	.667E-01
VA 8	.267	.200	.200	.667E-01	.333	.933	.667E-01	.667E-01
VA 9	.590	.667E-01	.133	1.00	.867	.867	.267	.933
VA10	.429	.867	.000	.933	.200	.000	.667E-01	.933
VA11	.448	.000	.000	.133	.933	.133	.933	1.00
VA12	.324	.000	1.00	.933	.667E-01	.667E-01	.000	.200
VA13	.352	.267	.667E-01	.133	.667E-01	.933	.667E-01	.933
VA14	.486	.667E-01	.867	.933	.200	.267	.200	.867
VA15	.400	.133	.133	.200	1.00	.133	1.00	.200
VA16	.381	.333	.133	.933	.667E-01	.200	.867	.133
VA17	.400	.000	1.00	.000	.867	.000	.133	.800
VA18	.667	.867	.933	.800	.867	.933	.000	.267
VA19	.467	1.00	.667E-01	1.00	.000	.133	.667E-01	1.00
VA20	.467	.000	1.00	.133	1.00	.200	.867	.667E-01

```
COMMANDE : HASBIN <> creation de tableaux binaires

        VALEUR DES PARAMETRES

    nombre d'individus                          105
    nombre de variables                          20
    nombre de classes                             7
    parametre de la loi binomiale              0.80

        EFFECTIF DES CLASSES

        classe   1  :  15
        classe   2  :  15
        classe   3  :  15
        classe   4  :  15
        classe   5  :  15
        classe   6  :  15
        classe   7  :  15

        TABLEAU DES VALEURS IDEALES

      v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v  v
      0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
      1  2  3  4  5  6  7  8  9  1  1  1  1  1  1  1  1  1  1  1  2
                                0  1  2  3  4  5  6  7  8  9  0
1     0  0  1  1  0  0  0  0  1  1  0  1  0  1  0  1  0  1  1  0
2     0  1  0  0  1  0  0  0  1  0  1  0  0  0  1  0  1  1  0  1
3     1  1  1  0  0  0  1  1  1  0  0  0  1  0  0  0  0  1  0  0
4     1  0  0  1  1  1  0  0  0  1  0  0  0  0  0  0  0  1  1  0
5     1  0  0  0  0  1  1  0  0  0  1  0  0  0  1  1  0  0  0  1
6     1  1  0  0  0  0  0  0  0  0  0  1  0  1  0  0  1  1  0  1
7     1  0  1  0  0  0  0  0  1  1  1  0  1  1  0  0  1  0  1  0

        CLASSES DE LA PARTITION SIMULEE

        classe numero :   1
           2    5    7    9   11   12   13   15   32   53   57   65   68   80  104
        classe numero :   2
          14   18   21   22   23   25   28   29   42   44   49   52   64   86   92
        classe numero :   3
          17   27   31   34   35   36   37   38   39   45   61   67   79   93   99
        classe numero :   4
           6   19   20   24   46   47   50   54   55   58   59   70   76   87   95
        classe numero :   5
           8   26   33   48   60   62   69   71   72   73   75   77   83   90  105
        classe numero :   6
           1    3   30   41   51   56   66   74   78   81   82   84   85   88   89
        classe numero :   7
           4   10   16   40   43   63   91   94   96   97   98  100  101  102  103
```

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires


        valeur du critere obtenu  :  404


        partition obtenue :

classe  1 :    14 elements

  8   26   33   48   60   62   69   71   72   73   75   77   83  105

classe  2 :    17 elements

  1    3   11   30   41   51   56   65   66   74   78   81   82   84   85
 88   89

classe  3 :    18 elements

 14   18   21   22   23   25   28   29   32   39   42   44   49   52   64
 86   92   93

classe  4 :    17 elements

  4   10   16   40   43   63   67   90   91   94   96   97   98  100  101
102  103

classe  5 :    13 elements

 17   27   31   34   35   36   37   38   45   61   79   80   99

classe  6 :    14 elements

  6   20   24   46   47   50   54   55   58   59   70   76   87   95

classe  7 :    12 elements

  2    5    7    9   12   13   15   19   53   57   68  104


        tableau des valeurs ideales


        VVVVVVVVVVVVVVVVVVV
        AAAAAAAAAAAAAAAAAAA
                 11111111112
        12345678901234567890

      1 1     11   1   11   1
      2 11          1 1  11 1
      3  1  1   1 1   1 11 1
      4 1 1     111 11  1 1
      5 111   111   1    1
      6 1  111   1        11
      7   11    11 1 1 1 11
```

COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables quantitatives

pourcentage d inertie expliquee en 7 classes : 37.85

partition obtenue :

classe numero 1 (effectif= 11)
18 22 25 29 32 49 52 64 86 89 92

classe numero 2 (effectif= 14)
8 26 33 48 60 62 69 71 72 73 75 77 83 105

classe numero 3 (effectif= 16)
4 10 36 40 43 63 90 91 94 96 97 98 100 101 102
103

classe numero 4 (effectif= 13)
2 5 7 9 12 13 15 19 53 57 67 68 104

classe numero 5 (effectif= 15)
6 16 20 24 46 47 50 54 55 58 59 70 76 87 95

classe numero 6 (effectif= 15)
14 17 21 23 28 31 34 35 37 39 42 45 61 80 93

classe numero 7 (effectif= 21)
1 3 11 27 30 38 41 44 51 56 65 66 74 78 79
81 82 84 85 88 99

coordonnees des centres de gravite des variables pour chaque classe de la partition :

variable (effectif)	population (105)	classe 1 (11)	classe 2 (14)	classe 3 (16)	classe 4 (13)	classe 5 (15)	classe 6 (15)	classe 7 (21)
VA 1	.590	.182	.786	.938	.769E-01	.800	.267	.810
VA 2	.486	.818	.214	.625E-01	.308	.200	.933	.810
VA 3	.448	.000	.214	.813	.923	.200	.467	.429
VA 4	.381	.273	.286	.125	1.00	1.00	.667E-01	.952E-01
VA 5	.419	.909	.214	.250	.769E-01	.867	.400	.333
VA 6	.410	.364	.929	.313	.154	.733	.133	.286
VA 7	.390	.182	.786	.313	.154	.267	.933	.143
VA 8	.352	.000	.143	.375	.231	.400	.733	.429
VA 9	.571	.727	.143	1.00	.846	.133	1.00	.286
VA10	.410	.364	.000	.750	.769	.800	.200	.952E-01
VA11	.505	.727	.857	.875	.308	.267	.533	.143
VA12	.410	.273	.286	.313	.846	.667E-01	.133	.810
VA13	.419	.273	.143	.938	.231	.267	.467	.476
VA14	.457	.273	.214	.875	1.00	.667E-01	.133	.571
VA15	.381	1.00	.857	.188	.308	.133	.467	.476E-01
VA16	.381	.000	.786	.125	.769	.333	.467	.238
VA17	.429	1.00	.714E-01	.688	.154	.200	.267	.619
VA18	.667	1.00	.357	.438	.538	.733	.733	.857
VA19	.438	.273	.214	.875	.769	.733	.667E-01	.190
VA20	.505	1.00	.857	.125	.769E-01	.200	.467	.810

CHAPITRE 3

CLASSIFICATION SUR TABLEAU DE CODAGE BINAIRE ADDITIF

1. INTRODUCTION

Nous proposons ici une méthode de classification pour les tableaux de données binaires résultant de la transformation, par le codage binaire additif, d'un tableau de modalités croisant individus et variables qualitatives ordinales. Celle-ci est construite en respectant un principe de fidélité des résultats par rapport à la forme initiale des données. La méthode de type nuées dynamiques va permettre de réaliser cet objectif.

A partir du tableau de codage, nous définissons un ensemble de points représentant l'ensemble à classifier dans l'espace $\mathbf{B}^m$ (où m est le nombre total de modalités). En fait, ces points ont une structure particulière : ce sont des vecteurs binaires de modalités (les composantes vérifient la contrainte de codage binaire additif). C'est sur cet ensemble que sera appliquée la méthode de type Nuées Dynamiques.

Dans un premier paragraphe, nous montrons principalement que l'algorithme MNDBIN peut être appliqué à ce type de données. Sans lui apporter aucune modification, il fournit des noyaux ayant une structure de vecteur binaire de modalités. Il nous reste alors à préciser la signification de ces noyaux et du critère associé à la partition obtenue. Nous montrons ensuite qu'il est possible d'écrire une version de cet algorithme utilisant directement le tableau de modalités et que nous appelons MNDORD. En fait, cette version correspond à une méthode de type Nuées Dynamiques dont nous préciserons les particularités.

Les variables étant qualitatives ordinales, elles peuvent être traitées comme des variables quantitatives. Les méthodes de classification utilisées dans le cas quantitatif peuvent être appliquées.

Dans le second paragraphe, nous nous intéressons plus particulièrement à l'utilisation de la méthode des Nuées Dynamiques utilisant la distance euclidienne usuelle et les centres de gravité des classes comme noyaux. Il s'agit de la méthode appelée MNDQAN (G. Celeux et al. 1989), pour laquelle deux applications sont envisagées. Nous l'appliquons tout d'abord aux données qualitatives ordinales, puis aux données binaires résultant du codage. Dans le second cas, l'application est très particulière et nous détaillons ici tous les résultats obtenus. Nous disposons ainsi de trois méthodes de classification (MNDBIN et les deux applications de MNDQAN) pour ce type de données. Une étude comparative est alors effectuée entre les différents critères associés.

Dans une dernière partie, nous présentons le programme MNDORD. Celui-ci a été intégré au logiciel d'analyse de données SICLA. Une application sur un exemple conséquent est également proposée.

2. LA MÉTHODE DE CLASSIFICATION

Dans le chapitre précédent, nous avons décrit le principe des méthodes des Nuées Dynamiques, ainsi que l'algorithme qui en découle. Sur ce principe, nous avons construit la méthode MNDBIN qui repose essentiellement sur le choix de la distance en valeurs absolues et de noyaux binaires. Cette méthode était plus précisément adaptée aux tableaux de variables binaires.

Ici, nous disposons d'un tableau de données binaires très particulier, puisqu'il résulte de la transformation, par le codage binaire additif, d'un tableau de modalités croisant individus et variables qualitatives ordinales. A partir du tableau binaire, on définit un nuage de points représentant l'ensemble à classifier dans l'espace $\mathbf{B}^m$ (où m est le nombre total de modalités). Par construction, les éléments du nuage ont une structure de vecteur binaire de modalités.

Il s'agit d'écrire une méthode de classification fournissant des résultats qui soient en accord avec la structure initiale des données.

Nous montrons ici que la méthode MNDBIN réalise cet objectif. En effet, cet algorithme fournit des noyaux ayant une structure de vecteur binaire de modalités. Après en avoir fait la démonstration, nous étudions le critère associé à la partition et précisons la signification des noyaux.

Nous proposons également une version de MNDBIN utilisant directement le tableau de modalités. En fait, il est possible de montrer que cette version correspond à une méthode de type Nuées Dynamiques utilisant la distance en valeurs absolues sur l'espace des vecteurs de modalités.

2.1 RAPPEL DES NOTATIONS

Soit $\mathbf{Z}(I,Q)$ le tableau de modalités croisant un ensemble $I=\{1,2,\ldots,n\}$ de n individus et un ensemble $Q=\{1,2,\ldots,p\}$ de p variables qualitatives ordinales. On note :

$$\mathbf{Z}(I,Q) = (z_i^q)$$

où z_i^q représente la modalité de la variable q choisie par l'individu i.

A chaque variable q correspond l'ensemble $J_q=\{1,2,\ldots,m_q\}$ de modalités. Nous définissons l'espace **E** comme le produit $J_1 x J_2 x \ldots x J_p$. Il s'agit de l'espace des vecteurs de modalités que nous munissons de la distance en valeurs absolues notée d_E. Les caractéristiques associées à d_E ont été définies dans le chapitre 1, paragraphe 4. En particulier, nous avons déjà montré que tout nuage de cet espace admet un vecteur de modalités pour centre médian.

A partir du tableau $\mathbf{Z}(I,Q)$, nous définissons le nuage de points $\mathbf{N_Z}(I)$, inclus dans l'espace **E**, par :

$$\mathbf{N_Z}(I) = \{ z_i, i \in I \}$$

$$\text{où} \quad z_i = (z_i^1, z_i^2, \ldots, z_i^p) \in \mathbf{E}$$

Soit $\mathbf{X}(I,J)$ le tableau de codage binaire additif correspondant au tableau de modalités $\mathbf{Z}(I,Q)$. C'est un tableau binaire à n lignes et m colonnes, où m est le nombre total de modalités. On utilise l'ensemble $J=\{1,2,\ldots,m\}$ pour indicer les colonnes de $\mathbf{X}(I,J)$.

On note :

$$\mathbf{X(I,J)} = (x_i^j)$$

En appelant q(j) l'indice de J correspondant à la modalité j de la variable q, on a :

$$\forall\, q \in Q,\ \forall\, j \in J_q \quad x_i^{q(j)} = \begin{cases} 1 \text{ si } j \le z_i^q \\ 0 \text{ si } j > z_i^q \end{cases}$$

$$\forall\, q \in Q \quad z_i^q = \sum_{j \in Jq} x_i^{q(j)}$$

Nous définissons l'espace **F** comme la restriction de $\mathbf{B^m}$ aux seuls vecteurs binaires de modalités. Les éléments de **F** résultent alors du codage des vecteurs de modalités de l'espace **E**. De même, à tout élément de **E** correspond un élément de **F**.

A partir du tableau **X(I,J)**, nous définissons le nuage **N(I)** par :

$$\mathbf{N(I)} = \{\, x_i,\ i \in I \,\}$$

$$\text{où} \quad x_i = (x_i^1, x_i^2, \ldots, x_i^m) \in \mathbf{F}$$

Ainsi, à chaque point x_i du nuage **N(I)** correspond un unique point z_i de $\mathbf{N_Z(I)}$.

2.2 LE PROBLEME

Il s'agit de déterminer une partition de l'ensemble des individus I en K classes, K étant fixé a priori.

L'ensemble à classifier est représenté dans l'espace **F** (inclus dans $\mathbf{B^m}$) par le nuage **N(I)**. Nous pouvons alors appliquer la méthode MNDBIN. En notant d la distance en valeurs absolues, le problème résolu par MNDBIN peut se formuler ainsi :

> *Trouver une partition $P=(P_1,P_2,\ldots,P_K)$ de I et un ensemble de noyaux $L=(a_1,a_2,\ldots,a_K)$ appartenant à $\mathbf{B^m}$ tels que le crirère :*
>
> $$W(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$
>
> *soit minimum.*

Dans la suite, nous montrons que les noyaux fournis par la méthode sont des vecteurs binaires de modalités appartenant à l'espace **F**.

2.3 L'ALGORITHME

Nous reprenons ici les fonctions d'affectation et de représentation qui caractérisent l'algorithme MNDBIN.

La fonction d'affectation f
L'individu i est affecté à la classe P_k dont il est le plus proche du noyau a_k au sens de la distance en valeurs absolues.

La fonction de représentation g
Les noyaux binaires $(a_1,a_2,\ldots,a_K)$ fournis par MNDBIN sont les centres médians des classes $(P_1,P_2,\ldots,P_K)$ de la partition. Le noyau a_k de la classe P_k est le point minimisant la quantité :

$$\sum_{i\in P_k} d(x_i,a_k)$$

Par définition, le noyau a_k est centre médian de l'ensemble $\{x_i, i\in P_k\}$ inclus dans **F**. Dans le chapitre 1, paragraphe 4, nous avons démontré la propriété suivante : un ensemble de points de **F** admet un centre médian appartenant à l'espace **F**. Nous avons également démontré que ce point est obtenu en utilisant simplement la règle de la majorité, de sorte que :

$$\forall\, j\in J \quad a_k^j = \text{médiane binaire } \{\, x_i^j,\ i \in P_k \,\}$$

Le noyau ainsi construit a donc une structure de vecteur binaire de modalités. Lors de cette étape de représentation, nous risquons de rencontrer un problème lié à la règle de la majorité : il s'agit du cas où toute valeur de l'intervalle [0,1] est médiane. Cela peut se produire ici pour plusieurs composantes successives (et correspondant à une même modalité) du noyau a_k. Dans ce cas, nous convenons de choisir pour médiane toujours une même valeur (0 par exemple), de manière à obtenir le codage d'une modalité. La valeur du critère ne dépend pas de ce choix.

La méthode MNDBIN peut donc être appliquée à un tableau de codage binaire additif. Elle fournit implicitement des noyaux appartenant à **F**. A ceux-ci sont associés des vecteurs de modalités de l'espace **E**. Chaque classe est ainsi résumée par un vecteur de modalités; la valeur représentative d'une variable dans une classe est aussi une modalité. Il nous reste maintenant à préciser la signification de tous ces résultats.

2.4 INTERPRETATION DES NOYAUX

Soit a_k le centre médian de l'ensemble $\{x_i, i\in P_k\}$. A cet ensemble correspond, dans l'espace **E**, l'ensemble $\{z_i, i\in P_k\}$ tel que :

$$\forall\, q\in Q,\ \forall\, i\in P_k \quad z_i^q = \sum_{j\in Jq} x_i^{q(j)}$$

où z_i^q est la modalité de la variable q choisie par l'individu i.

De même, au noyau a_k correspond un vecteur de modalités A_k appartenant à l'espace **E** et défini par :

$$A_k = (A_k^1, A_k^2, \ldots, A_k^p)$$

$$\text{où} \quad \forall\, q\in Q \quad A_k^q = \sum_{j\in Jq} a_k^{q(j)} \quad \text{et} \quad A_k^q \in J_q$$

La propriété d'égalité entre la distance en valeurs absolues d sur **F** et d_E sur **E** permet de démontrer que A_k est centre médian de l'ensemble $\{z_i, i\in P_k\}$ (chapitre 1, paragraphe 4). De plus, les composantes de ce point sont les médianes des variables dans la classe considérée, ce qui s'exprime ici par :

$$\forall\, q\in Q \quad A_k^q \text{ médiane de l'ensemble } \{z_i^q,\ i\in P_k\}$$

Si l'ensemble à classifier est inclus dans l'espace **F**, la méthode MNDBIN fournit des noyaux binaires de modalités appartenant à ce même espace. Ces modalités sont en fait des valeurs médianes. Il nous reste maintenant à étudier le critère assiocié à cette

2.5 EXPRESSION DU CRITERE À LA CONVERGENCE

A la convergence, nous pouvons exprimer le critère uniquement par rapport à la partition, de sorte que :

$$\mathbf{W}(P) = \mathbf{W}(P,g(P))$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$

Ce critère représente le nombre de fois où la situation obtenue s'écarte de la situation "idéale" (correspondant au cas où les vecteurs réponses des individus sont indentiques aux noyaux des classes auxquelles ils appartiennent).

Pour ces données binaires particulières, il nous faut préciser la signification du critère **W**(P). Pour cela, considérons les K éléments $(A_1,A_2,\ldots,A_K)$ de **E** correspondant aux K noyaux $(a_1,a_2,\ldots,a_K)$ fournis par l'algorithme. La propriété sur les distances permet d'exprimer le critère directement à partir des données non codées, soit :

$$\mathbf{W}(P) = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k) = \sum_{k=1}^{K} \sum_{i \in P_k} d_E(z_i,A_k)$$

Si on note :

$$\Im(P_k) = \sum_{i \in P_k} d(x_i,a_k) = \sum_{i \in P_k} d_E(z_i,A_k)$$

l'expression du critère devient :

$$\mathbf{W}(P) = \sum_{k=1}^{K} \Im(P_k)$$

On retrouve alors une expression rapelant celle de l'inertie intraclasse habituelle.

2.6 MÉTHODE DE CLASSIFICATION POUR TABLEAU DE MODALITÉS

La méthode MNDBIN fournit une partition P et un ensemble L de K noyaux binaires de modalités minimisant le critère **W**(P,L). Cette méthode utilise le tableau de codage et, d'un point de vue programmation, elle peut devenir coûteuse en place mémoire.

Nous avons démontré que les noyaux fournis par la méthode MNDBIN crrespondent à des vecteurs de modalités. D'autre part, il est possible d'exprimer le critère **W**(P,L) directement en fonction des données non codées. Il est alors possible d'écrire une version de l'algorithme utilisant directement le tableau de modalités.

Comme il apparait déjà dans les paragraphes précédents, cette version de l'algorithme n'est autre que la méthode des Nuées Dynamiques utilisant la distance en valeurs absolues sur l'espace **E**.

En effet, considérons le nuage $\mathbf{N_Z(I)}=\{z_i, i\in I\}$ associé au tableau de modalités $\mathbf{Z(I,Q)}$. Ce nuage représente, dans l'espace **E**, l'ensemble à classifier. Nous munissons **E** de la distance en valeurs absolues d_E. Nous appliquons alors la méthode des Nuées Dynamiques utilisant cette distance. Le critère associé s'exprime de la façon suivante :

$$\mathbf{W(P,L_E)} = \sum_{k=1}^{K} \sum_{i\in P_k} d_E(z_i,A_k)$$

où $L_E = (A_1, A_2, \ldots, A_k)$ est l'ensemble des noyaux.

Il apparait alors que le noyau A_k d'une classe P_k est son centre médian défini par :

$$\forall\, q\in Q \quad A_k^q = \text{médiane}\ \{\ z_i^q, i\in P_k\ \}$$

Comme nous l'avons demontré (chapitre 1, paragraphe 2), la valeur de la médiane correspond à l'une des modalités de la variable q. Le noyau A_k ainsi construit est un vecteur de modalités de l'espace **E**.

D'autre part, nous avons également démontré une propriété (chapitre 1, paragraphe 4) qui, ici, permet d'affirmer que : si A_k est centre médian de $\{z_i, i\in P_k\}$ alors le point a_k, transformation de A_k par le codage binaire additif, est centre médian de $\{x_i, i\in P_k\}$.

Dans ces conditions, le critère optimisé par la méthode est alors identique à celui optimisé par MNDBIN puisque :

$$\mathbf{W_E(P,L_E)} = \sum_{k=1}^{K} \sum_{i\in P_k} d_E(z_i,A_k)$$

$$\Leftrightarrow \quad \mathbf{W_E(P,L_E)} = \sum_{k=1}^{K} \sum_{i\in P_k} d(x_i,a_k)$$

$$\Leftrightarrow \quad \mathbf{W_E(P,L_E) = W(P,L)}$$

La méthode utilisant directement le tableau de modalités est appelée MNDORD (Méthode des Nuées Dynamiques sur variables qualitatives ORDinales).

2.7 INDICES D'AIDE À L'INTERPRÉTATION

Nous proposons ici des indices permettant une analyse plus fine des résutats obtenus par l'algorithme MNDORD (ou MNDBIN).

Il s'agit de fournir des indications sur la qualité globale de la partition, mais aussi sur le comportement des variables dans les différentes classes. Pour cela, nous définissons des indices à partir de l'expression du critère associé à la méthode.

La valeur du critère à la convergence constitue un premier indicateur : il mesure l'écart (somme des valeurs absolues des différences entre données initiales et noyaux correspondants) entre la situation obtenue et la situation idéale. De plus, comme nous l'avons déjà indiqué, il peut être interprété comme une inertie intraclasse (définie à partir

de la distance en valeurs absolues). En divisant cette quantité par le nombre total d'éléments du tableau initial (égal au produit np), on obtient une quantité exprimant la valeur moyenne des valeurs absolues des écarts entre valeurs initiales et noyaux correspondants.

D'autre part, nous proposons un indice permettant d'évaluer séparément l'homogénéité de chacune des classes de la partition. Pour toute classe P_k, il s'agit de l'inertie $\Im(P_k)$ représentant la somme des écarts (en valeurs absolues) entre données initiales et noyaux. Son expression est la suivante :

$$\Im(P_k) = \sum_{i \in P_k} d_E(z_i, A_k) = \sum_{i \in P_k} \sum_{q \in Q} |z_i^q - A_k^q|$$

On obtient un indice moyen en divisant $\Im(P_k)$ par le produit $n_k p$, où n_k est le cardinal de la classe P_k. Il représente alors la valeur moyenne de l'écart entre données initiales et noyaux pour une classe donnée.

Les noyaux de modalités permettent de distinguer rapidement les variables caractérisant le mieux chacune des classes de la partition. En complément, nous proposons pour chaque couple (classe, variable) un indice permettant d'étudier le comportement de la variable dans la classe. Il s'agit de la somme des valeurs absolues des écarts entre les valeurs prises par la variable et la valeur "idéale" fournie par le noyau correspondant. Nous l'appelons plus simplement écart de la variable dans la classe (nous n'avons supposé aucune pondération, on ne parle donc pas d'écart moyen). Il est noté, pour une classe P_k et une variable q, **EC**(k,q) et son expression est :

$$\mathbf{EC}(k,q) = \sum_{i \in P_k} |z_i^q - A_k^q|$$

En divisant cet indice par l'effectif n_k de la classe P_k, on obtient alors l'écart moyen de la variable dans la classe.

Les indices sont à étudier en regard de la taille du tableau (pour le critère) et des effectifs des classes (pour les inerties et les écarts de variables). Quant aux indices moyens, ils sont, eux, directement interprétables.

2.8 EXEMPLE SIMPLE D'APPLICATION

Soit un ensemble de 10 individus, identifiés par les nombres **1** à **10**, décrit par un ensemble de 5 variables qualitatives ordinales, identifiées par les lettres **a** à **e**.

Supposons, par exemple, que chaque variable est associée à un journal. L'individu répond alors à chacune d'elle par les modalités 1, 2 ou 3 suivant que :

- il ne lit pas du tout le journal associé à la variable (modalité1),
- il ne lit que quelquefois ce journal (modalité 2),
- il lit souvent ce journal (modalité 3).

Les réponses des individus sont représentées sous la forme d'un tableau de modalités (figure 1 de la page suivante).

Nous appliquons alors l'algorithme MNDORD en demandant 3 classes. Après plusieurs essais, la meilleure partition obtenue est :

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (\{\mathbf{2, 6, 7}\}, \{\mathbf{1, 4, 5, 8}\}, \{\mathbf{3, 9, 10}\})$$

Celle-ci est représentée (figure 2) en réordonnant les lignes du tableau initial de manière à respecter les classes obtenues.

Nous indiquons également :

figure 3 : les effectifs des classes ainsi que leur inertie,

figure 4 : les noyaux fournis par la méthode,

figure 5 : les écarts des variables dans chaque classe.

Enfin, la valeur du critère égale à 14 mesure l'écart entre la solution obtenue et la situation "idéale". L'indice moyen correspondant est de 0.28 (14/50), ce qui montre que, en moyenne, l'écart entre une donnée initiale et la valeur du noyau correspondant est de 0.28.

	a	b	c	d	e
1	1	2	2	3	2
2	3	2	1	2	1
3	2	3	3	1	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	1	2
7	3	3	2	1	1
8	1	1	1	3	3
9	1	2	2	1	1
10	1	3	2	2	2

figure 1
tableau initial

	a	b	c	d	e
2	3	2	1	2	1
6	3	2	1	1	2
7	3	3	2	1	1
1	1	2	2	3	2
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3
3	2	3	3	1	1
9	1	2	2	1	1
10	1	3	2	2	2

figure 2
tableau réordonné

	effectifs	**critères**
A	3	4
B	4	5
C	3	5

figure 3
descriptif des classes

	a	b	c	d	e
A	3	2	1	1	1
B	1	1	1	3	3
C	1	3	2	1	1

figure 4
les noyaux

	a	b	c	d	e
A	0	1	1	1	1
B	0	2	2	0	1
C	1	1	1	1	1

figure 5
tableau des écarts

3. APPLICATION D'UNE MÉTHODE POUR VARIABLES QUANTITATIVES

Les variables qualitatives ordinales peuvent être traitées comme des variables quantitatives. Ainsi, les méthodes de classification valables dans le cas quantitatif le restent dans le cas qualitatif ordinal.

A partir du tableau de modalités, on peut définir un nuage de points représentant l'ensemble à classifier dans $\mathbf{R}^p$ (où p est le nombre de variables). Nous pouvons alors appliquer la méthode MNDQAN (déja présentée et utilisée dans le premier chapitre). Cette méthode peut aussi être appliquée à partir du tableau de codage binaire additif. Dans ce paragraphe, nous décrivons les résultats et les particularités de ces deux

3.1 APPLICATION DE LA MÉTHODE AU TABLEAU DE MODALITÉS

Reprenons la notation $\mathbf{N}_Z(I)=\{z_i, i \in I\}$ pour le nuage associé au tableau de modalités $\mathbf{Z}(I,Q)$. Ce nuage est inclus dans l'espace $\mathbf{R}^p$ que nous munissons de la distance euclidienne notée D_e.

3.1.1 La méthode

Le problème de classification posé est de trouver une partition de l'ensemble I en K classes, K étant fixé a priori. L'ensemble à classifier est représenté ici, dans l'espace $\mathbf{R}^p$, par le nuage $\mathbf{N}_Z(I)$.

La méthode MNDQAN fournit une solution à ce problème en optimisant le critère $\mathbf{W}_1(P,L_G)$ défini par :

$$\mathbf{W}_1(P,L_G) = \sum_{k=1}^{K} \sum_{i \in P_k} D_e^2(z_i,G_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{q \in Q} (z_i^q - G_k^q)^2$$

où

$P = (P_1, P_2, \ldots, P_K)$ est une partition de I en K classes,

$L_G = (G_1, G_2, \ldots, G_K)$ est l'ensemble des K noyaux des classes.

Les fonctions caractérisant l'algorithme MNDQAN sont :

- la fonction d'affectation (**f**) : un individu est affecté à la classe P_k dont il est le plus proche du noyau G_k (au sens de la distance D_e).
- la fonction de représentation (**g**) : les noyaux sont les centres de gravité des classes. Si on note n_k le cardinal de la classe P_k, on a :

$$\forall\, k=1,2,\ldots,K,\ \forall\, q \in Q \quad G_k^q = \frac{1}{n_k} \sum_{i \in P_k} z_i^q$$

La valeur représentative d'une variable est donc la moyenne des modalités prises par les individus d'une classe. Les modalités étant ordonnées, l'interprétation de cette moyenne est possible, il s'agit d'une valeur comprise entre la première et la dernière modalité. D'autre part, le critère $\mathbf{W}_1(P,L_G)$ représente l'inertie intraclasse de la partition des individus.

3.1.2 La méthode avec contrainte sur les noyaux

Nous pouvons reconsidérer la méthode de classification en imposant aux noyaux d'être des vecteurs de modalités. Par rapport à l'algorithme existant, seule change la fonction de représentation.

Il s'agit ici de déterminer, pour tout k, le vecteur de modalités A_k minimisant la quantité :

$$\sum_{i \in P_k} d_e^2(z_i, A_k)$$

La composante A_k^q du noyau doit donc minimiser la quantité :

$$\sum_{i \in P_k} (z_i^q - A_k^q)^2$$

avec la contrainte $A_k^q \in J_q = \{1, 2, \ldots, m_q\}$.

Le problème se résoud assez simplement. La valeur minimisant cette quantité est la valeur entière, inférieure ou supérieure, la plus proche de la moyenne de la variable q dans la classe P_k. C'est une caractéristique différente de celles fournies par les méthodes MNDBIN (médiane) et MNDQAN (moyenne), sauf cas très particulier.

Pour illustrer ce résultat, considérons l'échantillon {1,1,1,4,4} d'une variable à 4 modalités. Les différentes caractéristiques de l'échantillon sont les suivantes : une moyenne égale à 2.2, une médiane égale à 1 et, comme caractéristique correspondant au problème avec contrainte, la valeur 2.

En pratique, cette méthode se révèle très voisine de la méthode sans contrainte MNDQAN. De plus lors du déroulement de l'algorithme, il se pose souvent le problème de noyaux égaux et donc de classes vides (les domaines de définition des variables sont restreints).

3.2 APPLICATION DE LA MÉTHODE AU TABLEAU DE CODAGE

Reprenons la notation $N(I)=\{x_i, i \in I\}$ pour le nuage associé au tableau de codage binaire additif $X(I,J)$. Ce nuage est inclus dans l'espace $\mathbf{R}^m$ (où m est le nombre total de modalités) que nous munissons de la distance euclidienne notée d_e.

3.2.1 La méthode

Le problème de classification posé est de trouver une partition de I en K classes, K fixé a priori. L'ensemble à classifier est représenté, dans l'espace $\mathbf{R}^m$, par le nuage N(I).

La méthode MNDQAN fournit une solution à ce problème en optimisant le critère suivant :

$$W_2(P,L_g) = \sum_{k=1}^{K} \sum_{i \in P_k} d_e(x_i, g_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (x_i^j - g_k^j)^2$$

où

$P = (P_1, P_2, \ldots, P_K)$ est une partition de I en K classes,

$L_g = (g_1, g_2, \ldots, g_K)$ est l'ensemble des K noyaux des classes.

L'algorithme est construit de manière habituelle :

- la fonction d'affectation : chaque individu est rangé dans la classe P_k dont il est le plus proche du noyau g_k (au sens de la distance d_e).
- la fonction de représentation : les noyaux sont les centres de gravité des classes. Si on note n_k le cardinal de la classe P_k, on a :

$$\forall\, k=1,2,\ldots,K,\ \forall\, j\in J \quad g_k^j = \frac{1}{n_k} \sum_{i\in P_k} x_i^j$$

Le critère $W_2(P,L_g)$ représente l'inertie intraclasse de la partition. Les noyaux, très particuliers, sont étudiés dans le paragraphe suivant.

3.2.2 Interprétation des noyaux

Considérons une classe P_k de la partition et les vecteurs G_k de $\mathbf{R}^p$ et g_k de $\mathbf{R}^m$ définis par :

$$G_k = \text{centre de gravité de } \{\, z_i,\ i\in P_k \,\}$$

$$g_k = \text{centre de gravité de } \{\, x_i,\ i\in P_k \,\}$$

Par construction, un point x_i est le codage du point z_i. Cela permet d'établir une relation entre les centres de gravité puisque :

$$\forall\, q\in Q \quad G_k^q = \frac{1}{n_k} \sum_{i\in P_k} z_i^q$$

$$\Leftrightarrow \quad \forall\, q\in Q \quad G_k^q = \frac{1}{n_k} \sum_{i\in P_k} \sum_{j\in Jq} x_i^{q(j)}$$

$$\Leftrightarrow \quad \forall\, q\in Q \quad G_k^q = \sum_{j\in Jq} g_k^{q(j)}$$

où q(j) est l'indice de J correspondant à la modalité j de la variable q.

Le noyau G_k peut donc être recalculé à partir du noyau g_k (la réciproque est fausse). Cependant, comme nous le constaterons dans la suite, les deux applications de la méthode MNDQAN aboutissent à des résultats différents.

Le centre de gravité g_k de la classe P_k posséde d'autres propriétés, qui sont toutes des conséquences du codage binaire additif.

Nous avons tout d'abord le résultat suivant :

$$\forall\, q\in Q \quad g_k^{q(1)} \geq g_k^{q(2)} \geq \ldots \geq g_k^{q(m_q)}$$

Les composantes de g_k fournissent des informations concernant les valeurs prises par les variables dans la classe P_k. Une composante q(j) de g_k représente, de par sa définition, la proportion d'individus de la classe P_k ayant choisis une modalité supérieure ou égale à la modalité j de la variable q.

Le point g_k fournit en outre des informations sur les proportions des modalités dans la classe considérée. En effet, si j et j+1 sont deux modalités successives d'une variable q, nous avons :

$$g_k^{q(j)} - g_k^{q(j+1)} = \frac{1}{n_k} \sum_{i \in P_k} x_i^{q(j)} - \frac{1}{n_k} \sum_{i \in P_k} x_i^{q(j+1)}$$

$$\Leftrightarrow \quad g_k^{q(j)} - g_k^{q(j+1)} = \frac{1}{n_k} \sum_{i \in P_k} (x_i^{q(j)} - x_i^{q(j+1)})$$

Toutes les valeurs 1 de la colonne q(j+1) se retrouvent, en même position, dans la colonne q(j) qui, elle, contient des valeurs 1 supplémentaires. La somme ci-dessus représente alors le nombre des valeurs 1 non communes aux deux colonnes q(j) et q(j+1) : la somme est donc égale au nombre d'individus de P_k ayant choisi la modalité j de la variable q.

Si on note $n_k^{q(j)}$ ce nombre, nous obtenons :

$$g_k^{q(j)} - g_k^{q(j+1)} = \frac{n_k^{q(j)}}{n_k}$$

ce qui correspond à la proportion d'individus de P_k ayant choisi la modalité j de la variable q. Connaissant le noyau (g_k) et l'effectif (n_k) d'une classe, il est alors possible de reconstituer exactement et de façon unique les échantillons des valeurs prises par les variables dans la classe considérée. Cela n'est pas concevable pour le noyau G_k dont les composantes sont des valeurs moyennes.

Nous allons illustrer toutes ces propriétés, en les appliquant à partir de l'échantillon {1,1,2,2,3} d'une variable qualitative ordinale à modalités dans {1,2,3}. Les deux centres G et g associés à l'échantillon sont calculés, les résultats obtenus sont les

modalités	codage
1	1 0 0
1	1 0 0
2	1 1 0
2	1 1 0
3	1 1 1
G = 1.8	$g = 1 \ \frac{3}{5} \ \frac{1}{5}$

A partir du centre g, nous reconstituons l'échantillon de la façon suivante :

- proportion de modalité 1 : $g^1 - g^2 = 1 - \frac{3}{5} = \frac{2}{5}$
- proportion de modalité 2 : $g^2 - g^3 = \frac{3}{5} - \frac{1}{5} = \frac{2}{5}$
- proportion de modalité 3 : $g^3 = \frac{1}{5}$

3.2.3 Exemple d'application

Reprenons le tableau de modalités, croisant 10 individus et 5 variables qualitatives ordinales, déjà décrit dans le paragraphe 2.8. Ce tableau est représenté dans la figure 1 ci-dessous.

Dans un paragraphe précédent, nous avons utilisé cet exemple pour illustrer la méthode MNDORD. Celle-ci avait fourni la partition suivante :

$$(\{2, 6, 7\}, \{1, 4, 5, 8\}, \{3, 9, 10\})$$

Nous appliquons la méthode MNDQAN sur le tableau de codage binaire additif de ce tableau. Trois classes sont demandées et la meilleure solution obtenue est :

$$(\{2, 6\}, \{1, 4, 5, 8\}, \{3, 7, 9, 10\})$$

Celle-ci est représentée (figure 2) en réordonnant les lignes de manière à respecter les classes de la partition. Notons enfin que cette partition permet d'expliquer 59% de l'inertie totale.

	a	b	c	d	e
1	1	2	2	3	2
2	3	2	1	2	1
3	2	3	3	1	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	1	2
7	3	3	2	1	1
8	1	1	1	3	3
9	1	2	2	1	1
10	1	3	2	2	2

figure 1
tableau initial

	a	b	c	d	e
2	3	2	1	2	1
6	3	2	1	1	2
1	1	2	2	3	2
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3
3	2	3	3	1	1
7	3	3	2	1	1
9	1	2	2	1	1
10	1	3	2	2	2

figure 2
tableau réordonné

3.2.4 La méthode avec contrainte de noyaux de modalités

Si nous reprenons la méthode de classification MNDQAN en imposant aux noyaux d'être des vecteurs de modalités de $\mathbf{B^m}$ (c'est à dire appartenant à l'espace **F**), le critère à optimiser devient :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} d_e^2(x_i, a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (x_i^j - a_k^j)^2$$

$$\Leftrightarrow \quad \mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in I} \sum_{j \in J} |x_i^j - a_k^j|$$

où $L = (a_1, a_2, \ldots, a_K)$ est l'ensemble des noyaux appartenant à **F**.

Nous retrouvons le critère optimisé par la méthode MNDBIN, celle-ci est donc équivalente à la méthode MNDQAN avec contrainte de noyaux de modalités.

4. ETUDE COMPARATIVE

Dans un premier temps, nous comparons les résultats obtenus par les deux applications de la méthode MNDQAN. Puis, nous étudions les différences entre cette méthode et la méthode MNDORD. Enfin, une application est proposée sur un tableau de petite taille.

4.1 COMPARAISON DES DEUX APPLICATIONS

Soit P une partition en K classes de l'ensemble des individus. Soient également les centres de gravité des classes $L_G=(G_1,G_2,\ldots,G_K)$ et $L_g=(g_1,g_2,\ldots,g_K)$, respectivement inclus dans $\mathbf{R}^p$ (pour les données non codées) et dans $\mathbf{R}^m$ (pour les données codées). Nous allons déterminer une relation liant les critères $\mathbf{W_1}(P,L_G)$ et $\mathbf{W_2}(P,L_g)$ associés aux deux applications de la méthode MNDQAN.

L'expression du critère $\mathbf{W_1}(P,L_G)$ peut être décomposé de la façon suivante :

$$\Leftrightarrow \quad \mathbf{W_1}(P,L_G) = \sum_{k=1}^{K} \sum_{i\in P_k} \sum_{q\in Q} (z_i^q - G_k^q)^2$$

$$\Leftrightarrow \quad \mathbf{W_1}(P,L_G) = \sum_{k=1}^{K} \sum_{i\in P_k} \sum_{q\in Q} \Big(\sum_{j\in Jq} (x_i^{q(j)} - g_k^{q(j)})^2 \Big)$$

$$\Leftrightarrow \quad \mathbf{W_1}(P,L_G) = \mathbf{W_2}(P,L_g) + \sum_{k=1}^{K} \sum_{i\in P_k} \sum_{q\in Q} \sum_{j\in Jq} \sum_{j'\neq j\in Jq} (x_i^{q(j)} - g_k^{q(j)})(x_i^{q(j')} - g_k^{q(j')})$$

Les problèmes d'optimisation des critères sont différents. Les deux applications de la méthode MNDQAN (sur tableau de modalités et tableau de codage) vont donc fournir des résultats différents.

Les propriétés des noyaux (de type g_k) permettent d'éclairer les différences entre les deux applications de la méthode MNDQAN. C'est ce que nous allons illustrer à travers un exemple simple.

Soit un ensemble de 9 individus décrit par une seule variable qualitative ordinale à modalités dans {1,2,3}. On suppose en outre que toutes les pondérations sont indentiques. Soit {1,1,2,2,3,3,3,3,3} l'échantillon des valeurs observées.

Nous nous intéressons ici la première étape d'affection de la méthode MNDQAN à partir de la partition initiale en deux classes constituées des valeurs suivantes :

$$(\{3, 3, 3, 2, 2\}, \{1, 3, 1, 3\})$$

Les noyaux de ces deux classes pour les deux applications sont les suivants :

- 1ére classe $\quad G_1 = 2.6 \qquad g_1 = (1, 1, \frac{3}{5})$
- 2éme classe $\quad G_2 = 2 \qquad g_2 = (1, \frac{1}{2}, \frac{1}{2})$

On constate que G_2 prend la valeur 2, alors que les individus de la seconde classe ne choisissent jamais la modalité 2. Considérons maintenant l'étape d'affectation et, en particulier, celle de l'individu représenté par la valeur z=2 dans **R**, par le point x=(1,1,0) dans $\mathbf{R}^3$.

Les distances entre les noyaux et l'individu sont alors les suivantes :

- $d_1^2(G_1,z) = 0.64$ $\qquad$ $d_2^2(g_1,x) = 0.36$
- $d_1^2(G_2,z) = 0$ $\qquad$ $d_2^2(g_2,x) = 0.5$

L'individu représenté par le point z est affecté à la deuxième classe et, dans l'autre cas, le même individu représenté par le point x est affecté à la première classe. Les noyaux des classes sont plus proches des données initiales lorsque les données sont codées. Ainsi, l'affectation tient compte des proportions des modalités dans les classes. Cela n'est pas le cas pour les données non codées, l'individu a été affecté à une classe ne contenant pas de modalité 2.

Nous pouvons alors situer l'application de la méthode MNDQAN au tableau de codage entre la méthode MNDBIN (la médiane est une modalité choisie par au moins un individu de la classe) et la méthode MNDQAN appliquée au tableau de modalités (la moyenne n'est pas nécessairement une modalité choisie par un individu de la classe).

4.2 COMPARAISON DES MÉTHODES

Il nous reste à comparer la méthode MNDORD avec les deux applications de la méthode MNDQAN. Pour cela, considérons toujours une partition P de I en K classes. Soit $L=(a_1,a_2,\ldots,a_K)$ les noyaux fournis par MNDBIN. Le critère associé s'exprime par :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i\in I} \sum_{j\in J} |x_i^j - a_k^j|$$

Plaçons nous maintenant dans l'espace $\mathbf{R}^m$ muni de la distance euclidienne d_e. La relation de Huyghens permet alors d'exprimer le critère $\mathbf{W}(P,L)$ en fonction de

$$\sum_{k=1}^{K} \sum_{i\in P_k} d_e^2(x_i,a_k) = \sum_{k=1}^{K} \sum_{i\in P_k} d_e^2(x_i,g_k) + \sum_{k=1}^{K} n_k d_2^2(g_k,a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P,L) = \mathbf{W_2}(P,L_g) + \sum_{k=1}^{K} n_k\, d_2^2(g_k,a_k)$$

où n_k est l'effectif de la classe P_k.

Cela montre bien la différence entre les problèmes d'optimisation de ces deux critères. Cette relation et celle liant $\mathbf{W_1}(P,L_G)$ et $\mathbf{W_2}(P,L_g)$ montrent alors que ces trois méthodes ne sont pas équivalentes.

4.3 APPLICATION ILLUSTRATIVE

Nous proposons ici d'appliquer les méthodes à partir d'un même tableau de données. Les résultats obtenus illustrent bien les différences entre les trois algorithmes décrits dans ce chapitre.

Considérons un ensemble de 10 individus, indentifiés par les nombres **1** à **10**, décrit par un ensemble de 5 variables qualitatives ordinales, identifiées par las lettres **a** à **e**. Nous supposons que toutes les variables sont à modalités dans {1,2,3}. Les réponses des individus sont représentées sous la forme d'un tableau de modalités (figure 1).

Nous appliquons les méthodes en demandant 3 classes. Plusieurs essais sont effectués et on ne retient, à chaque fois, que le meilleur résultat. On aboutit à trois partitions différentes représentées ci-dessous en réordonnant les lignes du tableau initial en respectant les classes obtenues. Le critère associé à la méthode MNDORD (figure 2) est égal à 17 (en moyenne, l'écart entre données initiales et noyaux est de 17/50=0.34). L'inertie expliquée par la partition fournie par MNDQAN (figure 3) appliqué au tableau de modalités est de 47%. Pour l'application au tableau de codages (figure 4), on obtient une inertie expliquée de 45%.

	a	b	c	d	e
1	1	2	3	2	2
2	2	2	3	2	3
3	2	3	2	3	1
4	3	3	1	1	3
5	3	1	1	1	1
6	1	3	2	2	2
7	1	3	1	1	3
8	2	1	2	3	3
9	3	3	1	3	2
10	2	2	2	3	3

figure 1
tableau initial

	a	b	c	d	e
1	1	2	3	2	2
3	2	3	2	3	1
6	1	3	2	2	2
7	1	3	1	1	3
2	2	2	3	2	3
8	2	1	2	3	3
10	2	2	2	3	3
4	3	3	1	1	3
5	3	1	1	1	1
9	3	3	1	3	2

figure 2
MNDORD

	a	b	c	d	e
5	3	1	1	1	1
1	1	2	3	2	2
2	2	2	3	2	3
8	2	1	2	3	3
10	2	2	2	3	3
3	2	3	2	3	1
4	3	3	1	1	3
6	1	3	2	2	2
7	1	3	1	1	3
9	3	3	1	3	2

figure 3
MNDQAN
sur modalités

	a	b	c	d	e
3	2	3	2	3	1
6	1	3	2	2	2
9	3	3	1	3	2
4	3	3	1	1	3
5	3	1	1	1	1
7	1	3	1	1	3
1	1	2	3	2	2
2	2	2	3	2	3
8	2	1	2	3	3
10	2	2	2	3	3

figure 4
MNDQAN
sur données codées

Pour chaque méthode, nous avons également effectué des essais en utilisant, comme points initiaux, les partitions fournies par les autres méthodes. Pour MNDORD, le critère obtenu est toujours supérieur à 17. Pour MNDQAN appliqué au tableau de modalités, on obtient une même inertie expliquée (47%) en partant de la partition fournie par l'application sur les données codées (la partition obtenue est celle de la figure 4). Pour MNDQAN appliqué au tableau de codage binaire additif, le critère n'est

5. PROGRAMME ET APPLICATION

5.1 PRÉSENTATION DU PROGRAMME

Comme MNDBIN, le programme MNDORD a été écrit en respectant le manuel du programmeur proposé dans le cadre du logiciel SICLA. Il s'applique à tout tableau croisant individus et variables qualitatives ordinales et offre les mêmes choix initiaux que la méthode MNDBIN.

5.2 APPLICATION DE LA MÉTHODE

Nous proposons d'étudier les données classiques de R.A. Fisher (1936) concerant 3 populations d'Iris : Setosa (**Seto**), Versicolor (**Vers**) et Virginica (**Virg**). On dispose d'un échantillon de 50 éléments pour chaque type d'Iris. Ces objets sont numérotés de **1** à **50** pour la population Setosa, de **51** à **100** pour Versicolor et de **101** à **150** pour Virginica. Chaque Iris est caractérisé par 4 variables quantitatives : la longueur du sépale (**LoSe**), la largeur du sépale (**LaSe**), la longueur du pétale (**LoPe**), la largeur du pétale (**LaPe**). Sur la page 91, nous présentons le tableau des mesures exprimées en centimètres.

Chaque variable quantitative est découpée en 3 intervalles de même taille, et à chaque intervalle est associée une modalité. Nous obtenons alors 4 variables qualitatives ordinales à 3 modalités chacune. Cette opération est réalisée très simplement en utilisant la commande CREQAL de SICLA. On indique, page 92, le détail de ce découpage et, page 93, le tableau de modalités obtenu. Les 4 variables créées sont identifiées par :

- **los** pour la longueur du sépale,
- **las** pour la largeur du sépale,
- **lop** pour la longueur du pétale,
- **lap** pour la largeur du pétale.

Dans un premier temps, nous étudions la partition naturelle des trois populations d'Iris. Nous essayons de determiner les variables qui discriminent le mieux cette partition. Deux approches sont envisagées, en considérant d'abord le tableau de mesures, puis le tableau de modalités. Des règles simples de décision seront aussi établies.

Approche MNDQAN

Pour les données quantitatives, nous effectuons une approche du type MNDQAN. La commande INPAQN de SICLA fournit alors toute une série d'indices d'aide à l'interprétation de la partition. Ces indices figurent en pages 94 et 95, où la population **Virg** correspond à la classe 1, **Vers** à la classe 2 et **Seto** à la classe 3. L'inertie expliquée par la partition est correcte : **86.89%**. Nous résumons ici les principaux résultats.

Les classes sont caractérisées par leur centre de gravité :

	Iris	**Virg**	**Vers**	**Seto**
LaPe	1.20	1.33	2.03	0.25
LoPe	3.76	4.26	5.55	1.46
LaSe	3.06	2.77	2.97	3.43
LoSe	5.84	5.94	6.59	5.01

A la lecture de ce tableau, il apparait que les variables relatives aux pétales, **LoPe** (surtout) et **LaPe**, permettent de distinguer très clairement la population **Seto**. Ces deux variables sont d'ailleurs celles ayant le plus fort pouvoir discriminant (d'après les indices **COR**, page 94). D'autre part, la variable **LaPe** joue un rôle important dans la formation des classes **Seto** et **Vers** (d'après les indices **COR** par classe et par variable, page 95). Elle permet de distinguer la population **Vers** de la population **Virg**. Enfin, **Seto** est la population la plus éloignée du comportement moyen des 150 Iris, alors que **Virg** est la plus proche (d'après les contributions relatives des classes à l'inertie interclasse de la partition, page 94).

Sans analyser d'avantage cette structure de données, nous pouvons utiliser la règle de décision suivante : on affecte un Iris à la population dont il est le plus proche du centre de gravité. Si on applique cela aux 150 Iris dont nous disposons, nous aboutissons à 16 éléments mal classés : 14 **Virg** sont affectés à **Vers** et 2 **Vers** sont affectés à **Virg**. La population **Seto** est parfaitement reconstituée.

Approche MNDORD

Nous allons maintenant effectuer une étude semblable en partant du tableau des variables qualitatives ordinales **lap, lop, las** et **los**. Comme nous allons le voir, les résultats obtenus sont plus explicites que les précédents.

Cette fois, chaque classe est caractérisée par son centre médian qui est un vecteur de modalités. Le critère de la partition est égal à **108** et indique que, en moyenne, l'écart entre une modalité initiale et la modalité correspondante du centre médian est de **0.18** (108/600). Ci-dessous, on indique le centre médian de l'ensemble des 150 Iris, ceux des classes, les écarts par classe et par variable et les effectifs par classe et par modalité.

les centres médians

	Iris	Virg	Vers	Seto
lap	2	3	2	1
lop	2	3	2	1
las	2	2	1	2
los	2	3	2	1

les écarts

	Iris	Virg	Vers	Seto
lap	84	5	1	0
lop	62	6	2	0
las	96	21	23	14
los	106	19	14	3

les effectifs

	lap			**lop**			**las**			**los**		
	1	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
Virg	0	5	45	0	6	44	19	29	2	1	17	32
Vers	0	49	1	0	48	2	27	23	0	11	36	3
Seto	50	0	0	50	0	0	1	36	13	47	3	0

Après examen du tableau des écarts, on constate que les deux variables **lop** et **lap** sont très homogènes dans toutes les classes. Elles permettent aussi de distinguer assez nettement les trois populations d'Iris : le genre **Seto** est caractérisé par la modalité 1 de ces variables, le genre **Vers** par la modalité 2 et le genre **Virg** par la modalité 3 (d'après le tableau des noyaux). De plus, la variable **lap** discrimine mieux les populations **Vers** et **Virg** (écart=5+1=6) que ne le fait la variable **lop** (écart=6+2=8).

Une règle de décision consiste à affecter chaque Iris à la population dont il est le plus proche de centre médian (au sens de la distance en valeurs absolues). Nous proposons ici une autre règle, simple et plus explicite, en reprenant les définitions des variables qualitatives ordinales. Les interprétations précédentes permettent de formuler la règle suivante :

Si $0.1 \leq$ **LoPe** ≤ 0.9 et $1. \leq$ **LaPe** ≤ 2.96

Alors genre **Seto**

Sinon Si **LaPe** ≤ 0.9

Alors genre **Vers** ou **Virg**

Sinon Si $0.9 <$ **LaPe** ≤ 1.7

Alors genre **Vers**

Sinon genre **Virg**

Le cas où il y a doute entre les genres **Vers** et **Virg** provient du fait que la modalité 1 de **lap** (et aussi de **lop**) n'apparait jamais dans ces populations (du moins pour les mesures dont nous disposons ici). Le cas échéant, il sera possible de compléter cette règle.

Comme cela apparait dans le tableau des effectifs, l'application aux 150 Iris aboutit finalement à 6 éléments mal classés : 5 **Virg** sont affectés à **Vers** et 1 **Vers** est affecté à **Virg**. On obtient donc un meilleur résulat que précédemment, tout du moins pour l'échantillon dont nous disposons ici.

La partition des trois populations n'est pas celle aboutissant au meilleur critère associé à la méthode MNDORD. La même remarque est valable pour la méthode MNDQAN. C'est ce que nous illustrons dans la suite en appliquant ces deux méthodes et en ne retenant, à chaque fois, que la meilleure solution.

Recherche de la meilleure partition

La meilleure partition en 3 classes fournie par MNDQAN est indiquée en page 96. L'inertie expliquée est de **88.43%**, ce qui correspond à une amélioration de **11.7%** de l'inertie intraclasse de la partition des genres. Au total, il y a 16 différences entre ces deux partitions. Les classes sont les mêmes que celles obtenues après les affectations correspondant à la règle de décision :

- **Classe 1** : 38 éléments dont 36 **Virg** et 2 **Vers**.
- **Classe 2** : 62 éléments dont 48 **Vers** et 14 **Virg**.
- **Classe 3** : 50 éléments du genre **Seto**.

L'application de MNDORD (pages 97 et 98) aboutit à un critère de **96**, ce qui correspond à un écart moyen de **0.16** (96/600) et à une amélioration de **11.1%** du critère de la partition naturelle. Ici, 9 Iris ne figurent pas dans leur classe d'origine. D'autre part, les 3 classes fournies ne correspondent pas à celles obtenues après la règle de décision. Elles sont composées des éléments suivants :

- **Classe 1** : 50 éléments du genre **Seto**.
- **Classe 2** : 45 éléments dont 43 **Virg** et 2 **Vers**.
- **Classe 3** : 55 éléments dont 48 **Vers** et 7 **Virg**.

La règle de décision a permis de bien classer 3 Iris sur les 9 différenciant la partition obtenue par MNDORD et la partition naturelle. D'ailleurs, ces 9 Iris sont ceux qui auraient été mal classés si nous avions utilisé la règle consistant à calculer les distances en valeurs absolues entre Iris et centres médians.

Enfin, signalons que la méthode MNDQAN a également été appliquée au tableau de codage binaire additif ainsi qu'au tableau de modalités. Dans les deux cas, on obtient les mêmes résultats que la méthode MNDORD.

LE TABLEAU DE MESURES EXPRIMEES EN CENTIMETRES

POPULATION SETO

	LoSe	LaSe	LoPe	LaPe
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1
11	5.4	3.7	1.5	0.2
12	4.8	3.4	1.6	0.2
13	4.8	3.0	1.4	0.1
14	4.3	3.0	1.1	0.1
15	5.8	4.0	1.2	0.2
16	5.7	4.4	1.5	0.4
17	5.4	3.9	1.3	0.4
18	5.1	3.5	1.4	0.3
19	5.7	3.8	1.7	0.3
20	5.1	3.8	1.5	0.3
21	5.4	3.4	1.7	0.2
22	5.1	3.7	1.5	0.4
23	4.6	3.6	1.0	0.2
24	5.1	3.3	1.7	0.5
25	4.8	3.4	1.9	0.2
26	5.0	3.0	1.6	0.2
27	5.0	3.4	1.6	0.4
28	5.2	3.5	1.5	0.2
29	5.2	3.4	1.4	0.2
30	4.7	3.2	1.6	0.2
31	4.8	3.1	1.6	0.2
32	5.4	3.4	1.5	0.4
33	5.2	4.1	1.5	0.1
34	5.5	4.2	1.4	0.2
35	4.9	3.1	1.5	0.2
36	5.0	3.2	1.2	0.2
37	5.5	3.5	1.3	0.2
38	4.9	3.6	1.4	0.1
39	4.4	3.0	1.3	0.2
40	5.1	3.4	1.5	0.2
41	5.0	3.5	1.3	0.3
42	4.5	2.3	1.3	0.3
43	4.4	3.2	1.3	0.2
44	5.0	3.5	1.6	0.6
45	5.1	3.8	1.9	0.4
46	4.8	3.0	1.4	0.3
47	5.1	3.8	1.6	0.2
48	4.6	3.2	1.4	0.2
49	5.3	3.7	1.5	0.2
50	5.0	3.3	1.4	0.2

POPULATION VERS

	LoSe	LaSe	LoPe	LaPe
51	7.0	3.2	4.7	1.4
52	6.4	3.2	4.5	1.5
53	6.9	3.1	4.9	1.5
54	5.5	2.3	4.0	1.3
55	6.5	2.8	4.6	1.5
56	5.7	2.8	4.5	1.3
57	6.3	3.3	4.7	1.6
58	4.9	2.4	3.3	1.0
59	6.6	2.9	4.6	1.3
60	5.2	2.7	3.9	1.4
61	5.0	2.0	3.5	1.0
62	5.9	3.0	4.2	1.5
63	6.0	2.2	4.0	1.0
64	6.1	2.9	4.7	1.4
65	5.6	2.9	3.6	1.3
66	6.7	3.1	4.4	1.4
67	5.6	3.0	4.5	1.5
68	5.8	2.7	4.1	1.0
69	6.2	2.2	4.5	1.5
70	5.6	2.5	3.9	1.1
71	5.9	3.2	4.8	1.8
72	6.1	2.8	4.0	1.3
73	6.3	2.5	4.9	1.5
74	6.1	2.8	4.7	1.2
75	6.4	2.9	4.3	1.3
76	6.6	3.0	4.4	1.4
77	6.8	2.8	4.8	1.4
78	6.7	3.0	5.0	1.7
79	6.0	2.9	4.5	1.5
80	5.7	2.6	3.5	1.0
81	5.5	2.4	3.8	1.1
82	5.5	2.4	3.7	1.0
83	5.8	2.7	3.9	1.2
84	6.0	2.7	5.1	1.6
85	5.4	3.0	4.5	1.5
86	6.0	3.4	4.5	1.6
87	6.7	3.1	4.7	1.5
88	6.3	2.3	4.4	1.3
89	5.6	3.0	4.1	1.3
90	5.5	2.5	4.0	1.3
91	5.5	2.6	4.4	1.2
92	6.1	3.0	4.6	1.4
93	5.8	2.6	4.0	1.2
94	5.0	2.3	3.3	1.0
95	5.6	2.7	4.2	1.3
96	5.7	3.0	4.2	1.2
97	5.7	2.9	4.2	1.3
98	6.2	2.9	4.3	1.3
99	5.1	2.5	3.0	1.1
100	5.7	2.8	4.1	1.3

POPULATION VIRG

	LoSe	LaSe	LoPe	LaPe
101	6.3	3.3	6.0	2.5
102	5.8	2.7	5.1	1.9
103	7.1	3.0	5.9	2.1
104	6.3	2.9	5.6	1.8
105	6.5	3.0	5.8	2.2
106	7.6	3.0	6.6	2.1
107	4.9	2.5	4.5	1.7
108	7.3	2.9	6.3	1.8
109	6.7	2.5	5.8	1.8
110	7.2	3.6	6.1	2.5
111	6.5	3.2	5.1	2.0
112	6.4	2.7	5.3	1.9
113	6.8	3.0	5.5	2.1
114	5.7	2.5	5.0	2.0
115	5.8	2.8	5.1	2.4
116	6.4	3.2	5.3	2.3
117	6.5	3.0	5.5	1.8
118	7.7	3.8	6.7	2.2
119	7.7	2.6	6.9	2.3
120	6.0	2.2	5.0	1.5
121	6.9	3.2	5.7	2.3
122	5.6	2.8	4.9	2.0
123	7.7	2.8	6.7	2.0
124	6.3	2.7	4.9	1.8
125	6.7	3.3	5.7	2.1
126	7.2	3.2	6.0	1.8
127	6.2	2.8	4.8	1.8
128	6.1	3.0	4.9	1.8
129	6.4	2.8	5.6	2.1
130	7.2	3.0	5.8	1.6
131	7.4	2.8	6.1	1.9
132	7.9	3.8	6.4	2.0
133	6.4	2.8	5.6	2.2
134	6.3	2.8	5.1	1.5
135	6.1	2.6	5.6	1.4
136	7.7	3.0	6.1	2.3
137	6.3	3.4	5.6	2.4
138	6.4	3.1	5.5	1.8
139	6.0	3.0	4.8	1.8
140	6.9	3.1	5.4	2.1
141	6.7	3.1	5.6	2.4
142	6.9	3.1	5.1	2.3
143	5.8	2.7	5.1	1.9
144	6.8	3.2	5.9	2.3
145	6.7	3.3	5.7	2.5
146	6.7	3.0	5.2	2.3
147	6.3	2.5	5.0	1.9
148	6.5	3.0	5.2	2.0
149	6.2	3.4	5.4	2.3
150	5.9	3.0	5.1	1.8

```
COMMANDE : DESQAL <> description de variables qualitatives


    1. variable  : los   longueur du sepale
       --------

            los1 : los1:=(LoSe>=4.3)&(LoSe<=5.5);

            los2 : los2:=(LoSe>5.5)&(LoSe<=6.7);

            los3 : los3:=(LoSe>6.7)&(LoSe<=7.9);


**********************
* modl * nbre *   %  *
**********************
* los1 *   59 *   39 *   ****************************************
* los2 *   71 *   47 *   ***********************************************
* los3 *   20 *   13 *   *************
**********************


    2. variable  : las   largeur du sepale
       --------

            las1 : las1:=(laSe>=2.)&(laSe<=2.8);

            las2 : las2:=(laSe>2.8)&(laSe<=3.6);

            las3 : las3:=(laSe>3.6)&(laSe<=4.4);


**********************
* modl * nbre *   %  *
**********************
* las1 *   47 *   31 *   *******************************
* las2 *   88 *   59 *   ************************************************************
* las3 *   15 *   10 *   **********
**********************


    3. variable  : lop   longueur du petale
       --------

            lop1 : lop1:=(LoPe>=1.)&(LoPe<=2.96);

            lop2 : lop2:=(LoPe>2.96)&(LoPe<=4.93);

            lop3 : lop3:=(LoPe>4.93)&(LoPe<=6.9);


**********************
* modl * nbre *   %  *
**********************
* lop1 *   50 *   33 *   *********************************
* lop2 *   54 *   36 *   ************************************
* lop3 *   46 *   31 *   *********************************
**********************


    4. variable  : lap   largeur du petale
       --------

            lap1 : lap1:=(laPe>=0.1)&(laPe<=0.9);

            lap2 : lap2:=(laPe>0.9)&(laPe<=1.7);

            lap3 : lap3:=(laPe>1.7)&(laPe<=2.5);


**********************
* modl * nbre *   %  *
**********************
* lap1 *   50 *   33 *   *********************************
* lap2 *   54 *   36 *   ************************************
* lap3 *   46 *   31 *   *********************************
**********************
```

LE TABLEAU DE MODALITÉS

	Seto					Vers					Virg				
		l o s	l a s	l o p	l a p		l o s	l a s	l o p	l a p		l o s	l a s	l o p	l a p
1		1	3	1	1	51	4	3	3	3	101	3	3	4	4
2		1	2	1	1	52	3	3	3	3	102	2	2	3	4
3		1	3	1	1	53	3	2	3	3	103	4	2	4	4
4		1	2	1	1	54	2	1	3	3	104	3	2	4	3
5		1	3	1	1	55	3	2	3	3	105	3	2	4	4
6		2	4	1	1	56	2	2	3	3	106	4	2	4	4
7		1	3	1	1	57	3	3	3	3	107	1	1	3	3
8		1	3	1	1	58	1	1	2	2	108	4	2	4	3
9		1	2	1	1	59	3	2	3	3	109	3	1	4	3
10		1	2	1	1	60	2	2	2	3	110	4	3	4	4
11		2	3	1	1	61	1	1	2	2	111	3	3	3	4
12		1	3	1	1	62	2	2	3	3	112	3	2	3	4
13		1	2	1	1	63	2	1	3	2	113	3	2	4	4
14		1	2	1	1	64	3	2	3	3	114	2	1	3	4
15		2	4	1	1	65	2	2	2	3	115	2	2	3	4
16		2	4	1	1	66	3	2	3	3	116	3	3	3	4
17		2	4	1	1	67	2	2	3	3	117	3	2	4	3
18		1	3	1	1	68	2	2	3	2	118	4	4	4	4
19		2	4	1	1	69	3	1	3	3	119	4	2	4	4
20		1	4	1	1	70	2	1	2	2	120	2	1	3	3
21		2	3	1	1	71	2	2	2	3	121	3	3	4	4
22		1	3	1	1	72	3	2	3	3	122	2	2	3	4
23		1	3	1	1	73	3	1	3	3	123	4	2	4	4
24		1	3	1	1	74	3	2	3	2	124	3	2	3	3
25		1	3	1	1	75	3	2	3	3	125	3	3	4	4
26		1	2	1	1	76	3	2	3	3	126	4	3	4	3
27		1	3	1	1	77	3	2	3	3	127	3	2	3	3
28		2	3	1	1	78	3	2	3	3	128	3	2	3	3
29		2	3	1	1	79	2	2	3	3	129	3	2	4	4
30		1	3	1	1	80	2	2	2	2	130	4	2	4	3
31		1	2	1	1	81	2	1	2	2	131	4	2	4	4
32		2	3	1	1	82	2	1	2	2	132	4	4	4	4
33		2	4	1	1	83	2	2	2	2	133	3	2	4	4
34		2	4	1	1	84	2	2	3	3	134	3	2	3	3
35		1	2	1	1	85	2	2	3	3	135	3	2	4	3
36		1	3	1	1	86	2	3	3	3	136	4	2	4	4
37		2	3	1	1	87	3	2	3	3	137	3	3	4	4
38		1	3	1	1	88	3	1	3	3	138	3	2	4	3
39		1	2	1	1	89	2	2	3	3	139	2	2	3	3
40		1	3	1	1	90	2	1	3	3	140	3	2	3	4
41		1	3	1	1	91	2	2	3	2	141	3	2	4	4
42		1	1	1	1	92	3	2	3	3	142	3	2	3	4
43		1	3	1	1	93	2	2	3	2	143	2	2	3	4
44		1	3	1	1	94	1	1	2	2	144	3	3	4	4
45		1	4	1	1	95	2	2	3	3	145	3	3	4	4
46		1	2	1	1	96	2	2	3	2	146	3	2	3	4
47		1	4	1	1	97	2	2	3	3	147	3	1	3	4
48		1	3	1	1	98	3	2	3	3	148	3	2	3	4
49		2	3	1	1	99	1	1	2	2	149	3	3	3	4
50		1	3	1	1	100	2	2	3	3	150	2	2	3	3

COMMANDE : INPAQN <> interpretation d'une partition par des variables quantitaitives

notes :

b(.,.)-->inertie interclasse de la partition
b(j,.)-->contribution de la variable j a l inertie interclasse de la partition
b(.,k)-->contribution de la classe k a l inertie interclasse de la partition
b(j,k)-->contribution de la classe k a l inertie interclasse de la partition pour la variable j

meme notation avec w -->inertie intraclasse de la partition
meme notation avec t -->inertie de la population

b(.,.) :	602.52
w(.,.) :	78.85
t(.,.) :	681.37

pourcentage d inertie expliquee : 88.43

indices de description des classes de la partition :

t(k) = t(.,k)/t(.,.) --> pourcentage d inertie extraite par la classe k
b(k) = b(.,k)/b(.,.) --> contribution relative de la classe k a l inertie interclasse de la partition
w(k) = w(.,k)/w(.,.) --> contribution relative de la classe k a l inertie intraclasse de la partition
e(k) = b(.,k)/t(.,.) --> pourcentage d inertie expliquee par la classe k

* classe	effectif	t(k)	b(k)	w(k)	e(k) *
1	62	10.9	5.7	50.5	5.1
2	50	53.7	58.2	19.2	51.5
3	38	35.4	36.0	30.3	31.8
					s=88.4

indice decrivant le pouvoir descriminant des variables pour la partition (les valeurs sont triees)

cor(j) = b(j,.)/t(j,.) --> pouvoir discriminant de la variable j
ctr(j) = b(j,.)/b(.,.) --> contribution relative de la variable j a l inertie interclasse

variable	libelle	cor	ctr
LoPe	longueur du petale	94.4	72.7
laPe	largeur du petale	89.8	12.9
LoSe	longueur du sepale	72.2	12.2
laSe	largeur du sepale	45.2	2.1

coordonnees des centres de gravite des variables pour chaque classe de la partition :

variable (effectif)	population (150)	classe 1 (62)	classe 2 (50)	classe 3 (38)
LoSe	5.84	5.90	5.01	6.85
laSe	3.06	2.75	3.43	3.07
LoPe	3.76	4.39	1.46	5.74
laPe	1.20	1.43	.246	2.07

COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables quantitatives

pourcentage d inertie expliquee : 88.43

partition obtenue :

classe numero 1 (effectif = 38)

53	78	101	103	104	105	106	108	109	110	111	112	113	116	117
118	119	121	123	125	126	129	130	131	132	133	135	136	137	138
140	141	142	144	145	146	148	149							

classe numero 2 (effectif = 62)

51	52	54	55	56	57	58	59	60	61	62	63	64	65	66
67	68	69	70	71	72	73	74	75	76	77	79	80	81	82
83	84	85	86	87	88	89	90	91	92	93	94	95	96	97
98	99	100	102	107	114	115	120	122	124	127	128	134	139	143
147	150													

classe numero 3 (effectif = 50)

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
46	47	48	49	50										

coordonnees des centres de gravite des classes de la partition :

variable (effectif)	population (150)	classe 1 (38)	classe 2 (62)	classe 3 (50)
LoSe	5.84	6.85	5.90	5.01
laSe	3.06	3.07	2.75	3.43
LoPe	3.76	5.74	4.39	1.46
laPe	1.20	2.07	1.43	.246

indices de description des classes de la partition :

classe	effectif	t(k)	b(k)	w(k)	e(k)
1	62	10.9	5.7	50.5	5.1
2	50	53.7	58.2	19.2	51.5
3	38	35.4	36.0	30.3	31.8
					s=88.4

indice decrivant le pouvoir descriminant des variables pour la partition (les valeurs sont triees)

variable	libelle	cor	ctr
LoPe	longueur du petale	94.4	72.7
laPe	largeur du petale	89.8	12.9
LoSe	longueur du sepale	72.2	12.2
laSe	largeur du sepale	45.2	2.1

indices decrivant les roles d une variable et d une classe :

	partition		classe 1		classe 2		classe 3	
variable	cor	ctr	cor	ctr	cor	ctr	cor	ctr
LoSe	72.2	12.2	0.2	0.6	34.3	10.0	37.7	17.7
laSe	45.2	2.1	20.9	17.1	24.3	2.0	0.0	0.0
LoPe	94.4	72.7	5.4	72.4	56.8	75.1	32.2	68.9
laPe	89.8	12.9	3.9	9.9	52.5	12.9	33.4	13.3

```
COMMANDE : MNDORD <> nuees dynamiques sur variables qualitatives ordinales utilisant la
                    distance L1


          valeur du critere obtenu :   96


          effectif des classes

          classe   1 :  50

          classe   2 :  45

          classe   3 :  55


          partition en ligne

          classe numero :  1

              1     2     3     4     5     6     7     8     9    10
             11    12    13    14    15    16    17    18    19    20
             21    22    23    24    25    26    27    28    29    30
             31    32    33    34    35    36    37    38    39    40
             41    42    43    44    45    46    47    48    49    50

          classe numero :  2

             71    78   101   102   103   104   105   106   108   109
            110   111   112   113   114   115   116   117   118   119
            121   123   125   126   128   129   130   131   132   133
            136   137   138   139   140   141   142   143   144   145
            146   147   148   149   150

          classe numero :  3

             51    52    53    54    55    56    57    58    59    60
             61    62    63    64    65    66    67    68    69    70
             72    73    74    75    76    77    79    80    81    82
             83    84    85    86    87    88    89    90    91    92
             93    94    95    96    97    98    99   100   107   120
            122   124   127   134   135

          tableau des valeurs ideales

              l  l  l  l
              o  a  o  a
              s  s  p  p

          1   1  2  1  1
          2   2  2  3  3
          3   2  1  2  2


          tableau des ecarts

              l   l   l   l
              o   a   o   a
              s   s   p   p

          1    3  14   0   0
          2   17  14   3   2
          3   15  21   4   3
```

COMMANDE : MNDORD <> nuees dynamiques sur variables qualitatives ordinalesutilisant la distance L1

tableau initial reordonne

Classe 1

	lo s	la s	lo p	la p
1	1	2	1	1
2	1	2	1	1
3	1	2	1	1
4	1	2	1	1
5	1	2	1	1
6	1	3	1	1
7	1	2	1	1
8	1	2	1	1
9	1	2	1	1
10	1	2	1	1
11	1	3	1	1
12	1	2	1	1
13	1	2	1	1
14	1	2	1	1
15	2	3	1	1
16	2	3	1	1
17	1	3	1	1
18	1	2	1	1
19	2	3	1	1
20	1	3	1	1
21	1	2	1	1
22	1	3	1	1
23	1	2	1	1
24	1	2	1	1
25	1	2	1	1
26	1	2	1	1
27	1	2	1	1
28	1	2	1	1
29	1	2	1	1
30	1	2	1	1
31	1	2	1	1
32	1	2	1	1
33	1	3	1	1
34	1	3	1	1
35	1	2	1	1
36	1	2	1	1
37	1	2	1	1
38	1	2	1	1
39	1	2	1	1
40	1	2	1	1
41	1	2	1	1
42	1	1	1	1
43	1	2	1	1
44	1	2	1	1
45	1	3	1	1
46	1	2	1	1
47	1	3	1	1
48	1	2	1	1
49	1	3	1	1
50	1	2	1	1

Classe 2

	lo s	la s	lo p	la p
71	2	2	2	3
78	2	2	3	2
101	2	2	3	3
102	2	1	3	3
103	3	2	3	3
104	2	2	3	3
105	2	2	3	3
106	3	2	3	3
108	3	2	3	3
109	2	1	3	3
110	3	2	3	3
111	2	2	3	3
112	2	1	3	3
113	3	2	3	3
114	2	1	3	3
115	2	1	3	3
116	2	2	3	3
117	2	2	3	3
118	3	3	3	3
119	3	1	3	3
121	3	2	3	3
123	3	1	3	3
125	2	2	3	3
126	3	2	3	3
128	2	2	2	3
129	2	1	3	3
130	3	2	3	2
131	3	1	3	3
132	3	3	3	3
133	2	1	3	3
136	3	2	3	3
137	2	2	3	3
138	2	2	3	3
139	2	2	2	3
140	3	2	3	3
141	2	2	3	3
142	3	2	3	3
143	2	1	3	3
144	3	2	3	3
145	2	2	3	3
146	2	2	3	3
147	2	1	3	3
148	2	2	3	3
149	2	2	3	3
150	2	2	3	3

Classe 3

	lo s	la s	lo p	la p
51	3	2	2	2
52	2	2	2	2
53	3	2	2	2
54	1	1	2	2
55	2	1	2	2
56	2	1	2	2
57	2	2	2	2
58	1	1	2	2
59	2	2	2	2
60	1	1	2	2
61	1	1	2	2
62	2	2	2	2
63	2	1	2	2
64	2	2	2	2
65	2	2	2	2
66	2	2	2	2
67	2	2	2	2
68	2	1	2	2
69	2	1	2	2
70	2	1	2	2
72	2	1	2	2
73	2	1	2	2
74	2	1	2	2
75	2	2	2	2
76	2	2	2	2
77	3	1	2	2
79	2	2	2	2
80	2	1	2	2
81	1	1	2	2
82	1	1	2	2
83	2	1	2	2
84	2	1	3	2
85	1	2	2	2
86	2	2	2	2
87	2	2	2	2
88	2	1	2	2
89	2	2	2	2
90	1	1	2	2
91	1	1	2	2
92	2	2	2	2
93	2	1	2	2
94	1	1	2	2
95	2	1	2	2
96	2	2	2	2
97	2	2	2	2
98	2	2	2	2
99	1	1	2	2
100	2	1	2	2
107	1	1	2	2
120	2	1	3	2
122	2	1	2	3
124	2	1	2	3
127	2	1	2	3
134	2	1	3	2
135	2	1	3	2

CHAPITRE 4

CLASSIFICATION SUR TABLEAU DE CODAGE DISJONCTIF COMPLET

1. INTRODUCTION

Le tableau binaire étudié ici résulte de la transformation, par le codage disjonctif complet, d'un tableau de modalités, croisant individus et variables qualitatives nominales. Nous proposons ici une méthode de classification qui tienne compte de la structure initiale des données.

A partir du tableau de codage, nous pouvons définir un nuage de points représentant l'ensemble à classifier dans l'espace $\mathbf{B}^m$ (où m est le nombre total de modalités). Ces points ont une structure particulière (les composantes codent des modalités), nous les appelons vecteurs binaires de modalités. Il est alors possible d'appliquer l'algorithme MNDBIN. Mais dans ce cas, les noyaux fournis par la méthode ne correspondent pas nécessairement à des codages de modalités. Nous envisageons alors de construire une nouvelle méthode, toujours sur le principe des Nuées Dynamiques, en imposant aux noyaux d'avoir la structure de vecteur binaire de modalités. Dans un premier paragraphe, nous décrivons cette méthode et ses caractéristiques (signification du crtière et interpétation des noyaux). Nous définissons ensuite une méthode équivalente, travaillant directement sur le tableau de modalités et utilisant des noyaux de modalités.

Pour ce type de données, une autre approche est possible. Celle-ci s'appuie sur la transformation des données initiales en un ensemble de "profils" des individus. Ce sont des vecteurs de l'espace $\mathbf{R}^m$, mais ils conservent, dans ce cas précis, une structure très proche de celle de vecteur binaire de modalités. Le nuage des profils représente alors l'ensemble à classifier. Pour ces éléments particuliers, on dispose d'une distance appropriée : la distance du Khi2 (distance qui permet d'atténuer l'importance des modalités à fort effectif).

C'est dans ce cadre que la méthode des Nuées Dynamiques sur variables QuALitatives nominales MNDQAL a été construite (H. Ralambondrainy 1988). Nous présentons rapidement cette méthode et, comme nous le verrons, les noyaux qu'elle fournit ne sont pas de la même forme que les éléments à classifier. Nous envisageons alors de reprendre cette méthode en imposant aux noyaux d'avoir la structure de profil. Après avoir construit et décrit ce nouvel algorithme, nous comparons les différentes méthodes de classification mises en évidence dans ce chapitre.

Enfin, dans un dernier paragraphe, nous présentons les programmes réalisés et intégrés au logiciel d'analyse de données SICLA. Une application sur un tableau de données a été effectuée. Les résultats obtenus par les différentes méthodes sont alors résumés.

2. LA MÉTHODE DE CLASSIFICATION

Nous disposons ici d'un tableau de données binaires, très particulier, puisqu'il résulte de la transformation, par le codage disjonctif complet, d'un tableau de modalités croisant individus et variables qualitatives nominales. A partir de ce tableau, nous définissons un nuage de points représentant, dans l'espace $\mathbf{B}^m$, l'ensemble à classifier (où m est le nombre total de modalités). En fait, les point de ce nuage ont une structure de vecteur de modalités codées. Nous les appelons ici vecteurs binaires de modalités.

La méthode MNDBIN (utilisant la distance en valeurs absolues et des noyaux binaires) peut être appliquée. Cependant, comme nous l'avons vu dans le chapitre 1, paragraphe 5, la règle de la majorité (appliquée aux valeurs 0 et 1 retenues pour le codage) ne fournit pas nécessairement des noyaux correspondant à des codages de modalités. Nous allons donc écrire une nouvelle méthode, en imposant aux noyaux d'être des vecteurs binaires de modalités.

2.1 RAPPEL DES NOTATIONS

Soit $\mathbf{Z}(I,Q)$ le tableau de modalités croisant un ensemble $I=\{1,2,\ldots,n\}$ de n individus et un ensemble $Q=\{1,2,\ldots,p\}$ de p variables qualitatives nominales. On note :

$$\mathbf{Z}(I,Q) = (z_i^q)$$

où z_i^q représente la modalité de la variable q choisie par l'individu i.

A chaque variable q correspond l'ensemble de modalités $J_q=\{1,2,\ldots,m_q\}$. Nous définissons alors l'espace **E** comme le produit $J_1 x J_2 x \ldots x J_p$, que munissons de la distance d_E, égale au nombre de composantes différentes entre les deux points considérés. Son expression sur **E** est la suivante :

$$\forall\ (x, y) \in \mathbf{E}^2 \quad d_E(x,y) = \sum_{q \in Q} \delta^q(x,y)$$

$$\text{où} \quad \delta^q(x,y) = \begin{cases} 1 \text{ si } x^q \neq y^q \\ 0 \text{ sinon} \end{cases}$$

A partir du tableau $\mathbf{Z}(I,Q)$, nous définissons le nuage $\mathbf{N_Z}(I)$, inclus dans l'espace **E**, par :

$$\mathbf{N_Z}(I) = \{z_i,\ i \in I\}$$

$$\text{où} \quad z_i = (z_i^1, z_i^2, \ldots, z_i^p)$$

Soit $\mathbf{X}(I,J)$ le tableau de codage disjonctif complet du tableau de modalités $\mathbf{Z}(I,Q)$. C'est un tableau binaire d'ordre (n,m), où m est le nombre total de modalités. L'ensemble $J=\{1,2,\ldots,m\}$ contient les indices des colonnes de $\mathbf{X}(I,J)$. On note :

$$\mathbf{X}(I,J) = (x_i^j)$$

$$\forall\ q \in Q,\ \forall\ j \in J_q \quad x_i^{q(j)} = \begin{cases} 1 \text{ si } j = z_i^q \\ 0 \text{ sinon} \end{cases}$$

où q(j) est l'indice de J correspondant à la modalité j de la variable q.

Nous définissons l'espace **F** comme la restriction de $\mathbf{B}^m$ aux seuls vecteurs binaires de modalités. Les éléments de **F** résultent du codage des vecteurs de modalités de l'espace **E**. De même, à tout élément de **E** correspond un élément de **F**.

Nous munissons **F** de la distance en valeurs absolues d. Les distances d et d_E sont liées par une relation (propriété démontrée dans le chapitre 1, paragraphe 5). Si X, Y sont deux vecteurs de modalites (appartenant à **E**) de codage x, y (appartenant à **F**), cette relation s'exprime par :

$$d(x,y) = 2\, d_E(X,Y)$$

A partir du tableau **X(I,J)**, nous définissons le nuage **N(I)**, inclus dans **F**, par :

$$\mathbf{N(I)} = \{ x_i, i \in I \} \quad \text{où} \quad x_i = (x_i^1, x_i^2, \ldots, x_i^m)$$

Ainsi, à tout point x_i du nuage **N(I)** correspond un unique point z_i de $\mathbf{N_Z(I)}$ et réciproquement.

2.2 LE PROBLEME

Il s'agit toujours de déterminer une partition de l'ensemble I des individus en K classes, K étant fixé a priori.

Chaque individu i est représenté par un point x_i du nuage **N(I)**. Dans ces conditions, le nuage est inclus dans **F**. Nous allons écrire un nouvel algorithme, toujours basé sur le principe des Nuées Dynamiques, mais respectant le principe d'homogénéité suivant : l'ensemble à classifier étant inclus dans **F**, nous imposons aux noyaux d'appartenir à ce même espace **F**. Nous appelons MNDDIJ cette nouvelle méthode (Méthode des Nuées Dynamiques sur tableau de codage DIsJonctif complet).

Le problème à résoudre est alors le suivant :

trouver une partition $P=(P_1,P_2,\ldots,P_K)$ de I et un ensemble $L=(a_1,a_2,\ldots,a_K)$ de K noyaux de F tels que le critère :

$$W(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j|$$

soit minimum.

2.3 L'ALGORITHME

Il s'agit de construire la fonction d'affectation **f**, telle que **W(f(L),L)** soit minimum, et la fonction de représentation **g**, telle que **W(P,g(P))** soit minimum.

La fonction d'affectation (f)
L'individu i est affecté à la classe P_k dont il est le plus proche du noyau a_k, au sens de la distance en valeurs absolues d.

La fonction de représentation (g)
Il s'agit de déterminer l'ensemble L des K noyaux optimisant le critère **W(P,g(P))**. Pour cela, il suffit de rechercher, toute classe P_k, le noyau a_k appartenant à **F** et minimisant quantité :

$$\sum_{i \in P_k} d(x_i,a_k)$$

Soit A_k appartenant à **E**, le vecteur de modalités codé par a_k. Les distances d_E sur **E** et d sur **F** étant liées par la relation suivante :

$$d(x_i,a_k) = 2\, d_E(z_i,A_k)$$

le problème est alors de trouver le point A_k de **E** minimisant :

$$\sum_{i\in P_k} d_E(z_i,A_k) = \sum_{i\in P_k}\sum_{q\in Q} \delta^q(z_i,A_k)$$

ce qui revient à déterminer, pour tout q, la composante A_k^q minimisant la quantité :

$$\sum_{i\in P_k} \delta^q(z_i,A_k)$$

qui représente le nombre d'individus de P_k n'ayant pas choisi la modalité A_k^q.

Dans le chapitre 1, paragraphe 5, nous avons vu comment résoudre un tel problème. Nous rappelons simplement le résultat : la solution est de choisir pour composantes q du point A_k, la modalité de la variable q qui est majoritaire relative dans la classe considée. Cette composante est donc déterminée de la façon suivante :

$$A_k^q = \text{modalité majoritaire relative de } \{\, z_i,\ i\in P_k \,\}$$

Finalement, le noyau a_k recherché est le transformé du vecteur de modalités dont les composantes sont les modalités majoritaires relatives des variables dans la classe P_k.

L'algorithme ainsi construit fournit un ensemble de noyaux du même type que les éléments à classifier. L'interprétation de ces vecteurs de modalités est simple, chacune des composante est la modalité la plus souvent choisie par les individus de la classes considérée.

Il nous reste à déterminer la signification du critère obtenu à la convergence de l'algorithme MNDDIJ.

2.4 EXPRESSION DU CRITERE À LA CONVERGENCE

A la convergence, on peut exprimer le critère uniquement par rapport à la partition, de sorte que :

$$\mathbf{W}(P) = \mathbf{W}(P,g(P))$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K}\sum_{i\in P_k} d(x_i,a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P) = \sum_{k=1}^{K}\sum_{i\in P_k}\sum_{j\in J} |\, x_i^j - a_k^q \,|$$

Ce critère représente le nombre d'éléments binaires du tableau **X**(I,J) qui sont différents des éléments du tableau correspondant à la situation "idéale" (situation où tous les individus sont identiques aux noyaux des classes auxquelles ils appartiennent).

Une autre expression de $\mathbf{W}(P)$, plus significative, peut être déterminée. Pour cela, considérons les K vecteurs de modalités $(A_1,A_2,\ldots,A_K)$ correspondant aux noyaux fournis par la méthode. La relation entre les distances d et d_E permet d'écrire les équivalences suivantes :

$$\mathbf{W}(P) = \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P) = 2 \sum_{k=1}^{K} \sum_{i \in P_k} d(z_i,A_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P) = 2 \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{q \in Q} \delta^q)(z_i,A_k)$$

$$\text{où} \quad \delta^q(z_i,A_k) = \begin{cases} 1 \text{ si } z_i^q \neq A_k^q \\ 0 \text{ sinon} \end{cases}$$

On peut alors écrire :

$$\mathbf{W}(P) = 2 \sum_{k=1}^{K} \sum_{q \in Q} D_k^q$$

où D_k^q est le nombre de modalités différentes de la modalité majoritaire relative dans la classe P_k pour la variable q.

Le critère représente alors simplement le nombre de fois (au facteur constant 2 près) où la situation obtenue s'écarte de la situation "idéale".

2.5 LA MÉTHODE SUR TABLEAU DE MODALITÉS

Comme cela apparait dans les paragraphes précédents, il est inutile de procéder au codage du tableau de modalités pour pouvoir appliquer la méthode MNDDIJ. Nous pouvons facilement écrire une version de cette méthode utilisant directement le tableau de modalités.

En fait, on peut montrer que cette version de MNDDIJ est équivalente à la méthode des Nuées Dynamiques appliquée au nuage $\mathbf{N}_Z(I)$, et utilisant la distance d_E et des noyaux de modalités appartenant à $\tilde{E}$. Si on note $L_E=(A_1,A_2,\ldots,A_K)$ ces noyaux, le critère optimisé s'écrit :

$$\mathbf{W}(P,L_E) = \sum_{k=1}^{K} \sum_{i \in P_k} d(z_i,A_k)$$

L'ensemble L_E est déterminé à partir de la règle de la majorité relative. La valeur du critère à la convergence représente exactement le nombre de désaccords entre la situation obtenue et la situation "idéale". Si on note $L=(a_1,a_2,\ldots,a_K)$ les codages des noyaux de L_E, on a alors :

$$\mathbf{W}(P,L_E) = \sum_{k=1}^{K} \sum_{i \in P_k} d(z_i,A_k) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i \in P_k} d(x_i,a_k)$$

$$\Leftrightarrow \quad \mathbf{W}(P,L_E) = \frac{1}{2} \mathbf{W}(P,L)$$

Cela montre l'équivalence avec MNDDIJ. C'est cette version plus simple de la méthode que nous retenons et que nous appelons toujours MNDDIJ.

2.6 INDICES D'AIDES À L'INTERPRÉTATION

Nous proposons ici des indices permettant une analyse plus fine de la partition obtenue par la méthode MNDDIJ.

La valeur du critère à la convergence constitue une première indication : elle mesure la qualité de la partition fournie. Cependant, cette valeur (égale au nombre de différences entre situation obtenue et situation "idéale") est à étudier en regard de la taille du tableau initial. A partir de ce critère, nous définissons un premier indice, égal au pourcentage de valeurs identiques entre noyaux et données initiales. L'utilisateur peut ainsi rapidement juger de la qualité de la partition obtenue. Sa définition est la suivante :

$$100 \times \frac{(np - \mathbf{W}(P))}{np}$$

$$\text{où} \quad \mathbf{W}(P) = \sum_{k=1}^{K} \sum_{q \in Q} D_k^q$$

et où D_k^q est le nombre de modalités non majoritaites dans la classe P_k pour la variable q.

D'autre part, pour chaque classe de la partition, nous proposons un indice premettant de juger de sa qualité. Il représente simplement le pourcentage de valeurs identiques au noyau de la classe. Il est alors possible de juger indépendamment de l'homogénéité de chacune des classes obtenues. Pour une classe P_k de cardinal n_k, la définition de l'indice est la suivante :

$$100 \times \frac{n_k p - \sum_{q \in Q} D_k^q}{n_k p}$$

Les noyaux de modalités premettent d'étudier rapidement les particularités de chaque classe de la partition. En complément, pour chaque couple (classe, variable), nous proposons un indice égal au pourcentage d'individus ayant choisis la modalité majoritaire relative. L'utilisateur dispose ainsi de Kxp indicateurs, directement inteprétables. Pour une classe P_k et une variable q, l'expression de l'indice est :

$$100 \times \frac{n_k - D_k^q}{n_k}$$

2.7 EXEMPLE SIMPLE D'APPLICATION

Soit un ensemble de 10 individus, identifiés par les nombres de **1** à **10**, décrit par un ensemble de 5 variables qualitatives nominales, identifiées par les lettres **a** à **e**. Supposons, par exemple, que ces variables possèdent chacune 3 modalités, représentées par les entiers {1,2,3}. Les données initiales sont représentées sous la forme d'un tableau de modalités (figure 1).

Nous appliquons alors la méthode MNDDIJ sur le tableau de modalités en demandant 3 classes. Après plusieurs essais, on aboutit à la finalement à la partition **(A,B,C)=({3,7,9,10},{1,4,5,8},{2,6})**.

Cette partition est représentée (figure 2) en réordonnant simplement les lignes du tableau de manière à respecter les classes obtenues. Nous indiquons également les indices d'aides à l'interprétation suivants :

- figure 3 les effectifs des classes et les indices d'homogénéité,
- figure 4 les noyaux de modalités fournis par la méthode,
- figure 5 la mesure d'homogénéité de chaque couple (classe, variable).

La valeur du critère à la convergence étant de 12, cela indique que sur 50 valeurs initiales, 12 ne sont pas égales à la valeur idéale. Et donc, la situation obtenue permet de représenter correctement 76% des données initiales.

	a	**b**	**c**	**d**	**e**
1	1	2	2	3	2
2	3	2	1	1	1
3	2	3	3	1	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	1	2
7	3	3	2	1	1
8	1	1	1	3	3
9	2	2	2	1	1
10	2	3	3	2	2

figure 1
tableau initial

	a	**b**	**c**	**d**	**e**
3	2	3	3	1	1
7	3	3	2	1	1
9	2	2	2	1	1
10	2	3	3	2	2
1	1	2	2	3	2
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3
2	3	2	1	1	1
6	3	2	1	1	2

figure 2
tableau réordonné

	effectifs	**homogénéité**
A	4	70
B	4	90
C	2	75

figure 3
caractéristiques des classes

	a	**b**	**c**	**d**	**e**
A	2	3	2	1	1
B	1	1	1	3	3
C	3	2	1	1	1

figure 4
les noyaux

	a	**b**	**c**	**d**	**e**
A	75	75	50	75	75
B	100	50	50	100	75
C	100	100	100	100	50

figure 5
homogénéité par classe
et par variable

3. CLASSIFICATION SUR LES PROFILS DES INDIVIDUS

Dans ce paragraphe, nous rappelons tout d'abord la notion de profil associé à un individu. A partir du tableau de codage disjonctif complet, on définit un ensemble de profils, inclus dans l'espace $\mathbf{R}^m$ (où m est le nombre total de modalités). Nous munissons $\mathbf{R}^m$ de la distance du Khi2 (qui convient particulièrement aux profils manipulés ici). Nous pouvons alors appliquer à cet ensemble la méthode de type nuées dynamiques MNDQAL (H. Ralambondrainy 1988) qui utilise cette distance. Pour cette méthode, les noyaux fournis sont les centres de gravité des classes. Chacune d'elle est donc caractérisée par un vecteur de $\mathbf{R}^m$, vecteur ayant une structure différente de celle des éléments de l'ensemble à classifier.

Nous proposons ici une méthode de même type, mais respectant toujours notre principe d'homogénéité : l'ensemble à classifier étant constitué de profils d'individus, on impose aux noyaux de la méthode d'être de même nature. Cette nouvelle approche est présentée en détail dans ce paragraphe et, pour mieux la situer, une comparaison avec les méthodes MNDDIJ et MNDQAL est proposée.

3.1 NOTATIONS ET DÉFINITIONS

Soit $\mathbf{X}(I,J)_{(n,m)}$ le tableau de codage disjonctif complet du tableau de modalités $\mathbf{Z}(I,Q)_{(n,p)}$.

La somme des éléments du tableau binaire ne dépend que du nombre d'individus n et du nombre de variables p puisque :

$$\sum_{i\in I}\sum_{j\in J} x_i^j = \sum_{i\in I}\sum_{q\in Q}\sum_{j\in Jq} x_i^{q(j)} = \sum_{i\in I}\sum_{q\in Q}(1) = np$$

où q(j) est l'indice de J correspondant à la modalité j de q.

On définit également les valeurs suivantes :

$$f_{ij} = \frac{x_i^j}{np} \qquad f_{i.} = \sum_{j\in J} f_{ij} = \frac{1}{np}\sum_{j\in J} x_i^j = \frac{1}{n}$$

$$f_{.j} = \sum_{i\in I} f_{ij} = \frac{1}{n}\sum_{i\in I} x_i^j = \frac{n^j}{np} \qquad \text{où} \quad n^j = \sum_{i\in I} x_i^j$$

Le profil de l'individu i, noté f_i (et appartenant à $\mathbf{R}^m$), est alors défini par :

$$f_i = \left(\frac{f_{i1}}{f_{i.}}, \frac{f_{i2}}{f_{i.}}, \ldots, \frac{f_{im}}{f_{i.}}\right)$$

Si on note x_i le représentant de l'individu i dans l'espace $\mathbf{B}^m$, nous avons :

$$f_i = \frac{1}{p}\ (x_i^1, x_i^2, \ldots, x_i^m)$$

$$f_i = \frac{1}{p}\ x_i$$

D'autre part, une pondération p_i est associée à chaque profil f_i. Elle est définie par :

$$p_i = f_{i.} = \frac{1}{n}$$

A partir du tableau **X**(I,J), nous définissons ainsi un nuage de n profils $\mathbf{N_f}(I)$ par :

$$\mathbf{N_f}(I) = \{ (f_i, \frac{1}{n}), i \in I \}$$

3.2 LA DISTANCE DU KHI2

On munit l'espace $\mathbf{R^m}$ de la distance du Khi2. Pour des vecteurs ayant une structure de profil, la distance apparait sous une forme simplifiée. En effet, considérons deux profils f_i et $f_{i'}$ du nuage, la distance s'exprime alors de la façon suivante :

$$d^2_{\chi 2}(f_i,f_{i'}) = \sum_{j \in J} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

$$\Leftrightarrow \quad d^2_{\chi 2}(f_i,f_{i'}) = \sum_{j \in J} \frac{np}{n^j} \left(\frac{x_i^j}{p} - \frac{x_{i'}^j}{p} \right)^2$$

$$\Leftrightarrow \quad d^2_{\chi 2}(f_i,f_{i'}) = \frac{n}{p} \sum_{j \in J} \frac{1}{n^j} \left(x_i^j - x_{i'}^j \right)^2$$

$$\Leftrightarrow \quad d^2_{\chi 2}(f_i,f_{i'}) = \frac{n}{p} \sum_{j \in J} \frac{1}{n^j} | x_i^j - x_{i'}^j |$$

La distance du Khi2 entre deux profils apparait alors comme une distance en valeurs absolues pondérée.

3.3 LA MÉTHODE

La méthode MNDQAL fournit une partition en K classes, K fixé a priori, de l'ensemble I des individus en optimisant le critère :

$$\mathbf{W_1}(P,L_g) = \sum_{k=1}^{K} \sum_{i \in P_k} f_{i.} d^2_{\chi 2}(f_i,g_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \frac{1}{n} \sum_{j \in J} \frac{np}{n^j} \left(\frac{x_i^j}{p} - g_k^j \right)^2$$

$$\Leftrightarrow \quad \mathbf{W_1}(P,L_g) = p \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \frac{1}{n^j} \left(\frac{x_i^j}{p} - g_k^j \right)^2$$

où $P = (P_1, P_2, \ldots, P_K)$ est une partition sur I en K classes,

et $L_g = (g_1, g_2, \ldots, g_K)$ est l'ensemble des noyaux des classes.

Les fonctions caractérisant l'algorithme MNDQAL sont les suivantes :

- la fonction de d'affectation (**f**) : qui minimise le critère $\mathbf{W}(\mathbf{f}(L_g),L_g)$ en affectant chaque individu à la classe P_k dont il est le plus proche du noyau g_k (au sens de la distance du Khi2);

- la fonction de représentation (**g**) : qui permet de déterminer les K noyaux minimisant le critère $\mathbf{W}(P,\mathbf{g}(P))$. On peut facilement montrer que ces noyaux sont les centres de gravité des classes. Si nous notons $\{g_1,g_2,\ldots,g_K\}$ l'ensemble de ces centres, on a :

$$\forall\, k=1,\ldots,K,\ \forall\, j\in J \qquad g_k^j = \frac{n_k^j}{n_k\, p}$$

où n_k est le nombre d'individus appartenant à la classe P_k,

et $n_k^j = \sum_{i\in P_k} x_i^j$ est le nombre d'individus de P_k ayant choisi la modalité j.

Ces noyaux ne sont ici pas directement interprétable par rapport aux données initiales.

3.4 LA MÉTHODE AVEC CONTRAINTE

Nous nous plaçons ici dans les mêmes conditions que précédemment : l'ensemble à classifier est représenté par le nuage $\mathbf{N}_r(I)$ des profils dans l'espace $\mathbf{R}^m$ muni de la distance du Khi2. Nous imposons ici aux noyaux de la méthode d'avoir une structure de profil. Si **F** est l'espace des vecteurs binaires de modalités, les noyaux devront être de la forme :

$\frac{1}{p}\, a$ où a est un vecteur binaire de modalités de **F**

De cette façon, une variable est caractérisée, dans une classe, par une modalité. Chaque classe est résumée par un vecteur de modalités. Nous appelons MNDDIK cette méthode (Méthodes des Nuées Dynamiques sur tableau de codage DIsjonctif complet utilisant la distance du Khi2 et des noyaux de profils).

3.4.1 La méthode

Le problème à résoudre est le suivant :

*trouver une partition $P=(P_1,P_2,\ldots,P_K)$ et un ensemble $L_a=(a_1,a_2,\ldots,a_K)$ de K noyaux appartenant à **F** tels que le critère :*

$$W_2(P,L_a) = \sum_{k=1}^{K} \sum_{i\in P_k} p_i d^2_{\chi^2}(f_i,\frac{1}{p} a_k)$$

soit minimum.

Les éléments manipulés ayant tous une structure de profil, la distance entre un noyau et le représentant d'un individu s'exprime simplement par :

$$d^2_{\chi^2}(f_i,\frac{1}{p} a_k) = \frac{n}{p} \sum_{j\in J} \frac{1}{n^j} |\, x_i^j - a_k^j \,|$$

L'expression du critère à minimiser est alors la suivante :

$$W_2(P,L_a) = \frac{1}{p} \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \frac{1}{n^j} | x_i^j - a_k^j |$$

Il nous reste à construire les fonctions d'affectation et de représentation caractérisant l'algorithme MNDDIK.

La fonction d'affectation (f)
Elle est définie à partir de la distance du Khi2.

La fonction de représentation (g)
L'étape de représentation consiste à déterminer, pour tout k, le noyau a_k minimisant la quantité :

$$\sum_{i \in P_k} \sum_{j \in J} \frac{1}{n^j} | x_i^j - a_k^j | = \sum_{i \in P_k} \sum_{q \in Q} \sum_{j \in Jq} \frac{1}{n^{q(j)}} | x_i^{q(j)} - a_k^{q(j)} |$$

Il suffit alors de trouver, pour tout q, les m_q composantes $(a_k^{q(j)}, j=1,\ldots,m_q)$ de a_k minimisant :

$$T = \sum_{i \in P_k} \sum_{j \in Jq} \frac{1}{n^{q(j)}} | x_i^{q(j)} - a_k^{q(j)} |$$

or :

$$\sum_{i \in P_k} | x_i^{q(j)} - a_k^{q(j)} | = \sum_{i \in P_k} (x_i^{q(j)}(1-a_k^{q(j)}) + a_k^{q(j)}(1-x_i^{q(j)}))$$

$$\Leftrightarrow \quad \sum_{i \in P_k} | x_i^{q(j)} - a_k^{q(j)} | = n_k^{q(j)} + n_k - 2a_k^{q(j)} n_k^{q(j)}$$

où $\quad n_k = Card(P_k) \quad$ et $\quad n_k^{q(j)} = \sum_{i \in P_k} x_i^{q(j)}$

d'où :

$$T = \sum_{j \in Jq} \frac{1}{n^{q(j)}} (n_k^{q(j)} + n_k - 2a_k^{q(j)} n_k^{q(j)})$$

La contrainte imposée se traduit par le fait qu'une seule composante est égale à 1, de sorte que, si on note r la modalité codée par ce noyau, T devient :

$$T = \sum_{j \in Jq} \frac{n_k^{q(j)}}{n^{q(j)}} + \frac{n_k}{n^{q(r)}} - 2\frac{n_k^{q(r)}}{n^{q(r)}}$$

Le premier terme de cette somme ne dépend pas de la modalité r. La solution est donc de choisir pour composante égale à 1, celle correspondant à la modalité r de la variable q et minimisant la quantité :

$$\frac{n_k - 2n_k^{q(r)}}{n^{q(r)}}$$

dont on donnera une interprétation dans un prochain paragraphe.

3.4.2 Expression du critère à la convergence

A la convergence, le critère $\mathbf{W_2}(P,L_a)$ peut s'exprimer uniquement en fonction de la partition, de sorte que :

$$\mathbf{W_2}(P) = \mathbf{W_2}(P,g(P)) = \frac{1}{p} \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \frac{1}{n^j} | x_i^j - a_k^j |$$

Il apparait que seules les composantes différentes de celles du noyau ont une contribution non nulle au critère. Contrairement à la méthode MNDDIJ (où pour toute différence, entre noyaux et individus, la contibution au critère était identique), la contribution dépend ici des effectifs des modalités. Il reste que le crière $\mathbf{W_2}(P)$ exprime toujours un écart entre la solution obtenue et la solution "idéale".

3.4.3 Interprétation des noyaux

Les noyaux fournis par la méthode MNDDIJ avait une interprétation simple : la caractéristique d'une variable était la modalité la plus souvent choisie par les individus de la classe. Considérons une variable quelconque q de Q, la modalité j choisie pour noyau était celle minimisant :

$$n_k - 2n_k^{q(j)}$$

Ici, la modalité choisie ne permet pas nécessairement une telle interprétation. Les effectifs des modalités jouent un rôle important dans la recherche des noyaux. La modalité j d'une variable q est choisie pour noyau si elle rend minimum la quantité :

$$T = \frac{n_k - 2n_k^{q(j)}}{n^{q(j)}}$$

Une première constatation se dégage de l'expression de T : si une modalité est choisie par au moins la moitié des individus d'une classe, celle-ci est alors prise pour noyau. En effet, dans ce cas, la quantité T à minimiser est strictement négative (si une modalité est majoritaire absolue) ou nulle (si exactement la moitié des individus l'ont choisie) et cela n'est possible que pour cette seule modalité (pour les autres, la quantité T est nécessairement positive).

Si les modalités des variables ont des effectifs voisins sur l'ensemble des individus, la méthode fournit alors des résultats analogues à ceux obtenus par l'algorithme MNDDIJ (les dénominateurs de la quantité T ont une influence moindre lors de la recherche des noyaux).

En pratique, si les effectifs des modalités ne sont pas trop disproportionnés, la modalité retenue est généralement celle qui est la plus souvent choisie par les individus d'une classe. Pour des effectifs assez disproprotionnés, la modalité retenue n'est pas nécessairement la modalité majoritaire relative. La situation suivante peut également se produire : dans certains cas, une modalité n'ayant jamais été choisie par les individus d'une classe peut être prise pour noyau.

Le noyau retenu dépend donc des effectifs sur I des modalités et, principalement, de celles à fort effectif. Ceci montre les limites de la méthode et de l'interprétation que l'on peut faire à partir de ces noyaux. Ci-après, nous donnons des exemples simples, illustrant ces différents cas de figure.

Exemple 1
Soit un ensemble **I** de 100 individus décrit par une variable qualitative nominale à 3 modalités dans {**1**,**2**,**3**}. Nous nous intéressons uniquement au comportement de cette variable dans une classe **A** de 20 individus. Dans la figure 1, on représente les effectifs de ces modalités sur **A** et sur **I**. Dans la figure 2, nous indiquons, pour chaque modalité, la valeur de la quantité (**T**) à minimiser. Nous constatons que le noyau retenu a pour valeur **2**, alors que la modalité **1** est celle de plus fort effectif sur **A**.

	1	**2**	**3**
A	9	8	3
I	20	60	20

figure 1
effectifs par modalité

	1	**2**	**3**
T	$\frac{1}{10}$	$\frac{1}{15}$	$\frac{11}{10}$

figure 2
valeurs de la quantité
à minimiser

Exemple 2
Soit un ensemble **I** de 100 individus décrit par une variable qualitative à 4 modalités dans {**1**,**2**,**3**,**4**}. Considérons encore une classe **A** de 20 individus. Dans la figure 3, on représente les effectifs de ces 4 modalités sur **I** et sur **A**. Nous indiquons également, figure 4, les valeurs de la quantité (**T**) à minimiser. Nous constatons que le noyau retenu prend la valeur **2**, alors que la modalité **2** n'est jamais choisie par les individus de la classe.

	1	**2**	**3**	**4**
A	7	0	7	6
I	13	60	13	14

figure 3
effectifs par modalité

	1	**2**	**3**	**4**
T	$\frac{6}{13}$	$\frac{1}{3}$	$\frac{6}{13}$	$\frac{4}{7}$

figure 4
valeurs de la quantité
à minimiser

3.4.4 Version de l'algorithme utilisant le tableau de modalités

Encore une fois, il est inutile de procéder au codage du tableau de modalités pour pouvoir appliquer la méthode MNDDIK. Il est possible d'exprimer la distance du Khi2 uniquement par rapport aux effectifs des modalités des variables. Si on note s(j) le nombre d'individus de I ayant choisi la modalité j, la distance entre profils et noyaux s'exprime alors de la façon suivante :

$$d^2_{\chi 2}(f_i,\frac{1}{p}a_k) = \frac{n}{p}\sum_{q\in Q}\sum_{j\in Jq}\frac{1}{n^{q(j)}}\,|\,x_i^{q(j)} - a_k^{q(j)}\,|$$

$$\Leftrightarrow \quad d^2_{\chi 2}(f_i,\frac{1}{p}a_k) = \frac{n}{p}\sum_{q\in Q}\left(\frac{1}{s(z_i^q)} + \frac{1}{s(A_k^q)}\right)\delta^q(z_i,A_k)$$

$$\text{où} \quad \delta^q(z_i,A_k) = \begin{cases} 1 \text{ si } z_i^q \neq A_k^q \\ 0 \text{ sinon} \end{cases}$$

Le critère de la méthode peut être réécrit à partir de la nouvelle expression de la distance.

3.4.5 Exemple simple d'application

Soit un ensemble de 10 individus, identifiés par les nombres de **1** à **10**, et décrit par un ensemble de 5 variables qualitatives nominales, identifiées par les lettres **a** à **e**. Ces variables sont à modalités dans {1,2,3}. Les données initiales sont représentées sous la forme d'un tableau (figure 1).

Après plusieurs essais effectués à partir de différents points initiaux, MNDDIK aboutit finalement à la partition (**A**, **B**, **C**) = ({**3**, **10**},{**1**, **2**, **6**, **7**, **9**}, {**4**, **5**, **8**}).

Celle-ci est représentée (figure 2) en réordonnant simplement les lignes de manière à respecter les classes obtenues. Nous présentons également les mesures d'homogénéité des classes et des variables par classes. Celles-ci sont simplement définies à partir du critère optmisé par la méthode et exprimées sous la forme de pourcentages (figure 3 et 5). Les noyaux obtenus à la convergence sont indiqués en figure 3. Pour ce tableau élémentaire, les composantes des noyaux sont aussi les modalités majoritaires relatives. Enfin, le critère (défini à partir de la distance du Khi2) indique un pourcentage d'accord de 74.5% entre données initiales et noyaux.

	a	**b**	**c**	**d**	**e**
1	1	2	2	3	2
2	3	2	1	1	1
3	2	3	3	1	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	1	2
7	3	3	2	1	1
8	1	1	1	3	3
9	2	2	2	1	1
10	2	3	3	2	2

figure 1
tableau initial

	a	**b**	**c**	**d**	**e**
3	2	3	3	1	1
10	2	3	3	2	2
1	1	2	2	3	2
2	3	2	1	1	1
6	3	2	1	1	2
7	3	3	2	1	1
9	2	2	2	1	1
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3

figure 2
tableau réordonné

	effectifs	**homogénéité**
A	2	76
B	5	65
C	3	87

figure 3
caractéristiques des classes

	a	**b**	**c**	**d**	**e**
A	2	3	3	1	1
B	3	2	2	1	1
C	1	1	1	3	3

figure 4
les noyaux

	a	**b**	**c**	**d**	**e**
A	100	100	100	25	46
B	62	75	60	78	56
C	100	74	67	100	100

figure 5
homogénéité par classe

4. COMPARAISON DES MÉTHODES

Dans ce chapitre, nous avons proposé deux nouvelles méthodes de classification appelées : MNDDIJ et MNDDIK. Nous avons également rappelé le principe de la méthode MNDQAL. Nous illustrons ici les différences entre ces trois algorithmes.

4.1 COMPARAISON DES CRITERES

La méthode MNDDIJ, de par l'utilisation de la distance en valeurs absolues et de noyaux de modalités, se distingue clairement des deux autres méthodes.

Pour comparer les critères associés aux méthodes MNDQAL et MNDDIK, on dispose, sur l'espace vectoriel $\mathbf{R}^m$, de la relation de Huyghens permettant d'exprimer l'inertie d'une classe par rapport à un point quelconque.

Soient g, le centre de gravité du nuage $\mathbf{N_f}(I)$, et P, une partition de I en K classes. Soient également $L_g=(g_1,g_2,\ldots,g_K)$, l'ensemble des centres de gravité des classes, et $L_a=(a_1,a_2,\ldots,a_K)$, l'ensemble des noyaux fournis par la méthode MNDDIK.

La relation de Huyghens permet alors d'écrire :

$$\sum_{k=1}^{K} \sum_{i\in P_k} p_i\, d^2_{\chi 2}(f_i,g_k) = \sum_{k=1}^{K} \sum_{i\in P_k} p_i\, d^2_{\chi 2}(f_i,\frac{1}{p}a_k) + \sum_{k=1}^{K} p_k\, d^2_{\chi 2}(\frac{1}{p}a_k,g_k)$$

$$\text{où} \quad p_k = \sum_{i\in P_k} p_i$$

d'où la relation liant les critères $\mathbf{W_1}(P,L_g)$ (méthode MNDQAL) et $\mathbf{W_2}(P,L_a)$ (méthode MNDDIK) :

$$\mathbf{W_1}(P,L_g) = \mathbf{W_2}(P,L_a) + \sum_{k=1}^{K} p_k\, d^2_{\chi 2}(\frac{1}{p}a_k,g_k)$$

qui exprime bien la différence entre les deux problèmes d'optimisation.

4.2 APPLICATION ILLUSTRATIVE

Nous illustrons les différences entre ces trois méthodes en les appliquant sur un même tableau de données.

Considérons un ensemble de 10 individus, identifiés par les nombres **1** à **10**. Ceux-ci sont décrits par un ensemble de 55 variables qualitatives nominales, identifiées par les lettres **a** à **e**. Supposons que toutes ces variables soient à modalités dans {1,2,3}.

Les réponses des individus sont représentées sous la forme d'un tableau de modalités (figure 1 de la page suivante).

Nous appliquons les trois méthodes MNDDIJ, MNDQAL et MNDDIK, en demandant 3 classes. A chaque fois, on effectue plusieurs essais et on ne retient que la meilleure solution. Les méthodes aboutissent alors à 3 partitions différentes.

Celles-ci sont représentées en réordonnant les lignes du tableau initial en respectant les partitions obtenues par ces méthode (figure 2 pour la solution obtenue par MNDDIJ, figure 3 pour celle obtenue par MNDQAL et figure 5 pour MNDDIK).

Les critères obtenus à la convergence et exprimés en pourcentage d'accord sont de 74% pour la méthode MNDDIJ (il y a 13 différences entre données initiales et noyaux), et 70.5% pour MNDDIK. Enfin, l'inertie expliquée par la partition fournie par MNDQAL est de 46%.

	a	b	c	d	e
1	1	2	2	3	2
2	3	2	1	2	1
3	2	3	3	2	1
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	3	2
7	3	3	2	1	1
8	1	1	1	3	3
9	1	2	2	1	1
10	2	2	3	3	2

figure 1
tableau initial

	a	b	c	d	e
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3
3	2	3	3	2	1
7	3	3	2	1	1
9	1	2	2	1	1
1	1	2	2	3	2
2	3	2	1	2	1
6	3	2	1	3	2
10	2	2	3	3	2

figure 2
MNDDIJ

	a	b	c	d	e
3	2	3	3	2	1
10	2	2	3	3	2
1	1	2	2	3	2
4	1	1	2	3	3
5	1	2	1	3	3
6	3	2	1	3	2
8	1	1	1	3	3
2	3	2	1	2	1
7	3	3	2	1	1
9	1	2	2	1	1

figure 3
MNDQAL

	a	b	c	d	e
2	3	2	1	2	1
3	2	3	3	2	1
7	3	3	2	1	1
1	1	2	2	3	2
6	3	2	1	3	2
9	1	2	2	1	1
10	2	2	3	3	2
4	1	1	2	3	3
5	1	2	1	3	3
8	1	1	1	3	3

figure 4
MNDDIK

Pour chaque méthode, nous avons également effectué des essais en utilisant, comme points initiaux, les partitions fournies par les deux autres méthodes. Les résultats obtenus ne sont pas meilleurs que ceux présentés ci-dessus.

5. PROGRAMMES ET APPLICATION

5.1 PROGRAMMES

Comme les programmes précédents, MNDDIJ et MNDDIK ont été construits à l'aide de la bibliothèque de procédures du logiciel SICLA. Ils deviennent ainsi des commandes de ce logiciel. Encore une fois, l'utilisateur a la possibilité de définir les différents paramètres initiaux.

5.2 APPLICATION

Nous proposons d'appliquer ces méthodes au tableau de données croisant 30 felins et 15 variables qualitatives nominales (le tableau figure en page 117). L'une de ces variables correspond au classement par les zoologistes (**zool**). Elle possède 4 modalités relative aux genres **panthera (pant)**, **neofelis (nefe)**, **felis (feli)** et **acinonyx (acin)**. Nous avons ici la distribution suivante :

- **pant** effectif 5, identifiés par les chiffres de **1** à **5**,
- **nefe** effectif 1, identifié par **6**,
- **feli** effectif 23, identifié par **7,9** à **30**,
- **acin** effectif 1, identifié par **8**.

Nous choisissons de supprimer les deux felins du type **nefe** et **acin**. La conséquence est que la variable à 2 modalités associée au caractère des griffes est également supprimée (sur les 30 félins, seul celui du genre **acin** a des griffes retractibles, tous les autres ne les ont pas).

Nous allons donc appliquer les méthodes présentées dans ce chapitre sur le tableau croisant 13 variables (nous n'utilisons pas la variable **zool**) et 28 félins. Nous essayons de caractériser les deux genres **pant** et **feli** en recherchant des partitions.

La commande DESQAL de SICLA permet de détailler les effectifs des modalités de toutes les variables. Les résultats sont indiqués à partir de la page 118, on pourra s'y référer pour connaître les variables manipulées ici.

Une première constatation est que la variable **lary** discrimine parfaitement la partition naturelle. A deux félins près, il en va de même pour la variable **dent**. On a donc les caractèristiques suivantes :

pant : absence de l'os hyaoïde (modalité **lar2** de **lary**),
canines peut développées (modalité **den2** de **dent**).

feli : absence de l'os hyaoïde (modalité **lar1** de **lary**),
canines peut développées (modalité **den1** de **dent**).

Considérons que la valeur représentative d'une variable dans une classe soit sa modalité majoritaire relative (approche MNDDIJ). Chaque classe est alors résumée par un vecteur de modalités. Dans ces conditions, on compte **91** désaccords entre données initiales et noyaux, soit un pourcentage d'accords de **76.7%**.

Application de MNDQAL

Cette méthode fournit une partition en 2 classes expliquant **23.067%** de l'inertie initiale des données (les résultats sont en page 121).

Les effectifs des deux classes obtenues sont les suivants :

Classe 1 : 4 félins du genre **pant**,

Classe 2 : 23 félins du genre **feli** et le **pant 5**.

Par ordre d'importance, les variables caractérisant le mieux les deux classes sont les suivantes :

Classe 1 : absence de l'os hyaoïde (modalité **lar2** de **lary**),
canines peut développées (modalité **den2** de **dent**),
taille supérieure à 70 cm (modalité **tai3** de **tail**),
poids supérieur à 80 kg (modalité **pod3** de **poid**).

Classe 2 : présence de l'os hyaoïde (modalité **lar1** de **lary**).
canines très développées (modalité **den1** de **dent**),
petite proie (modalité **pro3** de **proi**).

Les variables caractérisant la première classe sont toutes parfaitement homogènes. Dans la seconde, on compte une différence pour la variable **lary**, et 3 pour chacune des variables **dent** et **proi**, soit un total de 7 provenant des trois félins identifiés par **5,7** et **11**. Les trois variables cités ci-dessus sont d'ailleurs celles discriminant le mieux la partition. Remarquons que, au total, le nombre de désaccords entre données et modalités majoritaires relatives est de **95**, d'où un pourcentage d'accords de **75.64%** (inférieur à la partition naturelle).

Application de MNDDIJ

Cette méthode fournit une partition et un critère associé égal à **87** (page 122). Celui-ci représente le nombre de différences entre données initiales et noyaux de modalités, il indique aussi un pourcentage d'accords de **77,7%**.

La partition fournie est la suivante :

Classe 1 : 21 félins du genre **feli**,

Classe 2 : 5 félins du genre **pant** et les **feli 7** et **11**.

A la lecture du tableau des noyaux et des homogénéités (page 122), on constate que la variable **dent** discrimine parfaitement la partition : les félins de la première classe sont tous caractérisés par des canines peu développées (modalité 1 de la variable **dent**), ceux de la deuxième classe par des canines très développées (modalité 2 de **dent**). A un niveau moindre, on peut également citer les variables **lary** et **proie**. On retrouve alors les trois mêmes variables discriminantes que pour l'application précédente (mais dans un ordre d'importance différent). Au niveau de la constitution des classes, MNDDIJ et MNDQAL diffèrent des trois éléménts **5,7** et **11**. Signalons également que la méthode MNDDIK fournit ici les mêmes résultats que MNDDIJ.

Autres applications

Nous avons poursuivi l'étude en recherchant des partitions en 4 classes. La méthode MNDQAL (page 121) fournit, entre autres, une classe (numéro 3) ne contenant que les **pant 1** et **2**, une autre (numéro 1) contenant les trois autres **pant** et le **feli 7**. La méthode MNDDIJ (page 123) isole mieux le genre **pant** puisque 4 d'entre eux sont regoupés en une seule classe (numéro 4), le **pant 5** figurant avec les **feli 7, 9, 10** et **11** dans la classe 3. Enfin, contrairement au cas précédent, la méthode MNDDIK aboutit, à la même partition que la méthode MNDQAL.

LE TABLEAU DE DONNEES :

	type	poil	graif	comp	orei	lary	tail	poid	long	queu	dent	prob	arbr	chas	zool
1	1	1	2	1	1	2	3	3	3	2	2	1	1	2	1
2	3	1	2	3	1	2	3	3	3	2	2	1	1	1	1
3	2	1	2	3	1	2	3	3	2	1	2	1	2	1	1
4	2	1	2	3	1	2	3	3	2	2	2	2	2	1	1
5	2	2	2	1	1	2	2	2	2	3	2	2	2	1	1
6	2	1	1	1	1	1	3	2	2	3	1	2	1	2	4
7	1	1	2	2	1	1	2	3	2	3	2	2	2	1	3
8	4	1	2	3	1	2	2	2	2	3	2	3	2	1	2
9	2	1	2	1	2	1	2	2	2	1	1	3	2	2	3
10	2	1	2	2	1	1	2	2	2	2	1	3	2	1	3
11	2	2	2	2	2	1	2	2	2	1	2	2	2	1	3
12	1	1	2	2	2	1	2	2	1	1	1	3	2	2	3
13	2	1	2	2	1	1	1	1	2	2	1	3	1	1	3
14	1	1	2	2	1	1	1	2	2	3	1	3	2	1	3
15	1	2	2	3	2	1	1	2	1	2	1	3	2	1	3
16	1	1	2	3	1	1	1	1	1	1	1	3	2	1	3
17	2	1	2	3	1	1	1	1	1	2	1	3	2	1	3
18	1	2	2	2	1	1	1	1	1	2	1	3	1	1	3
19	3	1	2	3	1	1	1	1	1	2	1	3	2	2	3
20	1	1	2	2	2	1	1	1	1	1	1	3	2	1	3
21	2	1	2	3	1	1	1	1	1	2	1	3	2	1	3
22	2	1	2	2	1	1	1	1	1	2	1	3	2	1	3
23	1	2	2	3	1	1	1	1	1	1	1	3	2	1	3
24	1	1	2	3	1	1	1	1	1	2	1	3	2	1	3
25	2	1	2	2	1	1	1	1	1	1	1	3	2	2	3
26	1	2	2	3	1	1	1	1	1	1	1	3	2	1	3
27	4	1	2	3	1	1	1	1	1	3	1	3	2	1	3
28	2	1	2	3	1	1	1	1	1	2	1	3	2	1	3
29	1	1	2	3	1	1	1	1	1	2	1	3	2	1	3
30	2	2	2	3	1	1	1	1	2	2	1	2	2	1	3

```
COMMANDE : DESQAL <> description de variables qualitatives


      1. variable  : type  aspect du pelage
         --------

            typ1 : sans tache

            typ2 : tachete

            typ3 : raye

            typ4 : marbre<

 ***********************
 * modl * nbre *   %  *
 ***********************
 * typ1 *   12 *   43 *   *******************************************
 * typ2 *   13 *   46 *   ********************************************
 * typ3 *    2 *    7 *   *******
 * typ4 *    1 *    4 *   ****
 ***********************



      2. variable  : poil  fourrure
         --------

            poi1 : poils ras

            poi2 : poils longs

 ***********************
 * modl * nbre *   %  *
 ***********************
 * poi1 *   21 *   75 *   ***************************************************************************
 * poi2 *    7 *   25 *   *************************
 ***********************



      3. variable  : comp  comportement predateur
         --------

            com1 : diurne

            com2 : diurne et nocturne

            com3 : nocturne

 ***********************
 * modl * nbre *   %  *
 ***********************
 * com1 *    3 *   11 *   ***********
 * com2 *   10 *   36 *   ************************************
 * com3 *   15 *   54 *   ****************************************************
 ***********************



      4. variable  : orei  forme des oreilles
         --------

            ore1 : rondes et arrondies

            ore2 : en pointe(type caracal)

 ***********************
 * modl * nbre *   %  *
 ***********************
 * ore1 *   23 *   82 *   ***************************************************************************
 * ore2 *    5 *   18 *   ******************
 ***********************



      5. variable  : lary  presence ou non de l'os hyaoide
         --------

            lar1 : presence

            lar2 : absence

 ***********************
 * modl * nbre *   %  *
 ***********************
 * lar1 *   23 *   82 *   ***************************************************************************
 * lar2 *    5 *   18 *   ******************
 ***********************
```

```
COMMANDE : DESQAL <> description de variables qualitatives
```

6. variable : tail taille du garrot

tail : taille < 50 cm

tai2 : 50 cm < taille < 70 cm

tai3 : taille > 70 cm

modl	nbre	%	
tail	18	64	**********
tai2	6	21	**********
tai3	4	14	**********

7. variable : poid poids de l'animal

pod1 : poids < 10 kg

pod2 : 10 kg < poids < 80 kg

pod3 : poids > 80 kg

modl	nbre	%	
pod1	16	57	**********
pod2	7	25	**********
pod3	5	18	**********

8. variable : long longueur du corps

lon1 : longueur < 80 cm

lon2 : 80 cm < longueur < 150 cm

lon3 : longueur > 150 cm

modl	nbre	%	
lon1	16	57	**********
lon2	10	36	**********
lon3	2	7	**********

9. variable : queu longueur de la queue % a celle du corps

que1 : petite

que2 : moyenne

que3 : longue

modl	nbre	%	
que1	9	32	**********
que2	15	54	**********
que3	4	14	**********

```
COMMANDE : DESQAL <> description de variables qualitatives


     10. variable  : dent  canines developpees
         --------

               den1 : tres developpees

               den2 : peu developpees

 ***********************
 * modl * nbre *   %  *
 ***********************
 * den1 *    21 *   75 *   ***************************************************************************
 * den2 *     7 *   25 *   *************************
 ***********************



     11. variable  : proi  type de proie
         --------

               pro1 : grosse

               pro2 : grosse ou petite

               pro3 : petite

 ***********************
 * modl * nbre *   %  *
 ***********************
 * pro1 *     3 *   11 *   ***********
 * pro2 *     5 *   18 *   ******************
 * pro3 *    20 *   71 *   ***************************************************************************
 ***********************



     12. variable  : arbr  monte ou non aux arbres
         --------

               arb1 : monte aux arbres

               arb2 : ne monte pas

 ***********************
 * modl * nbre *   %  *
 ***********************
 * arb1 *     4 *   14 *   **************
 * arb2 *    2  *       *  ***************************************************************************
 ***********************



     13. variable  : chas  chasse a courre ou a l'affut
         --------

               cha1 : oui

               cha2 : non

 ***********************
 * modl * nbre *   %  *
 ***********************
 * cha1 *    23 *   82 *   ***************************************************************************
 * cha2 *     5 *   18 *   ******************
 ***********************
```

```
COMMANDE : MNDQAL <> methode des nuees dynamiques sur variables qualitatives


   partition en 2 classes :

   classe  1  :  effectif   4   14%,  variance  0.45678E-01
      1     2     3     4

   classe  2  :  effectif  24   86%,  variance  0.44019E-01
      5     7     9    10    11    12    13    14    15    16    17    18    19    20    21
     22    23    24    25    26    27    28    29    30


   inertie totale   :  1.6154

   inertie intra    :  1.2428

   inertie inter    :  0.37262

   --->pourcentage d'inertie explique  :  23.067         (Inertie inter/Inertie totale)



COMMANDE : MNDQAN <> methode des nuees dynamiques sur variables qualitatives

   partition en 4 classes :

   classe  1  :  effectif   4   14%,  variance  0.44124E-01
      3     4     5     7

   classe  2  :  effectif   4   14%,  variance  0.30763E-01
      9    10    11    12

   classe  3  :  effectif   2    7%,  variance  0.23593E-01
      1     2

   classe  4  :  effectif  18   64%,  variance  0.29905E-01
     13    14    15    16    17    18    19    20    21    22
     23    24    25    26    27    28    29    30


   inertie totale   :  1.6154

   inertie intra    :  0.88502

   inertie inter    :  0.73036

   --->pourcentage d'inertie explique  :  45.213 %         (Inertie inter/Inertie totale)
```

```
COMMANDE : MNDDIJ <> nuees dynamiques sur tableau disjonctif complet


          valeur du critere obtenu :   87


          effectif des classes :

          classe   1 :  21
          classe   2 :   7


          partition :

          classe numero :  1

              9    10    12    13    14    15    16    17    18    19
             20    21    22    23    24    25    26    27    28    29
             30

          classe numero :  2

              1     2     3     4     5     7    11


          tableau des valeurs ideales

              t  p  c  o  l  t  p  l  q  d  p  a  c
              y  o  o  r  a  a  o  o  u  e  r  r  h
              p  i  m  e  r  i  i  n  e  n  o  b  a
              e  l  p  i  y  l  d  g  u  t  i  r  s

          1   1  1  3  1  1  1  1  1  2  1  3  2  1
          2   2  1  3  1  2  3  3  2  2  2  2  2  1


          homogeneite par classe

              t   p   c   o   l   t   p   l   q   d   p   a   c
              y   o   o   r   a   a   o   o   u   e   r   r   h
              p   i   m   e   r   i   i   n   e   n   o   b   a
              e   l   p   i   y   l   d   g   u   t   i   r   s

        1    48  76  57  81 100  86  76  76  57 100  95  90  81
        2    57  71  43  86  71  57  71  71  43 100  57  71  86


      tableau initial reordone

      t p c o l t p l q d p a c
      y o o r a a o o u e r r h
      p i m e r i i n e n o b a
      e l p i y l d g u t i r s
 ---------------------------------
   9  2 1 1 2 1 2 2 2 1 1 3 2 2
  10  2 1 2 1 1 2 2 2 2 1 3 2 1
  12  1 1 2 2 1 2 2 1 1 1 3 2 2
  13  2 1 2 1 1 1 1 2 2 1 3 1 1
  14  1 1 2 1 1 1 2 2 3 1 3 2 1
  15  1 2 3 2 1 1 2 1 2 1 3 2 1
  16  1 1 3 1 1 1 1 1 1 1 3 2 1
  17  2 1 3 1 1 1 1 1 2 1 3 2 1
  18  1 2 2 1 1 1 1 1 2 1 3 1 1
  19  3 1 3 1 1 1 1 1 2 1 3 2 2
  20  1 1 2 2 1 1 1 1 1 1 3 2 1
  21  2 1 3 1 1 1 1 1 2 1 3 2 1
  22  2 1 2 1 1 1 1 1 2 1 3 2 1
  23  1 2 3 1 1 1 1 1 1 1 3 2 1
  24  1 1 3 1 1 1 1 1 2 1 3 2 1
  25  2 1 2 1 1 1 1 1 1 1 3 2 2
  26  1 2 3 1 1 1 1 1 1 1 3 2 1
  27  4 1 3 1 1 1 1 1 3 1 3 2 1
  28  2 1 3 1 1 1 1 1 2 1 3 2 1
  29  1 1 3 1 1 1 1 1 2 1 3 2 1
  30  2 2 3 1 1 1 1 2 2 1 2 2 1
 ---------------------------------
   1  1 1 1 1 2 3 3 3 2 2 1 1 2
   2  3 1 3 1 2 3 3 3 2 2 1 1 1
   3  2 1 3 1 2 3 3 2 1 2 1 2 1
   4  2 1 3 1 2 3 3 2 2 2 2 2 1
   5  2 2 1 1 2 2 2 2 3 2 2 2 1
   7  1 1 2 1 1 2 3 2 3 2 2 2 1
  11  2 2 2 2 1 2 2 2 1 2 2 2 1
 ---------------------------------
```

```
COMMANDE : MNDDIJ <> nuees dynamiques sur tableau disjonctif complet

        valeur du critere obtenu :   62

        effectif des classes :

        classe   1  :   8

        classe   2  :  11

        classe   3  :   5

        classe   4  :   4

        partition :

        classe numero :  1

            12    14    16    18    20    23    25    26

        classe numero :  2

            13    15    17    19    21    22    24    27    28    29    30

        classe numero :  3

             5     7     9    10    11

        classe numero :  4

             1     2     3     4

        tableau des valeurs ideales

          type  poil  comp  orei  lary  tail  poid  long  quen  dent  prob  arbr  chas
        1    1     1     2     1     1     1     1     1     1     1     3     2     1
        2    2     1     3     1     1     1     1     1     2     1     3     2     1
        3    2     1     2     1     1     2     2     2     1     2     2     2     1
        4    2     1     3     1     2     3     3     2     2     2     1     1     1

        homogeneite par classe

          type  poil  comp  orei  lary  tail  poid  long  quen  dent  prob  arbr  chas
        1   88    63    63    75   100    88    75    88    75   100   100    88    75
        2   55    82    82    91   100   100    91    82    91   100    91    91    91
        3   80    60    60    60    80   100    80   100    40    60    60   100    80
        4   50   100    75   100   100   100   100    50    75   100    75    50    75

        tableau initial reordone

          type poil comp orei lary tail poid long quen dent prob arbr chas
        ------------------------------------------------------------------
          12    1    1    2    2    1    2    2    1    1    1    3    2    2
          14    1    1    2    1    1    1    2    2    3    1    3    2    1
          16    1    1    3    1    1    1    1    1    1    1    3    2    1
          18    1    2    2    1    1    1    1    1    2    1    3    1    1
          20    1    1    2    2    1    1    1    1    1    1    3    2    1
          23    1    2    3    1    1    1    1    1    1    1    3    2    1
          25    2    1    2    1    1    1    1    1    1    1    3    2    2
          26    1    2    3    1    1    1    1    1    1    1    3    2    1
        ------------------------------------------------------------------
          13    2    1    2    1    1    1    1    2    2    1    3    1    1
          15    1    2    3    2    1    1    2    1    2    1    3    2    1
          17    2    1    3    1    1    1    1    1    2    1    3    2    1
          19    3    1    3    1    1    1    1    1    2    1    3    2    2
          21    2    1    3    1    1    1    1    1    2    1    3    2    1
          22    2    1    2    1    1    1    1    1    2    1    3    2    1
          24    1    1    3    1    1    1    1    1    2    1    3    2    1
          27    4    1    3    1    1    1    1    1    3    1    3    2    1
          28    2    1    3    1    1    1    1    1    2    1    3    2    1
          29    1    1    3    1    1    1    1    1    2    1    3    2    1
          30    2    2    3    1    1    1    1    2    2    1    2    2    1
        ------------------------------------------------------------------
           5    2    2    1    1    2    2    2    2    3    2    2    2    1
           7    1    1    2    1    1    2    3    2    3    2    2    2    1
           9    2    1    1    2    1    2    2    2    1    1    3    2    2
          10    2    1    2    1    1    2    2    2    2    1    3    2    1
          11    2    2    2    2    1    2    2    2    1    2    2    2    1
        ------------------------------------------------------------------
           1    1    1    1    1    2    3    3    3    2    2    1    1    2
           2    3    1    3    1    2    3    3    3    2    2    1    1    1
           3    2    1    3    1    2    3    3    3    2    1    2    1    2
           4    2    1    3    1    2    3    3    2    2    2    2    2    1
        ------------------------------------------------------------------
```

CHAPITRE 5

INERTIE SUR L'ESPACE BINAIRE ET APPLICATION À LA CLASSIFICATION

1. INTRODUCTION

Dans la cadre de l'espace $\mathbf{R}^p$ muni de la distance euclidienne usuelle, on dispose des notions de centre de gravité et d'inertie d'un nuage de points. Celles-ci permettent notamment d'établir une relation très utilisée : la relation de Huyghens.

Dans le chapitre 1, paragraphe 3.3.4, nous avons proposé des notions analogues dans le cadre de l'espace binaire $\mathbf{B}^p$ muni de la distance en valeurs absolues ou distance L_1. Nous avons défini la notion de centre médian (qui joue le rôle du centre de gravité) et d'inertie associée à cette distance. Cependant, celles-ci n'ont pas les mêmes propriétés que les notions analogues définies sur $\mathbf{R}^p$: d'une part la propriété de conservation du centre médian n'est pas vérifiée, d'autre part une relation de type Huyghens fait défaut.

Dans ce chapitre, nous envisageons de reprendre toutes ces notions, sous de nouvelles hypothèses, de manière à obtenir sur l'espace $\mathbf{B}^p$ des propriétés analogues à celles établies sur l'espace $\mathbf{R}^p$. Ensuite, sous cette nouvelle approche, nous reprenons les méthodes de classification de données binaires. Celles-ci apparaissent alors sous une forme plus habituelle.

Dans un premier paragraphe, nous présentons les nouvelles notions de centre médian et d'inertie définie à partir de la distance en valeurs absolues. Celles-ci reposent sur la définition d'une pondération particulière associée à un élément de l'espace. Ici, nous définissons plus précisemment des vecteurs de pondérations. A partir de cette nouvelle hypothèse, nous démontrons d'une part que la propriété de conservation du centre médian est maintenant vérifiée, d'autre part une relation de type Huyghens et, plus généralement, une relation de décomposition de l'inertie (du type habituel : inertie totale = inertie intraclasse + inertie interclasse).

Dans un second paragraphe, nous reprenons la méthode de classification sur tableau de variables binaires MNDBIN (chapitre 2). En utilisant les notions précédentes, le critère optimisé apparait sous la forme d'une inertie intraclasse. La méthode MNDBIN devient alors analogue à la méthode des Nuées Dynamiques utilisant la distance euclidienne usuelle et les centres de gravité des classes comme noyaux (méthode appelée MNDQAN dans le chapitre 2). Nous proposons également une extension, plus générale, de cet algorithme. Enfin, à partir de ces nouvelles notions, nous définissons des indices permettant d'analyser les résultats de la classification. Ceux-ci sont définis de façon analogue à ceux proposés dans le cas quantitatif.

Dans un troisième paragraphe, nous reprenons la méthode de classification croisée d'un tableau binaire CROBIN (G. Govaert 1983). Son objectif est de trouver un tableau résumé le plus proche possible du tableau initial. Ainsi défini, cet algorithme ne suit pas le principe général de la classification croisée, principe qui repose sur la conservation d'une mesure d'information associée à un tableau (G. Govaert 1983). Dans ce travail, nous montrons que l'on peut définir une mesure d'information, associée à un tableau de

données binaires, et replacer ainsi cet algorithme dans le contexte général de la classification croisée.

Enfin, dans un dernier paragraphe, nous proposons un algorithme de classification ascendante hiérarchique pour les données binaires. A partir de la nouvelle notion d'inertie, nous définissons des indices d'agrégation (analogues à ceux définis dans le cas quantitatif à partir de l'inertie habituelle). Certains peuvent être utilisés pour indicer une hiérarchie, d'autres, comme nous le verrons, posent des problèmes d'inversion.

2. INERTIE SUR L'ESPACE BINAIRE

2.1 LA PREMIERE APPROCHE

Soit **A** un nuage de points inclus dans $\mathbf{B}^p$. Supposons que chaque élément x de **A** soit muni d'une pondération $\alpha(x)$ appartenant à $\mathbf{R}^+$.

La proximité entre deux points est mesurée par la distance L_1 notée d :

$$\forall\, x \text{ et } y \in \mathbf{A} \qquad d(x,y) = \sum_{j=1}^{p} |x^j - y^j|$$

Nous résumons ici les notions déjà définies dans le chapitre 1, paragraphe 3.

Le centre médian
Le centre médian a de du nuage **A** de $\mathbf{B}^p$ est défini par :

$$\forall\, j=1,\ldots,p \qquad a^j = \text{médiane binaire } \{\, (x^j, \alpha(x)),\ x \in \mathbf{A} \,\}$$

L'inertie
L'inertie du nuage **A** par rapport à un point b de $\mathbf{B}^p$ est définie par :

$$\mathfrak{I}_b(\mathbf{A}) = \sum_{x \in A} \alpha(x)\, d(x,b) = \sum_{x \in A} \alpha(x) \sum_{j=1}^{p} |x^j - b^j|$$

Par définition, le centre médian est le point d'inertie minimale.

Comme nous l'avons vu dans dans le chapitre 1, la propriété de conservation du centre médian n'est pas vérifiée : si on remplace une partie de **A** par son centre médian, on change le centre médian de l'ensemble. En fait, le problème que nous avions rencontré, était celui de la pondération à associer au centre médian.

2.2 EXTENSION DE CETTE APPROCHE

2.2.1 La nouvelle pondération

Dans le paragraphe précédent, nous avons muni chaque élément x de la pondération $\alpha(x)$ appartenant à $\mathbf{R}^+$. Nous proposons ici d'utiliser comme pondération, non plus un simple réel, mais un vecteur de pondérations.

Définition
A tout point x de $\mathbf{B}^p$, nous associons le vecteur de pondérations $\alpha(x)$ défini par :

$$\alpha(x) = (\alpha^1(x), \alpha^2(x), \ldots, \alpha^p(x))$$

$$où \quad \forall j=1, \ldots, p \quad \alpha^j(x) \in \boldsymbol{R}^+$$

A partir de cette nouvelle hypothèse, nous allons définir une extension de la notion de centre médian et d'inertie.

2.2.2 Centre médian

La différence avec la situation précédente est que chaque composante d'un point est munie de sa propre pondération. Le centre médian de **A** est, maintenant, tout point a minimisant la quantité :

$$\sum_{x \in A} \sum_{j=1}^{p} \alpha^j(x) | x^j - a^j |$$

Propriété
Le centre médian a du nuage A est défini par :

$$\forall j=1, \ldots, p \quad a^j = médiane\ binaire\ \{ (x^j, \alpha^j(x)), x \in A \}$$

Preuve
Il s'agit de rechercher, pour tout j, la composante a^j minimisant :

$$\sum_{x \in A} \alpha^j(x) | x^j - a^j |$$

La solution évidente est de choisir pour a^j la médiane de $\{(x^j, \alpha^j(x)), x \in \mathbf{A}\}$.

Remarquons que, si toutes les pondérations des éléments de **A** sont égales, on retrouve la notion précédente.

2.2.3 Pondération associée au centre médian

Jusqu'à présent, nous n'avons associé aucune pondération particulière au centre médian d'un nuage **A** de $\mathbf{B}^p$. La seule possibilité envisagée a été de munir ce point de la somme des pondérations des éléments de **A**. Cette approche ne permet pas de vérifier la propriété de conservation du centre médian.

Nous définissons ici une nouvelle pondération permettant de remédier à cela. C'est à partir de cette nouvelle notion que sera définie l'inertie sur l'espace binaire.

Définition

Nous associons au centre médian a de A le vecteur de pondérations $\alpha(a)$ défini par :

$$\forall j=1,\ldots,p \quad \alpha^j(a) = | n_A^j(1) - n_A^j(0) |$$

où $n_A^j(1) = \sum_{x \in A} \alpha^j(x)\, x^j$ *somme des pondérations des valeurs 1,*

et $n_A^j(0) = \sum_{x \in A} \alpha^j(x)\, (1 - x^j)$ *somme des pondérations des valeurs 0.*

Lorsque toutes les pondérations sont égales à 1, la composante j du vecteur de pondérations exprime la différence entre le nombre d'éléments de **A** ayant une composante j à 1 et le nombre d'éléments de **A** ayant une composante j à 0.

2.2.4 Propriété de conservation du centre médian

Soit la partition $(\mathbf{A_1},\mathbf{A_2},\ldots,\mathbf{A_K})$ de **A**. Pour tout k, le centre médian a_k de $\mathbf{A_k}$ est défini par :

$$\forall\ j=1,\ldots,p \quad a_k^j = \text{médiane binaire de } \{\ (x^j, \alpha^j(x)),\ x \in \mathbf{A_k}\ \}$$

$$\forall\ j=1,\ldots,p \quad \alpha^j(a_k) = | n_{A_k}^j(1) - n_{A_k}^j(0) |$$

$$\text{où} \quad n_{A_k}^j(1) = \sum_{x \in A_k} \alpha^j(x)\, x^j$$

$$\text{et} \quad n_{A_k}^j(0) = \sum_{x \in A_k} \alpha^j(x)(1 - x^j)$$

Propriété

Si $(\mathbf{A_1},\mathbf{A_2},\ldots,\mathbf{A_K})$ est une partition de A, alors le centre médian de l'ensemble des centres médians $\{a_1,a_2,\ldots,a_K\}$ (munis de leur vecteur de pondérations respectif) des parties $(\mathbf{A_1},\mathbf{A_2},\ldots,\mathbf{A_K})$ est le centre médian de A.

Preuve

Notons b le centre médian de l'ensemble $\mathbf{B}=\{(a_k, \alpha(a_k)), k=1,\ldots,K\}$ des centres médians des parties $\mathbf{A_1},\mathbf{A_2},\ldots,\mathbf{A_K}$.

Le point b est défini par :

$$\forall\ j=1,\ldots,p \quad b^j = \text{médiane binaire de } \{\ (a_k^j, \alpha^j(a_k)),\ a_k \in \mathbf{B}\ \}$$

$$\forall\ j=1,\ldots,p \quad \alpha^j(b) = | n_B^j(1) - n_B^j(0) |$$

$$\text{où} \quad n_B^j(1) = \sum_{k=1}^{K} \alpha^j(a_k)\, a_k^j$$

$$\text{et} \quad n_B^j(0) = \sum_{k=1}^{K} \alpha^j(a_k)(1 - a_k^j)$$

Soit a le centre médian de **A**. Il s'agit de démontrer que b = a et $\alpha(b) = \alpha(a)$.

Pour tout j, on a :

$$\alpha^j(b) = | n_B^j(1) - n_B^j(0) |$$

$$\Leftrightarrow \quad \alpha^j(b) = | \sum_{k=1}^{K} \alpha^j(a_k)\, a_k^j \; - \sum_{k=1}^{K} \alpha^j(a_k)(1 - a_k^j) |$$

$$\Leftrightarrow \quad \alpha^j(b) = | \sum_{k=1}^{K} (n_{A_k}^j(1) - n_{A_k}^j(0))\, a_k^j \; - \sum_{k=1}^{K} (n_{A_k}^j(0) - n_{A_k}^j(1))(1 - a_k^j) |$$

$$\Leftrightarrow \quad \alpha^j(b) = | \sum_{k=1}^{K} n_{A_k}^j(1) \; - \sum_{k=1}^{K} n_{A_k}^j(0) |$$

$$\Leftrightarrow \quad \alpha^j(b) = | n_A^j(1) \; - n_A^j(0) |$$

$$\Leftrightarrow \quad \alpha^j(b) = \alpha^j(a)$$

De ce résultat, on peut déduire l'égalité entre les vecteurs a et b puisque :

$$b^j = 1 \Leftrightarrow n_B^j(1) > n_B^j(0) \Leftrightarrow n_A^j(1) > n_A^j(0)$$

$$b^j = 1 \Leftrightarrow a^j = 1$$

et :

$$b^j = 0 \Leftrightarrow n_B^j(0) \geq n_B^j(1) \Leftrightarrow n_A^j(0) > n_A^j(1)$$

$$b^j = 0 \Leftrightarrow a^j = 0$$

La propriété de conservation du centre médian est ainsi démontrée.

2.3 INERTIE SUR L'ESPACE BINAIRE

En utilisant les notions du paragraphe 2.2, nous allons construire une nouvelle inertie sur l'espace $\mathbf{B}^p$.

Définitions
*Nous définissons l'inertie du nuage **A** par rapport à un point b de $\mathbf{B}^p$ par :*

$$\mathcal{I}_b(A) = \sum_{x \in A} \sum_{j=1}^{p} \alpha^j(x) \,/\, x^j - b^j/$$

*Nous appelons inertie du nuage **A** l'inertie de **A** par rapport à son centre médian a. On note simplement :*

$$\mathcal{I}(A) = \mathcal{I}_a(A)$$

Remarquons que si toutes les pondérations des éléments de **A** sont égales, on retrouve là aussi la première définition.

2.4 PSEUDO-THÉOREME DE HUYGHENS

Dans le chapitre 1 paragraphe 3.3.4, nous avons déjà tenté de construire une relation de type Huyghens. La notion d'inertie utilisée alors ne permettait pas d'aboutir à un tel résultat.

A partir de la nouvelle notion d'inertie sur l'espace $\mathbf{B}^p$, cela devient possible.

Pseudo-théorème de Huyghens
*Soient **A** un nuage de $\mathbf{B}^p$ et a son centre médian. L'inertie de **A** et l'inertie de **A** par rapport à un point quelconque sont liées par une relation qui à la forme habituelle de la relation de Huyghens :*

$$\forall b \in \mathbf{B}^p \quad \mathfrak{I}_b(A) = \mathfrak{I}(A) + \mathfrak{I}_b(\{a\})$$

$$où \quad \mathfrak{I}_b(\{a\}) = \sum_{j=1}^{p} \alpha^j(a) \, / \, b^j - a^j /$$

Preuve
Par définition :

$$\mathfrak{I}_b(\mathbf{A}) = \sum_{x \in A} \sum_{j=1}^{p} \alpha^j(x) | x^j - b^j | = \sum_{j=1}^{p} b^j n_A^j(0) + (1-b^j) n_A^j(1)$$

$$\mathfrak{I}(\mathbf{A}) = \sum_{x \in A} \sum_{j=1}^{p} \alpha^j(x) | x^j - a^j | = \sum_{j=1}^{p} a^j n_A^j(0) + (1-a^j) n_A^j(1)$$

D'où :

$$\mathfrak{I}_b(\mathbf{A}) - \mathfrak{I}(\mathbf{A}) = \sum_{j=1}^{p} (b^j - a^j)(n_A^j(0) - n_A^j(1))$$

Cette différence est toujours positive ou nulle puisque, pour tout j, on a :

- si $a^j = 0$ alors $n_A^j(0) - n_A^j(1) \geq 0$ et $b^j(n_A^j(0) - n_A^j(1)) \geq 0$
- si $a^j = 1$ alors $n_A^j(0) - n_A^j(1) \leq 0$ et $(b^j - 1)(n_A^j(0) - n_A^j(1)) \geq 0$

On peut alors écrire :

$$\mathfrak{I}_b(\mathbf{A}) - \mathfrak{I}(\mathbf{A}) = \sum_{j=1}^{p} | n_A^j(0) - n_A^j(1) | | b^j - a^j |$$

$$\Leftrightarrow \quad \mathfrak{I}_b(\mathbf{A}) - \mathfrak{I}(\mathbf{A}) = \sum_{j=1}^{p} \alpha^j(a) | b^j - a^j |$$

$$\Leftrightarrow \quad \mathfrak{I}_b(\mathbf{A}) - \mathfrak{I}(\mathbf{A}) = \mathfrak{I}_b(\{a\})$$

Le pseudo-théorème de Huyghens est donc démontré.

Propriété
Le centre médian du nuage est le point d'inertie minimale.

Ce résultat est une simple conséquence du pseudo-théorème de Huyghens.

2.5 RELATION DE DÉCOMPOSITION DE L'INERTIE

Dans le chapitre 2 paragraphe 3.5, nous avons établi une relation voisine de la relation de décomposition de l'inertie. Cependant, nous n'avons pu l'exprimer sous la forme habituelle : inertie totale = inertie intraclasse + inertie interclasse.

La nouvelle notion d'inertie et le pseudo-théorème de Huyghens vont nous permettre de construire une telle relation.

Propriété
Soient b un point de B^p et $(A_1,A_2,\ldots,A_K)$ une partition de A. Nous avons la relation de décomposition suivante :

$$\mathfrak{I}_b(A) = \sum_{k=1}^{K} \mathfrak{I}(A_k) + \sum_{k=1}^{K} \mathfrak{I}_b(\{a_k\})$$

où, pour tout k, a_k représente le centre médian de la classe A_k.

Si on exprime cette relation par rapport au centre médian du nuage A, on obtient la relation habituelle :

inertie totale = inertie intraclasse + inertie interclasse

Preuve
Le pseudo-théorème de Huyghens permet de démontrer facilement la relation de décomposition. En effet, on a :

$$\mathfrak{I}_b(A) = \sum_{k=1}^{K} \sum_{x \in A_k} \sum_{j=1}^{p} \alpha^j(x) | x^j - b^j |$$

$$\Leftrightarrow \quad \mathfrak{I}_b(A) = \sum_{k=1}^{K} \mathfrak{I}_b(A_k)$$

$$\Leftrightarrow \quad \mathfrak{I}_b(A) = \sum_{k=1}^{K} \mathfrak{I}(A_k) + \mathfrak{I}_b(\{a_k\})$$

$$\Leftrightarrow \quad \mathfrak{I}_b(A) = \sum_{k=1}^{K} \mathfrak{I}(A_k) + \sum_{k=1}^{K} \mathfrak{I}_b(\{a_k\})$$

Lorsque b est le centre médian du nuage A, on retrouve les notions d'inertie intraclasse et d'inertie interclasse :

$$\mathfrak{I}(A) = \sum_{k=1}^{K} \mathfrak{I}(A_k) + \mathfrak{I}(\{(a_k, \alpha(a_k)), k = 1, \ldots, K\})$$

La notion d'inertie binaire va nous permettre de donner une nouvelle interprétation de la méthode de classification MNDBIN et de la méthode de classification croisée CROBIN.

C'est ce que nous montrons dans les paragraphes suivants.

3. LA MÉTHODE DE CLASSIFICATION MNDBIN

Nous conservons ici la notation $N(I)=\{x_i, i \in I\}$ pour le nuage de points de B^p, défini à partir d'un tableau X(I,J) croisant un ensemble I de n individus et un ensemble J de p variables binaires.

3.1 LA MÉTHODE MNDBIN

C'est un algorithme de type Nuées Dynamiques appliqué au nuage N(I) de B^p et utilisant la distance en valeurs absolues d. Il fournit une solution locale au problème d'optimisation suivant :

trouver une partition $P=(P_1,\ldots,P_K)$ de I et un ensemble $L=(a_1,\ldots,a_K)$ de K noyaux de B^p tels que le critère :

$$W(P,L) = \sum_{k=1}^{K} \sum_{i \in I} d(x_i,a_k)$$

soit minimum.

La résolution (chapitre 2) a mis en évidence une fonction d'affectation habituelle (elle repose sur la distance d) et une fonction de représentation particulière : les noyaux sont les centres médians des classes. Ainsi, chaque classe est décrite par un vecteur binaire facilement interprétable.

3.2 LA NOUVELLE APPROCHE

Considérons le nuage N(I) dont les points sont munis de vecteurs de pondérations toutes égales à 1 :

$$\forall\, i \in I, \forall\, j \in J \quad \alpha^j(x_i) = 1$$

La relation de décomposition de l'inertie permet d'écrire :

$$\Im(N(I)) = W(P,L) + \Im(\{(a_k, \alpha(a_k)), k=1, \ldots, K\})$$

Le critère W(P,L) apparait alors comme l'inertie intraclasse de la partition. De plus, le problème de la minimisation de W(P,L) est équivalent au problème de la maximisation de l'inertie interclasse du nuage des centres médians des classes. Ainsi présentée, cette méthode est analogue à la méthode des Nuées Dynamiques utilisant la distance euclidienne et les centres de gravités des classes comme noyaux.

Si on note a la centre médian du nuage N(I), l'inertie interclasse de la partition s'écrit :

$$\Im(\{(a_k, \alpha(a_k)), k=1, \ldots, K\}) = \sum_{k=1}^{K} \sum_{j \in J} \alpha^j(a_k)\, |\, a^j_k - a^j|$$

De par sa définition, cette inertie permet de donner un nouveau sens à la méthode. Il apparait que seules les variables ayant une médiane locale (valeur 0 ou 1 majoritaire dans la classe) différente de la médiane globale (valeur 0 ou 1 majoritaire sur I) ont une contribution non nulle à l'inertie interclasse. Plus précisemment, l'inertie interclasse s'écrit comme la somme des pondérations des médianes locales qui sont différentes des médianes globales. La méthode maximisant cette somme, une partition sera d'autant meilleure qu'elle contiendra des classes regroupant, le mieux possible, les valeurs

minoritaires des variables initiales (afin d'obtenir des pondérations associées les plus grandes possibles).

Remarques
L'introduction de l'inertie binaire ne change pas l'algorithme MNDBIN (les pondérations des centres médians des classes n'interviennent pas). Simplement, cette inertie binaire nous permet d'interpréter la méthode MNDBIN en terme d'optimimastion d'une inertie intraclasse et d'une inertie interclasse.

Lorsque toutes les pondérations initiales sont égales, la méthode MNDBIN fournit une solution au problème de classification posé. Cependant, celui-ci ne permet pas de prendre en compte des pondérations initiales différentes. Dans la suite, nous proposons une généralisation de l'algorithme MNDBIN, de sorte que celui-ci puisse être appliqué à un tableau de données binaires muni d'un tableau de pondérations quelconques.

Une utilisation intéressante de l'algortihme MNDBIN généralisé sera vue dans le paragraphe traitant de la méthode de classification croisée CROBIN.

3.3 GÉNÉRALISATION DE LA MÉTHODE

Nous proposons ici d'étendre le champ d'application de la méthode MNDBIN à un nuage de points munis de vecteurs de pondérations quelconques.

Nous supposons donc que chaque élément du tableau initial est muni de sa propre pondération, de sorte que :

$\mathbf{X}(I,J) = (x_i^j)$ le tableau binaire,

$\alpha(I,J) = (\alpha_i^j)$ le tableau des pondérations,

$\forall\, i \in I, \forall\, j \in J$ α_i^j est la pondération associée à x_i^j.

Chaque élément x_i du nuage $\mathbf{N}(I)$ est alors muni du vecteur de pondérations :

$$\alpha(x_i) = (\alpha_i^1, \alpha_i^2, \ldots, \alpha_i^p)$$

Pour mesurer correctement la proximité entre individus et noyaux, nous définissons une nouvelle distance tenant compte de ces pondérations. Il s'agit d'une forme adaptée de la distance en valeurs absolues. On la note **D** et elle est définie par :

$$\forall\, k=1,\ldots,K, \forall\, i \in I \quad \mathbf{D}(x_i,a_k) = \sum_{j \in J} \alpha_i^j |\, x_i^j - a_k^j|$$

où a_k représente le noyau de la classe k.

Le critère associé à la méthode s'exprime alors par :

$$\mathbf{W}(P,L) = \sum_{k=1}^{K} \sum_{i \in P_k} \mathbf{D}(x_i,a_k) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \alpha_i^j |\, x_i^j - a_k^j|$$

Il nous reste à définir les fonctions d'affectation et de représentation associées.

La fonction d'affectation repose sur la mesure **D** : un individu i est affecté à la classe k dont il est le plus proche du noyau au sens de la mesure **D**.

La fonction de représentation met encore en évidence le même type de noyaux; ce sont les centres médians des classes définis à partir de la notion de médiane mais en tenant compte des pondérations :

$$\forall\, k=1,\ldots,K,\ \forall\, j\in J \quad a_k^j = \text{médiane binaire de l'ensemble } \{\ (x_i^j, \alpha_i^j),\ i\in I\ \}$$

Remarques
Si on reprend la notion d'inertie binaire, la critère **W**(P,L) optimisé par l'algorithme généralisé apparait alors comme une inertie intraclasse. La relation de décomposition démontrée dans le cas général s'exprime ici de la façon suivante ici :

$$\Im(\mathbf{N}(I)) = \mathbf{W}(P,L) + \Im(\{(a_k, \alpha(a_k)), k=1, \ldots, K\})$$

Cela montre l'équivalence avec le problème de la maximisation de l'inertie interclasse.

Ce résultat va permettre une approche différente de la méthode de classification croisée CROBIN. En effet, après avoir donné une interprétation en terme d'inertie du critère optimisé par cette méthode, il nous sera possible de démontrer que les deux algorithmes intermédiaires utilisés par CROBIN correspondent en fait à un seul et même algorithme : l'algorithme MNDBIN généralisé. Cette étude fera l'objet d'un prochain

Enfin, et c'est ce que nous développons dans le paragraphe suivant, l'introduction de la notion d'inertie binaire va nous permettre de définir des indices de description d'une partition.

3.4 INDICES DE DESCRIPTION D'UNE PARTITION

Dans le chapitre 2, nous avons déjà proposé des indices de description d'une partition obtenue par l'algorithme MNDBIN. Ceux-ci fournissent des indications sur la qualité globale de la partition, sur la qualité de chacune des classes et, enfin, sur l'homogénéité de chaque variable dans chaque classe. Ils sont exprimés en terme de pourcentage d'accords entre données initiales et valeurs "idéales". A partir de la nouvelle approche, nous pouvons définr de nouveaux indices, exprimant différemment la qualité des résultats fournis par la méthode.

Lorsque les variables sont quantitatives, on se place dans l'espace $\mathbf{R}^p$ muni de la distance euclidienne usuelle. On peut alors appliqué la méthode des Nuées Dynamiques utilisant cette distance et minimisant l'inertie intraclasse de la partition. A partir de la relation de décomposition de l'inertie, on définit alors des indices de description d'une partition (G. Celeux et al. 1989).

Nous disposons ici d'une méthode et d'une relation analogue sur l'espace binaire. Nous allons utiliser cette relation pour définir des indices de description d'une partition obtenue par l'algorithme MNDBIN généralisé.

3.4.1 Notations

Considérons toujours le nuage **N**(I) de centre médian a, une partition P de I en K classes et l'ensemble L des K noyaux correspondants.

Dans la suite, pour définir les indices, nous utiliserons les notations suivantes :

- relation de décomposition de l'inertie : $T = W + B$

- pour tout j et tout k on a : $T^j = W^j + B^j \quad T_k = W_k + B_k \quad T_k^j = W_k^j + B_k^j$

- inertie totale : $T = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \alpha_i^j |x_i^j - a^j| = \sum_{j \in J} T^j = \sum_{k=1}^{K} T_k = \sum_{k=1}^{K} \sum_{j \in J} T_k^j$

- inertie intraclasse : $W = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \alpha_i^j |x_i^j - a_k^j| = \sum_{j \in J} W^j = \sum_{k=1}^{K} W_k = \sum_{k=1}^{K} \sum_{j \in J} W_k^j$

- inertie interclasse : $B = \sum_{k=1}^{K} \sum_{j \in J} \alpha^j(a_k) |a_k^j - a^j| = \sum_{j \in J} B^j = \sum_{k=1}^{K} B_k = \sum_{k=1}^{K} \sum_{j \in J} B_k^j$

3.4.2 Définition des indices

D'une manière générale, on retrouve, dans le cas binaire, le même type d'interprétation de ces indices que dans le cas quantitatif. En outre, ceux-ci peuvent apporter des indications supplémentaires, vu la particularité des données.

Dans la suite, nous utiliserons les appellations suivantes :

- Lorsque le noyau d'une classe est identique au centre médian du nuage, nous dirons que la classe a un comportement "médian".

- De la même manière, lorsque la médiane locale d'une variable (dans une classe) est égale à sa médiane globale (sur tous les individus), nous dirons que la classe en question a un comportement "médian" vis à vis de cette variable. Nous dirons également que la variable a un comportement local "médian".

- Lorsque toutes les composantes de tous les vecteurs de pondérations sont égales à 1, la médiane globale d'une variable est la valeur 0 ou 1 qui est majoritaire sur l'ensemble des individus. Lorsque les pondérations sont quelconques, la médiane global est la valeur 0 ou 1 de plus forte pondération associée. Dans les deux cas, nous parlerons toujours de "valeur majoritaire" et

Indice général
Il représente la part d'inertie conservée ou expliquée en assimilant les individus aux K centres médians :

$$\mathbf{R} = \frac{B}{T}$$

Un indice **R** proche de 1 signifie que le nuage des individus se résume au nuage des K centres médians.

L'inertie T mesure la présence de valeurs minoritaires dans le tableau initial. L'inertie B est un indicateur de la répartition de ces valeurs dans les classes de la partition. Un indice **R** égal à 1 indique alors que les valeurs minoritaires sont parfaitement regroupées dans certaines classes.

Contribution des variables
On définit ici un indice équivalent à **R** et représentant la part d'inertie de la variable j prise en compte par la partition :

$$\mathbf{COR}(j) = \frac{B^j}{T^j}$$

Lorsque les données sont quantitatives, cet indice est une mesure du pouvoir discriminant de la variable par rapport à la partition. L'indice **R** représente alors le pouvoir discriminant moyen des variables vis à vis de la partition. Ces remarques restent

Un indice égal à 1 indique d'une part que la partition contient des classes regroupant parfaitement les valeurs minoritaires de la variable j, d'autre part que cette variable à un comportement homogène dans chacune des classes.

L'indice est donc d'autant plus élevé qu'il existe des classes regroupant le mieux possible les valeurs minoritaires de la variable. Par contre, si la variable à un comportement médian (homogène ou non) dans toutes les classes de la partition, elle n'a aucun pouvoir discriminant ($\mathbf{COR}(j) = 0$).

On définit également la contribution relative de la variable j à l'inertie interclasse de la partition :

$$\mathbf{CTR}(j) = \frac{B^j}{B} \quad \text{et} \quad \sum_{j \in J} \mathbf{CTR}(j) = 1$$

Celui-ci est complémentaire au premier et varie généralement dans le même sens. Comme dans le cas quantitatif, le fait suivant peut se produire : dans certains cas, on a **COR**(j) faible (resp. fort) et **CTR**(j) fort (resp. faible), cela signifie que la variable j malgré une forte (resp. faible) contribution à l'inertie interclasse est peu (resp. très)

Description des classes
Pour chaque classe k de la partition, on définit les indices suivants :

$\mathbf{T}(k) = \frac{T_k}{T}$: représente le poucentage d'inertie extrait de la classe k. Il représente également la part de valeurs minoritaires extraite par la classe k.

$\mathbf{B}(k) = \frac{B_k}{B}$: la contribution relative de la classe k à l'inertie interclasse. Cet indice permet de situer la position de la classe par rapport au centre médian global. Plus il est élevé, plus la classe occupe une position excentrée par rapport au centre médian global.

$\mathbf{W}(k) = \frac{W_k}{W}$: la contribution relative de la classe k à l'inertie intraclasse. Il indique la concentration de la classe indépendamment de sa position. Toutefois, il est à étudier en regard du cardinal de la classe : un indice fort est d'autant plus significatif que le cardinal est élevé. L'indice **W**(k) représente la part de désaccords extraite de la classe k.

Description des classes par variables
Pour chaque variable j et chaque classe k, on définit un indice représentant le pouvoir discriminant de la variable j pris en compte par la classe k :

$$\mathbf{COR}(j,k) = \frac{B_k^j}{T^j} \quad \text{et} \quad \sum_{k=1}^{K} \mathbf{COR}(j,k) = \mathbf{COR}(j)$$

Un indice **COR**(j,k) élevé indique d'une part que la variable à un comportement local non médian, d'autre part qu'elle est homogène dans la classe k. L'indice montre alors la répartition, entre les classes, du pouvoir discriminant de la variable. Encore une fois, si la variable ne se distingue pas de son comportement médian, l'indice est nul et n'offre aucune contribution au pouvoir discriminant (même si la variable est parfaitement homogène dans la classe k).

On définit également un indice complémentaire au précédent et représentant la contribution relative de la variable j et de la classe k à l'inertie interclasse :

$$\mathbf{CTR}(j,k) = \frac{B_k^j}{B_k} \quad \text{et} \quad \sum_{k=1}^{K} \mathbf{CTR}(j,k) = \mathbf{CTR}(j)$$

Il permet de voir, parmi les variables ayant un comportement local non médian, celles qui caractérisent le plus chaque classe. De nouveau, les variables ayant un comportement local identique au comportement global n'ont aucune contribution.

3.4.3 Remarque sur les indices

Les indices font apparaitre le résultat suivant : les classes et les variables ne sont significatives (contributions non nulles) que dans la mesure où elles se distinguent du comportement médian. Si cette différence de comportement existe, une classe ou une variable apporte alors une information permettant de caractériser certains individus par rapport à tout l'ensemble.

Considérons, par exemple, un nuage dont le centre médian est l'origine (toutes les composantes sont égales à 0) et sur lequel on applique la méthode. La partition obtenue sera d'autant meilleure qu'elle regroupe le mieux possible les valeurs 1 du tableau initial. Une variable a un pouvoir discriminant (non nul) si, dans une classe au moins, sa médiane est égale à 1.

3.4.4 Exemple d'application

Soit le tableau croisant l'ensemble **I={1,2,3,4,5,6,7,8,9,10}** des individus et l'ensemble **J={a,b,c,d,e}** des variables (figure 1 de la page suivante). On suppose, en outre, que toutes les pondérations initiales sont égales à 1.

On applique la méthode MNDBIN en demandant 3 classes. Après plusieurs essais, on obtient la partition suivante :

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) = \{(\mathbf{1}, \mathbf{6}, \mathbf{7}, \mathbf{9}\}, \{\mathbf{3}, \mathbf{4}, \mathbf{8}\}, \{\mathbf{2}, \mathbf{5}, \mathbf{10}\})$$

que l'on représente (figure 2) en réordonnant les lignes du tableau initial de manière à respecter les classes obtenues. Dans la figure 3, on représente le centre médian du nuage

des individus ainsi que les noyaux des classes. Dans la figure 4, on indique les vecteurs de pondérations de ces centres.

L'indice général **R** montre que la partition permet d'expliquer 67% de l'inertie initiale. Dans la figure 5, nous indiquons les contributions des variables (indice **COR**(j)), ainsi que les contributions des variables par classes (indice **COR**(j,k)). Pour permettre une interprétation plus rapide, ces indices sont exprimés en pourcentage. On constate alors que la variable **a** n'a aucun pouvoir discriminant (**COR(a)**=0) alors que la variable **c** discrimine parfaitement la partition (**COR(c)**=100). Dans la figure 6, on indique les indices de description des classes. Des trois classes, une même part de valeurs minoritaires est extraite (**T(A)=T(B)=T(C)**=33%). Les classes **B** et **C** sont celles ayant les plus fortes contributions relatives à l'inertie interclasse. La classe **A** est celle occupant la position la moins excentrée.

	a	b	c	d	e
1	1	1	0	0	0
2	1	0	0	0	1
3	1	0	1	1	0
4	1	0	1	1	1
5	1	0	0	0	1
6	1	1	0	1	0
7	0	1	0	1	0
8	1	0	1	1	1
9	1	0	0	1	0
10	1	0	0	0	0

figure 1
tableau initial

	a	b	c	d	e
1	1	1	0	0	0
6	1	1	0	1	0
7	0	1	0	1	0
9	1	0	0	1	0
3	1	0	1	1	0
4	1	0	1	1	1
8	1	0	1	1	1
2	1	0	0	0	1
5	1	0	0	0	1
10	1	0	0	0	0

figure 2
tableau initial
réordonné

	a	b	c	d	e
I	1	0	0	1	0
A	1	1	0	1	0
B	1	0	1	1	1
C	1	0	0	0	1

figure 3
centres médians

	a	b	c	d	e
I	8	4	4	2	2
A	2	2	4	2	4
B	3	3	3	3	1
C	3	3	3	3	1

figure 4
pondérations

	a	b	c	d	e
I	0	67	100	75	50
A	0	67	0	0	0
B	0	0	100	0	25
C	0	0	0	75	25

figure 5
les indices **COR**

	A	B	C
T(k)	33	33	33
B(k)	20	40	40
W(k)	60	20	20

figure 6
description des classes

3.4.5 Programme

Nous avons déja présenté le programme MNDBIN et les résultats qu'il fournit à l'utilisateur. En complément, nous avons écrit un programme fournissant les indices d'aide à l'interprétation que nous avons définis dans un paragraphe précédent. Celui-ci

est appelé INPABN (INdices de description d'une PArtition sur données BiNaires) et intégré au logiciel SICLA. Ce logiciel ainsi complété contient alors trois programmes de classification et trois programmes d'aide à l'interprétation pouvant être appliqués à trois types de données :

- MNDQAN et INPAQN pour les données quantitatives,
- MNDQAL et INPAQL pour les données qualitatives,
- MNDBIN et INPABN pour les données binaires.

3.5 INFLUENCE DE LA TRANSFORMATION DU TABLEAU INITIAL

Nous nous posons ici la question suivante : quel sera le comportement de la méthode MNDBIN lorsque, dans le tableau initial, on inverse les valeurs d'une ou plusieurs colonnes ?

Définition de la transformation
Considérons le tableau binaire $\mathbf{X}(I,J)$ et le tableau de pondérations $\alpha(I,J)$. Nous appelons transformation de ce tableau par rapport à un point b de $\mathbf{B}^p$ le tableau $\mathbf{X}(I(b),J(b))$ définit par :

$$\mathbf{X}(I(b),J(b)) = (x_i^j(b))$$

$$\text{où } \forall\, i \in I,\ \forall\, j \in J \quad x_i^j(b) = | x_i^j - b^j |$$

Ainsi, une colonne j est inversée si et seulement si la composante j de b est égale à 1. Lorsque le point b est le centre médian du nuage associé au tableau, on obtient ce que nous appelons un tableau centré.

A partir du tableau $\mathbf{X}(I(b),J(b))$, on définit alors le nuage $\mathbf{N}(I(b))$ par :

$$\mathbf{N}(I(b)) = \{ (x_i(b), \alpha(x_i(b))), i \in I \}$$

$$\text{où} \quad x_i(b) = | x_i - b |$$

$$\text{et} \quad \alpha(x_i(b)) = \alpha(x_i)$$

Considérons la base canonique usuelle $(e_1,\ldots,e_p)$ et l'origine $O=(0,\ldots,0)$ de $\mathbf{B}^p$. Les coordonnées des points du nuage $\mathbf{N}(I)$ sont exprimées par rapport à ce repère. Par contre, la transformation effectuée sur les données initiales revient à placer l'origine du repère au point b. Les coordonnées des points du nuage $\mathbf{N}(I(b))$ sont alors exprimées par rapport au nouveau repère.

Le centre médian
On peut facilement démontrer le résultat suivant : si a est le centre médian de $\mathbf{N}(I)$, alors la transformation $a(b)=|a - b|$ de a est le centre médian du nuage $\mathbf{N}(I(b))$. Ce résultat s'applique également au centre médian d'un sous-ensemble du nuage et, plus généralement, aux centres médians des classes d'une partition.

Application de la méthode
Considérons une partition P de I en K classes. Dans le repère d'origine O ou dans celui d'origine b, les quantités mesurées par les différentes inerties (intervenant dans la relation de décomposition) sont les mêmes.

Par exemple, pour l'inertie du nuage on a :

$$\Im(\mathbf{N}(I(b))) = \sum_{i\in I}\sum_{j\in J} \alpha^j(x_i)\, |x_i^j(b) - a^j(b)|$$

$$\Leftrightarrow \quad \Im(\mathbf{N}(I(b))) = \sum_{i\in I}\sum_{j\in J} \alpha^j(x_i)\, ||x_i^j - b^j| - |a^j - b^j||$$

$$\Leftrightarrow \quad \Im(\mathbf{N}(I(b))) = \sum_{i\in I}\sum_{j\in J} \alpha^j(x_i)\, |x_i^j - a^j|$$

$$\Leftrightarrow \quad \Im(\mathbf{N}(I(b))) = \Im(\mathbf{N}(I))$$

De façon analogue, on démontre que les inerties interclasse et intraclasse ne subissent pas l'influence du changement de repère. La méthode MNDBIN peut donc être appliquée indifféremment à partir de tout tableau transformé. Changer les valeurs d'une ou plusieurs colonnes du tableau initial n'influe pas sur les résultats fournis par

Remarque
Notons $\underline{J}$ l'ensemble des p variables égales aux p variables inversées de J : une variable de $\underline{J}$ prend la valeur 1 (resp. 0) si et seulement si la variable correspondante de J prend la valeur 0 (resp. 1). Considérons un ensemble Q de p variables appartenant à $J \cup \underline{J}$. Toute variable de cet ensemble apparait soit sous sa forme initiale, soit sous une forme inversée, mais jamais sous les deux formes. Le tableau **X**(I,Q) correspond alors à une transformation du tableau initial **X**(I,J). Les résultats démontrés dans ce paragraphe peuvent être interpétés de la façon suivante : la méthode MNDBIN fournit toujours un même résultat à toutes les applications sur des tableaux de type **X**(I,Q).

4. LA MÉTHODE DE CLASSIFICATION CROISÉE CROBIN

4.1 LE PRINCIPE DE LA CLASSIFICATION CROISÉE (G. GOVAERT 1983)

Rappelons rapidement le principe de la classification croisée : ayant un tableau croisant deux ensembles I (les individus) et J (les variables), il s'agit de trouver simultanément une partition P de I et une partition Q de J de manière à obtenir, comme en analyse factorielle, des résultats en même temps sur les deux ensembles.

Lorsque le tableau envisagé est un tableau de contingence, un tableau de variables qualitatives ou encore un tableau de variables quantitatives, la méthode de classification croisée repose sur la notion de mesure d'information associée à un tableau. Cette mesure peut être le Khi2 de contingence, dans le cas de tableau de contingence, ou l'inertie, dans le cas de données quantitatives. La méthode fournit un tableau résumant le tableau initial et maximisant cette mesure d'information. Elle procède de façon itérative en effectuant, à chaque étape, une classification sur l'ensemble I ou l'ensemble J. Dans les deux cas, l'algorithme intermédiaire utilisé est le même et il maximise, à chaque fois, la mesure d'information.

La méthode de classification croisée sur données binaires CROBIN ne suit pas ce principe. Elle n'utilise pas une mesure d'information, mais un critère mesurant l'écart entre le tableau initial et le tableau résumé. Le problème est alors de trouver le couple de partitions conduisant à l'écart le plus faible possible.

4.2 LA MÉTHODE CROBIN (G. GOVAERT 1983)

Soit **X**(I,J) un tableau croisant un ensemble I de n individus et un ensemble J de p variables binaires.

La méthode CROBIN fournit une solution locale au problème d'optimisation suivant :

trouver une partition P de I en K classes, une partition Q de J en M classes et un tableau binaire à K lignes et M colonnes

$$A = (a_k^m)$$

tels que le critère

$$W(P,Q,A) = \sum_{k=1}^{K} \sum_{m=1}^{M} \sum_{i \in P_k} \sum_{j \in Q^m} / x_i^j - a_k^m /$$

soit minimum.

Pour résoudre ce problème, la méthode CROBIN utilise deux algorithmes, voisins l'un de l'autre, et tous deux construits sur le principe des Nuées Dynamiques :

- le premier permet de déterminer, à Q fixé, une partition P de I et un tableau **A** améliorant le critère **W**(P,Q,A),
- le second permet de déterminer, à P fixé, une partition Q de J et un tableau **A** améliorant encore le critère **W**(P,Q,A).

En partant de deux nombres K et M et d'un point initial (P ou Q), l'algorithme CROBIN fournit, à la convergence, une solution au problème posé. Notons également que ces deux algorithmes recherchent des noyaux particuliers ayant, dans les deux cas, une forme spécifique : chaque composante binaire de ces noyaux est pondérée (par les cardinaux des classes de Q dans le premier algorithme, par les cardinaux des classes de P dans le second). Le tableau **A** recherché lors d'une itération de l'une ou l'autre étape intermédiaire est construit de la façon suivante :

$$\forall\, k=1,\ldots,K,\ \forall\, m=1,\ldots,M \quad a_k^m = \text{médiane de } \{\, x_i^j,\ i \in P_k \text{ et } j \in Q^m \,\}$$

Nous allons montrer que cette approche peut se ramener à la recherche d'un tableau résumé maximisant une mesure d'information. Ainsi, toutes les méthodes de classification croisée seront unifiées. Nous allons plus loin en proposant une forme généralisée de l'algorithme CROBIN.

4.3 LA MESURE D'INFORMATION

A partir des tableaux **X**(I,J) et α(I,J), on peut définir les nuages **N**(I) de $\mathbf{B}^p$ et **N**(J) de $\mathbf{B}^n$ par :

$$\mathbf{N}(I) = \{\, (x_i, \alpha_i),\ i \in I \,\}$$
$$\mathbf{N}(J) = \{\, (x^j, \alpha^j),\ j \in J \,\}$$

où x_i, α_i sont les vecteurs lignes et x^j, α^j les vecteurs colonnes des tableaux de départ.

L'inertie du nuage **N**(I) par rapport à l'origine de $\mathbf{B}^p$ est égale à l'inertie du nuage **N**(J) par rapport à l'origine de $\mathbf{B}^n$. En effet, on a :

$$\Im_0(\mathbf{N}(\mathrm{I})) = \sum_{i\in I}\sum_{j\in J} \alpha_i^j x_i^j = \sum_{j\in J}\sum_{i\in I} \alpha_i^j x_i^j = \Im_0(\mathbf{N}(\mathrm{J}))$$

Définition
Nous définissons la mesure d'information du tableau initial X(I,J) comme l'inertie commune des nuages des lignes et des colonnes par rapport à leur origine respective :

$$\mathit{Info}(X(I,J)) = \Im_0(N(I)) = \Im_0(N(J))$$

Cette information est une mesure de la présence de la valeur 1 dans le tableau **X**(I,J). Si toutes les pondérations initiales sont égales à 1, il s'agit exactement du nombre de valeurs 1 contenues dans le tableau.

Si on prend soin de centrer la nuage des individus, c'est à dire de prendre pour origine de l'espace le centre médian du nuage, l'information est alors une mesure de la présence des valeurs minoritaires du tableau **X**(I,J).

4.4 TABLEAU ASSOCIÉ À UN COUPLE DE PARTITIONS

Soient P une partition de I en K classes et Q une partition de J en M classes.

Définition
Au couple de partition (P,Q), nous associons le tableau X(P,Q) d'ordre (K,M) défini par :

$$X(P,Q) = (x_k^m(P,Q))$$

où

$$\forall m=1,\ldots,M,\ \forall k=1,\ldots,K \quad x_k^m(P,Q) = \text{médiane}\ \{\ (x_i^j,\ \alpha_i^j),\ i\in P_k \text{ et } j\in Q^m\ \}.$$

On définit comme tableau de pondérations associé au tableau X(P,Q) le tableau α(P,Q) suivant :

$$\alpha_k^m(P,Q) = |\ n_k^m(1) - n_k^m(0)\ |$$

où

$n_k^m(1) = \sum_{i\in P_k}\sum_{j\in Q^m} \alpha_i^j x_i^j$ *somme des pondérations des valeurs 1 de* $X(P_k,Q^m)$,

$n_k^m(0) = \sum_{i\in P_k}\sum_{j\in Q^m} \alpha_i^j (1 - x_i^j)$ *somme des pondérations des valeurs 1 de* $X(P_k,Q^m)$.

Remarques sur la construction du tableau X(P,Q)
Le tableau **X**(P,Q) peut être obtenu en deux étapes, en construisant tout d'abord le tableau **X**(I,Q) puis, à partir de ce dernier, le tableau **X**(P,Q) lui même.

De façon équivalente, on peut également obtenir le tableau **X**(P,Q) en construisant tout d'abord le tableau **X**(P,J).

Le tableau **X**(I,Q) est défini à partir du tableau initial **X**(I,J) de la façon suivante :

$$\forall\, i \in I,\ \forall\, m=1,\ldots,M \quad x_i^m(I,Q) = \text{médiane de } \{\ (x_i^j, \alpha_i^j),\ j \in Q^m\ \}$$

$$\text{et} \quad \alpha_i^m(I,Q) = \left| \sum_{j \in Q^m} \alpha_i^j x_i^j - \sum_{j \in Q^m} \alpha_i^j (1 - x_i^j) \right|$$

A partir de **X**(I,Q), nous pouvons contruire le tableau **X**(P,Q) suivant :

$$\forall\, K=1,\ldots,K,\ \forall\, m=1,\ldots,M \quad x_k^m(P,Q) = \text{médiane de } \{\ (x_i^m(I,Q), \alpha_i^m(I,Q)),\ i \in P_k\ \}$$

$$\text{et} \quad \alpha_k^m(P,Q) = \left| \sum_{i \in P_k} \alpha_i^m(I,Q) x_i^m(I,Q) - \sum_{i \in P_k} \alpha_i^m(I,Q)(1 - x_i^m(I,Q)) \right|$$

La propriété de conservation de la médiane permet de montrer que le tableau, ainsi construit, correspond exactement à la définition du tableau **X**(P,Q).

Remarque sur les nuages associées aux tableaux X(I,Q) et X(P,J)
A partir des lignes du tableau **X**(I,Q) et des pondérations associés, nous pouvons construire le nuage **N**(I/Q) inclus dans $\mathbf{B}^{\mathbf{M}}$. Si P est une partition en K classes de ce nuage, les centres médians des classes sont, par construction, les vecteurs lignes du tableau **X**(P,Q). De plus, les vecteurs de pondérations de ces centres sont les vecteurs lignes du tableau de pondérations associé α(P,Q).

A partir des colonnes du tableau **X**(P,J) et des pondérations associées, nous pouvons construire le nuage **N**(J/Q) inclus dans $\mathbf{B}^{\mathbf{K}}$. Si Q est une partition de ce nuage, les centres médians des classes et leurs vecteurs de pondérations sont, cette fois, les vecteurs colonnes de **X**(P,Q) et du tableau de pondérations associé α(P,Q).

Mesure d'information associée au tableau X(P,Q)
Le tableau **X**(P,Q) étant un tableau binaire muni de pondérations, nous pouvons lui associer une mesure d'information. Pour cela, considérons les deux nuages **N**(P/Q) de $\mathbf{B}^{\mathbf{M}}$ et **N**(Q/P) de $\mathbf{B}^{\mathbf{K}}$ suivant :

$$\mathbf{N}(P/Q) = \{\ (x_k, \alpha_k),\ k=1,\ldots,K\ \} \text{ et } \mathbf{N}(Q/P) = \{\ (x^m, \alpha^m),\ m=1,\ldots,M\ \}$$

où :

- les x_k sont les vecteurs lignes et les x^m les vecteurs colonnes de **X**(P,Q),
- les α_k sont les vecteurs lignes et les α^m les vecteurs colonnes de α(P,Q).

La mesure d'information associée au tableau **X**(P,Q) est alors la suivante :

$$\mathbf{Info}(\mathbf{X}(P,Q)) = \Im_0(\mathbf{N}(P/Q)) = \Im_0(\mathbf{N}(Q/P))$$

$$\Leftrightarrow \quad \mathbf{Info}(\mathbf{X}(P,Q)) = \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_k^m(P,Q)\, x_k^m(P,Q)$$

Il nous reste maintenant à démontrer que les deux algorithmes intermédiaires utilisés par CROBIN ne sont autre que l'algorithme MNDBIN généralisé, et que la méthode CROBIN optimise, en fait, la mesure d'information **Info** du tableau **X**(P,Q). C'est ce que nous montrons dans le paragraphe suivant.

4.5 LA NOUVELLE APPROCHE

Dans un premier temps, nous considérons que le tableau initial **X**(I,J) est associé à un tableau de pondérations toutes égales à 1. C'est à ce type de tableau que s'applique la méthode CROBIN existante.

Puis, après avoir montré que CROBIN s'inscrit dans le cadre général des méthodes de classification croisée, nous proposons un algorithme CROBIN généralisé qui permet de prendre en compte des pondérations initiales quelconques.

4.5.1 Les deux algorithmes intermédiaires

Soit Q une partition de J en M classes, Q initialement fixée. Le premier algorithme intermédiaire de CROBIN recherche une partition P de I en K classes et un tableau binaire **A** d'ordre (K,M) minimisant le critère **W**(P,Q,**A**) que nous avons déjà présenté. Cet algorithme recherche en fait un ensemble de noyaux particuliers que l'on peut qualifier de pondérés.

Le tableau des noyaux **A** recherché par CROBIN n'est autre que le tableau **X**(P,Q) associé au couple de partition (P,Q). Cet algorithme intermédiaire n'est autre que l'algorithme MNDBIN généralisé appliqué au nuage **N**(I/Q) défini à partir du tableau **X**(I,Q) et du tableau de pondérations α(I,Q).

Alors que l'algorithme intermédiaire de CROBIN recherche des noyaux pondérés, l'algorithme MNDBIN généralisé recherche, lui, des noyaux binaires et travaille sur des données initiales qui, elles, sont pondérés. Le critère optimisé est alors l'inertie intraclasse **W**(P/Q) de la partition P du nuage **N**(I/Q). C'est ce que nous montrons dans

Soient (P,Q) un couple de partitions en K et M classes de (I,J). Le critère correspondant à le méthode CROBIN s'écrit **W**(P,Q,**X**(P,Q)). Aprés avoir décomposé l'expression de **W**(P,Q,**X**(P,Q)), nous avons fait apparaitre un lien avec l'inertie intraclasse **W**(P/Q), lien qui se traduit par la relation suivante :

$$\mathbf{W}(P,Q,\mathbf{X}(P,Q)) = \mathbf{W}(P/Q) + \mathbf{W}(Q)$$

où **W**(Q) est l'inertie intraclasse de la partition Q du nuage des variables **N**(J). Si Q est fixée, cette relation montre bien le lien existant entre l'algorithme intermédiaire et la méthode MNDBIN généralisée.

De la même manière, pour une partition P fixée, le second algorithme intermédiaire de CROBIN n'est autre que la méthode MNDBIN généralisée appliquée, cette fois, au nuage **N**(J/P) défini à partir des colonnes du tableau **X**(P,J) et du tableau de pondérations α(P,J). Le crière correspondant s'exprime alors comme l'inertie intraclasse **W**(Q/P) de la partition Q du nuage **N**(J/P). A nouveau, nous pouvons construire une relation du même type que la précédente :

$$\mathbf{W}(P,Q,\mathbf{X}(P,Q)) = \mathbf{W}(Q/P) + \mathbf{W}(P)$$

où **W**(P) est l'inertie intraclasse de la partition P du nuages des individus **N**(I). Cela montre le lien entre ces deux algorithmes

Il nous reste maintenant à montrer que l'algorithme CROBIN optimise en fait la mesure d'information **Info** associée au tableau binaire **X**(P,Q).

4.5.2 La méthode CROBIN et la mesure d'information

Soit (P,Q) un couple de partitions en K et M classes de (I,J). Nous savons alors construire le tableau **X**(P,Q) ainsi que les tableaux **X**(I,Q) et **X**(P,J) et les nuages associés **N**(I/Q) (inclus dans $\mathbf{B}^M$) et **N**(J/P) (inclus dans $\mathbf{B}^K$).

Puisque les centres médians des classes de P sont les vecteurs lignes du tableau **X**(P,Q), l'inertie intraclasse de la partition P de **N**(I/Q) s'écrit :

$$\mathbf{W}(P/Q) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{m=1}^{M} \alpha_i^m(I,Q) \left| x_i^m(I,Q) - x_k^m(P,Q) \right|$$

De manière symétrique, l'inertie intraclasse de la partition Q de **N**(J/P) s'écrit :

$$\mathbf{W}(Q/P) = \sum_{m=1}^{M} \sum_{j \in Q^m} \sum_{k=1}^{K} \alpha_k^j(P,J) \left| x_k^j(P,J) - x_k^m(P,Q) \right|$$

Propriétés
*Les inerties intraclasses **W**(P/Q) et **W**(Q/P) sont liées aux mesures d'information des tableaux **X**(I,Q), **X**(P,J) et **X**(P,Q) par les relations suivantes :*

$$(i) \qquad \textbf{\textit{Info}}(\mathbf{X}(I,Q)) = \mathbf{W}(P/Q) + \textbf{\textit{Info}}(\mathbf{X}(P,Q))$$

$$(ii) \qquad \textbf{\textit{Info}}(\mathbf{X}(P,J)) = \mathbf{W}(Q/P) + \textbf{\textit{Info}}(\mathbf{X}(P,Q))$$

Preuve
Nous démontrons ici la relation (i) (la seconde relation se démontre de la même façon). La mesure d'information du tableau **X**(P,Q) est égale à l'inertie du nuage **N**(P/Q) (défini dans le paragraphe 4.4) par rapport à l'origine. Or, ce nuage n'est autre que le nuage des centres médians des classes de la partition P de **N**(I/Q). Si on décompose l'inertie du nuage **N**(I/Q) par rapport à l'origine, on obtient alors :

$$\Im_0(\mathbf{N}(I/Q)) = \mathbf{W}(P/Q) + \sum_{k=1}^{K} \sum_{m=1}^{M} \alpha_k^m(P,Q)\, x_k^m(P,Q)$$

$$\Leftrightarrow \quad \Im_0(\mathbf{N}(I/Q)) = \mathbf{W}(P/Q) + \Im_0(\mathbf{N}(P/Q))$$

Cette relation n'est autre que la relation (i) puisque, par définition, on a :

$$\textbf{Info}(\mathbf{X}(I,Q)) = \Im_0(\mathbf{N}(I/Q))$$

$$\text{et} \quad \textbf{Info}(\mathbf{X}(P,Q) = \Im_0(\mathbf{N}(P/Q))$$

La nouvelle interprétation de la méthode CROBIN
Dans un premier temps, nous avons démontré que l'algorithme CROBIN minimise, tour à tour, les inerties interclasses **W**(P/Q) puis **W**(Q/P). Pour cela, il utilise l'algorithme MNDBIN généralisée qui apparait comme l'unique algorithme intermédiaire de la méthode. Enfin, l'introduction de la mesure d'information pour tableaux binaires nous a conduit aux deux relations (i) et (ii). Celles-ci permettent alors de montrer que l'algorithme CROBIN recherche une suite de couple (P,Q) de partitions maximisant la mesure d'information **Info** du tableau **X**(P,Q).

Nous pouvons maintenant aller plus loin, et proposer un algorithme CROBIN plus général travaillant à partir d'un tableau initial dont les éléments sont munis de pondérations, non plus toutes égales à 1, mais quelconques.

4.6 L'ALGORITHME GÉNÉRALISÉ

Considérons le problème de classification croisée suivant :

> *Soient X(I,J) le tableau de données binaires et α(I,J) le tableau de pondérations. Il s'agit de trouver un couple de partitions (P,Q) de (I,J) tel que la mesure d'information associée au tableau X(P,Q) soit maximale.*

Tous les résultats vus précédemment, et en particulier les relations (i) et (ii), peuvent facilement être redémontrées en tenant compte de pondérations initiales quelconques. Ces deux relations nous permettent alors de construire un algorithme CROBIN généralisé qui utilise, comme algorithme intermédiaire, l'algorithme MNDBIN

En partant d'un point initial (P ou Q) il optimise, à chaque étape, la mesure d'information du tableau X(P,Q). Pour cela, il construit une suite de couple (P^r,Q^r) en procédant de la façon habituelle :

- Soient Q^r une partition de J et $X(I,Q^r)$ le tableau associé au couple (I,Q^r). La méthode MNDBIN généralisée fournit une partition P^{r+1} de I et un tableau $X(P^{r+1},Q^r)$ en maximisant la mesure d'information associée au tableau $X(P^{r+1},Q^r)$ (d'après la relation (i)).

- Soient P^{r+1} une partition de I et $X(P^{r+1},J)$ le tableau associé au couple (P^{r+1},J). La méthode MNDBIN généralisée fournit une partition Q^{r+1} de J et un tableau $X(P^{r+1},Q^{r+1})$ en maximisant la mesure d'information associée au tableau $X(P^{r+1},Q^{r+1})$ (d'après la relation (ii)).

Dans le cas où toutes les pondérations sont égales à 1, on retrouve, bien sûr, l'algorithme CROBIN initial.

5. CLASSIFICATION ASCENDANTE HIÉRARCHIQUE SUR DONNÉES BINAIRES

Lorsque les données sont binaires, on dispose d'un grand nombre d'indices de dissimilarité entre individus. A partir de ceux-ci, on peut construire des indices de proximité entre classes comme l'indice du lien minimum ou single linkage, l'indice du lien maximum ou complete linkage ou encore l'indice du lien moyen ou average linkage (E. Diday et al. 1982).

Une autre approche consiste à plonger les données binaires dans un espace du type $\mathbf{R}^p$. On peut alors utiliser les indices de proximités définis dans le cas quantitatif. Le plus utilisé est l'indice de d'augmentation d'inertie ou indice de Ward (1963). Nous pouvons également citer l'indice de l'inertie de la réunion de deux classes (M. Jambu 1978).

Nous proposons ici une nouvelle approche. A partir des propriétés démontrées sur l'espace $\mathbf{B}^p$, nous pouvons définir un certain nombre d'indices d'agrégation. En fait, les indices que nous proposons sont définis de façon analogue à ceux proposés dans le cas quantitatif. Parmi ceux-ci, certains ne peuvent être utilisés pour indicer une hiérarchie.

C'est le cas notamment de l'indice analogue à l'indice de Ward. Comme l'indice de la distance euclidienne entre centres de gravités des classes (Sokal et Michener 1958) dans le cas quantitatif, l'indice de la distance (en valeurs absolues) entre les centres médians des classes pose également un problème d'inversion.

Par contre, l'indice de l'inertie binaire convient parfaitement. Nous proposons, en outre, un nouvel indice d'agrégation défini comme une mesure de l'homogénéité de la réunion de deux classes. Pour ces indices, une relation de récurrence n'a pu être démontrée.

5.1 NOTATIONS

On reprend ici les notations suivantes :

- **X(I,J)** le tableau croisant un ensemble I de n individus et un ensemble J de p variables binaires. Nous supposons ici que toutes les pondérations sont égales à 1. La généralisation à des pondérations quelconques ne pose aucun problème particulier.
- **N(I)** le nuage défini à partir du tableau **X**(I,J).
- **N(A)** le nuage défini à partir d'une partie A de I et du sous-tableau **X**(A,J) correspondant.

5.2 LES LIMITES DE L'ANALOGIE

Lorsque les données sont quantitatives, on dispose de l'indice de Ward qui peut également être utilisé pour indicer la hiérarchie. Nous définissons ici un indice analogue dans le cadre de l'espace $\mathbf{B}^p$ muni de la distance en valeurs absolues.

Cependant, celui-ci pose des problèmes d'inversion et ne peut être utilisés pour indicer une hiérarchie. Le même problème se pose si on prend comme critère d'agrégation la distance en valeurs absolues entres les centres médians des classes. C'est ce que nous montrons dans ce paragraphe.

5.2.1 Indice analogue à l'indice de Ward

Le pseudo-indice de Ward
Soient I_1 et I_2 deux parties disjointes de I. Nous définissons le pseudo-indice de Ward comme l'indice de l'augmentation de l'inertie binaire aprés agrégation de I_1 et I_2, soit :

$$\Delta(I_1,I_2) = \Im(\mathbf{N}(I_1 \cup I_2)) - \Im(\mathbf{N}(I_1)) - \Im(\mathbf{N}(I_2))$$

Par définition, cet indice est égal à l'inertie interclasse du couple (I_1,I_2) et :

$$\Delta(I_1,I_2) = \Im(\{a_1,a_2\})$$

où

a_1 est le centre médian du nuage $\mathbf{N}(I_1) = \{ (x_i, \alpha(x_i)), i \in I_1 \}$,

a_2 est le centre médian du nuage $\mathbf{N}(I_2) = \{ (x_i, \alpha(x_i)), i \in I_2 \}$.

Autre interprétation de l'indice
Les inerties binaires intervenant dans la définition de l'indice peuvent s'interpréter de la façon suivante :

$\Im(\mathbf{N}((I_1 \cup I_2))$ est égale au nombre de valeurs minoritaires de $\mathbf{X}(I_1 \cup I_2, J)$,

$\Im(\mathbf{N}(I_1))$ est égale au nombre de valeurs minoritaires de $\mathbf{X}(I_1, J)$,

$\Im(\mathbf{N}(I_2))$ est égale au nombre de valeurs minoritaires de $\mathbf{X}(I_2, J)$.

L'indice Δ apparait alors comme l'indice de l'augmentation du nombre de valeurs minoritaires après agrégation de deux classes.

L'indice n'est pas croissant
L'indice Δ ne peut pas être utilisé pour indicer la hierarchie. A travers l'exemple ci-dessous, nous montrons que l'indice n'est pas croissant et qu'un problème d'inversion peut surgir.

Soit le tableau binaire, croisant 3 individus $\{\mathbf{1},\mathbf{2},\mathbf{3}\}$ et trois variables binaires $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$, défini par :

	a	**b**	**c**
1	1	1	0
2	1	0	1
3	0	1	1

avec des pondérations initiales toutes égales à 1.

L'indice Δ, entre deux quelconques de ces points, est toujours égal à 2. Agrégeons, par exemple, les deux premiers individus :

$$\Delta(\{\mathbf{1}\},\{\mathbf{2}\}) = 2$$

Le nombre de valeurs minoritaires, initialement égal à 0, est donc augmenté de 2 (une valeur minoritaire pour **b** et une pour **c**).

Après agrégation du troisième point, on obtient l'indice :

$$\Delta(\{\mathbf{1},\mathbf{2}\},\{\mathbf{3}\}) = 1$$

L'agrégation de $\{\mathbf{1},\mathbf{2}\}$ et de $\{\mathbf{3}\}$ induit donc une augmentation d'une seule valeur minoritaire (celle prise par **a** sur l'individu **3**), c'est-à-dire moins que lors de l'étape précédente. Les variables (**b**,**c**) n'avaient pas de majorité dans la partie $\{\mathbf{1},\mathbf{2}\}$. L'agrégation avec $\{\mathbf{3}\}$ a permis de leur en fournir une, mais cela ne conduit à aucune augmentation du nombre de valeurs minoritaires pour ces deux variables.

Remarque
Dans la pratique, l'utilisation du pseudo-indice de Ward aboutit parfois à une hiérarchie convenablement indicée. Nous en donnerons un exemple dans la suite.

5.2.2 L'indice de la distance entre centres médians

Comme l'indice précédent, l'indice de la distance en valeurs absolues entre les centres médians des deux classes agrégées peut poser des problèmes d'inversion.

Notons Δ cet indice, on a :

$$\Delta(I_1,I_2) = \sum_{j\in J} | a_1^j - a_2^j |$$

où a_1 et a_2 sont les centres médians des deux parties I_1 et I_2.

Nous donnons, ci-dessous, un exemple simple montrant que cet indice ne peut être utilisé pour indicer une hiérarchie.

Considérons le tableau croisant 3 individus {**1**,**2**,**3**} et 3 variables {**a**,**b**,**c**} défini par :

	a	**b**	**c**
1	0	0	1
2	0	1	0
3	1	0	0

avec des pondérations initiales toutes égales à 1.

La distance entre deux quelconques de ces points est toujours de 2. Si on agrège {**1**} et {**2**}, le centre médian de {**1**,**2**} est alors (0,0,0) (on choisit la valeur 0 si aucune valeur n'est majoritaire). L'indice $\Delta(\{\mathbf{1},\mathbf{2}\},\{\mathbf{3}\})$ égal à 1 est inférieur au précédent.

L'indice ne tient pas compte des vecteurs de pondérations et donc du comportement des variables. Deux parties I_1 et I_2 ayant un même centre médian mais des pondérations associées différentes sont agrégées, même si il existe une partie I_3, ayant un centre médian différent, mais où les variables peuvent avoir un comportement global se rapprochant d'avantage de celui de l'une ou l'autre partie I_1 ou I_2.

5.3 INDICE DE L'INERTIE

Pour des données quantitatives, M. Jambu (1978) propose d'utiliser, comme indice d'agrégation, l'inertie de la réunion de deux classes. Nous définissons ici un indice analogue à partir de l'inertie binaire.

Définition
Nous définissons l'indice de l'inertie de la réunion de deux classes I_1 et I_2 de la façon suivante :

$$\Delta(I_1,I_2) = \Im(N(I_1 \cup I_2))$$

Supposons encore que toutes les pondérations initiales soient égales à 1. Comme nous l'avons déjà vu, cet indice représente alors le nombre de valeurs minoritaires de la partie obtenue après agrégation. Il apparait évident que ce nombre se trouve soit inchangé, soit augmenté après chaque agrégation. L'inertie peut alors être utilisée pour indicer la hiérarchie.

Ce résultat se démontre simplement à partir de la relation de décomposition de l'inertie binaire. Si on note a_1 le centre médian du nuage $\mathbf{N}(I_1)$ et a_2 celui de $\mathbf{N}(I_2)$, on a :

$$\Delta(I_1,I_2) = \Im(\mathbf{N}(I_1 \cup I_2))$$

$$\Leftrightarrow \quad \Delta(I_1 \cup I_2) = \Im(\mathbf{N}(I_1)) + \Im(\mathbf{N}(I_2)) + \Im(\{a_1,a_2\})$$

Cette relation montre que l'indice de hiérarchie correspondant à $I_1 \cup I_2$ est supérieur à l'indice de hiérarchie correspondant à I_1 et à celui correspondant à I_2.

La généralisation à des pondérations initiales quelconques est évidente. L'inertie s'interprète alors comme la somme des pondérations des valeurs minoritaires.

5.4 UN NOUVEL INDICE

Nous proposons ici un indice représentant, cette fois, la part de valeurs minoritaires après agrégation. Ce n'est pas un indice équivalent à l'indice de l'inertie puisqu'il prend en compte les effectifs des parties à agréger.

Définition de l'indice
Soient I_1 et I_2 deux parties de I. Nous définissons le critère d'agrégation, égal à la part de valeurs minoritaires après agrégation, de la façon suivante :

$$\Delta(I_1 \cup I_2) = \frac{2\, \Im(N(I_1 \cup I_2))}{p\,(n_1 + n_2)}$$

où p est le nombre de variables, n_1 le cardinal de I_1 et n_2 le cardinal de I_2.

L'indice Δ varie de la valeur minimale 0 à la valeur maximale 1, de sorte que :

$\Delta=0$ si la réunion $I_1 \cup I_2$ contient des individus ayant tous répondus de façon identique aux variables,

$\Delta=1$ si, après agrégation de I_1 et I_2, aucune variable n'a une majorité

Au cours de l'algorithme de classification ascendante hiérarchique, l'indice Δ croît. Après agrégation, le nombre de valeurs minoritaires se trouve augmenté. L'indice croît également sauf dans le cas extrême où les centres médians des deux parties I_1 et I_2 sont égaux. Cela ne peut se produire que si la classe I_1 (par exemple) contient des individus de vecteurs réponses tous identiques et si la classe I_2 est réduite à un élément se comportant comme ceux de I_1 (I_2 ne peut contenir plus d'un élément car cela serait synonyme d'une mauvaise agrégation lors d'une étape précédente). Dans ce cas de figure, les indices $\Delta(I_1,I_2)$, $\Delta(I_1)$ et $\Delta(I_2)$ sont tous trois égaux à 0 (où, pour k=1,2, $\Delta(I_k)$ représente la part de valeurs minoritaire de I_k).

Généralisation à des pondérations initiales quelconques
En utilisant la définition de l'inertie par rapport à l'origine et de l'inertie par rapport au point dont les coordonnées sont toutes égales à 1, nous obtenons l'expression suivante de l'indice :

$$\Delta(I_1,I_2) = \frac{2\, \Im(N(I_1 \cup I_2))}{\Im_0(N(I_1 \cup I_2)) + \Im_1(N(I_1 \cup I_2))}$$

Cette nouvelle expression permet de prendre en compte des pondérations initiales quelconques : il représente la part de valeurs minoritaires, calculée en tenant compte des pondérations de ces valeurs.

5.5 PROGRAMME ET APPLICATIONS

Le programme CAHBIN a également été conçu dans le cadre du logiciel SICLA. Il propose à l'utilisateur le choix entre les différents indices étudiés ici : le pseudo-indice de Ward (paragraphe 5.1), l'indice de l'inertie binaire (paragraphe 5.3) et l'indice de la part de valeurs minoritaires (paragraphe 5.4).

Nous avons conservé le pseudo-indice de Ward dans la programmation car, pour certains tableaux binaires, il aboutit à une hiérarchie convenablement construite. C'est le cas, par exemple, des données MERO (déjà étudiées dans le chapitre 2). Nous allons d'ailleurs utiliser ce jeu de données pour illustrer le programme CAHBIN.

Rappellons tout d'abord que l'étude de MERO a mis en évidence deux goupes **A** et **B** de plaques-boucles. L'ensemble **A** semblait contenir des plaques-boucles construites avec des techniques plus anciennes que celles du groupe **B**. La méthode MNDBIN a aussi permis de déterminer 5 sous-groupes constitués comme suit :

Groupe $\mathbf{A_1}$: 21 elements

03 04 09 14 20 21 22 24 25 26 28 30
31 32 33 35 6 46 49 52 53

Groupe $\mathbf{A_2}$: 9 elements

05 15 23 47 54 55 56 57 58

Groupe $\mathbf{A_3}$: 9 elements

07 11 17 27 34 48 51 50 59

Groupe $\mathbf{B_1}$: 11 elements

01 10 16 37 38 39 40 41 42 43 45

Groupe $\mathbf{B_2}$: 9 elements

02 06 08 12 13 18 19 29 44

Le critère associé à cette partition est égal à **100** (le nombre de désaccords entre noyaux et données initiales). Pour la méthode MNDQAN, la partition obtenue est identique à une permutation près : l'élément **52** passe de $\mathbf{A_1}$ à $\mathbf{A_3}$. Notons que le nombre de désaccords est alors de **101**.

Les hiérarchies contruites à partir des différents indices sont les suivantes :

- $\mathbf{H_1}$ utilisant le pseudo-indice de Ward (page 153),
- $\mathbf{H_2}$ l'indice de l'inertie binaire (page 154),
- $\mathbf{H_3}$ l'indice de la part de valeurs minoritaires (page 155),
- $\mathbf{H_4}$ l'indice de Ward (page 156).

Les coupures en 2 classes de ces hiérarchies sont toutes identiques : on retrouve parfaitement les groupes **A** et **B**.

Les coupures en 5 classes sont assez proches de la partition fournie par MNDBIN. Les deux groupes inclus dans **B** se retrouvent parfaitement. Pour le groupe **A**, des différences apparaissent :

- coupure de $\mathbf{H_1}$: **24** passe de $\mathbf{A_3}$ à $\mathbf{A_1}$, le critère augmente de 1 et passe à **101**.
- coupure de $\mathbf{H_2}$: **34** passe de $\mathbf{A_3}$ à $\mathbf{A_1}$ et **35** de $\mathbf{A_1}$ à $\mathbf{A_2}$, le critère augmente de 3 et passe à **103**.
- coupure de $\mathbf{H_3}$: **7** et **50** passent de $\mathbf{A_3}$ à $\mathbf{A_1}$ et **52** de $\mathbf{A_1}$ à $\mathbf{A_3}$, le critère audmente de 5 et passe à **105**.
- coupure de $\mathbf{H_4}$: on retrouve la partition fournie par MNDQAN, le critère est donc de **101**.

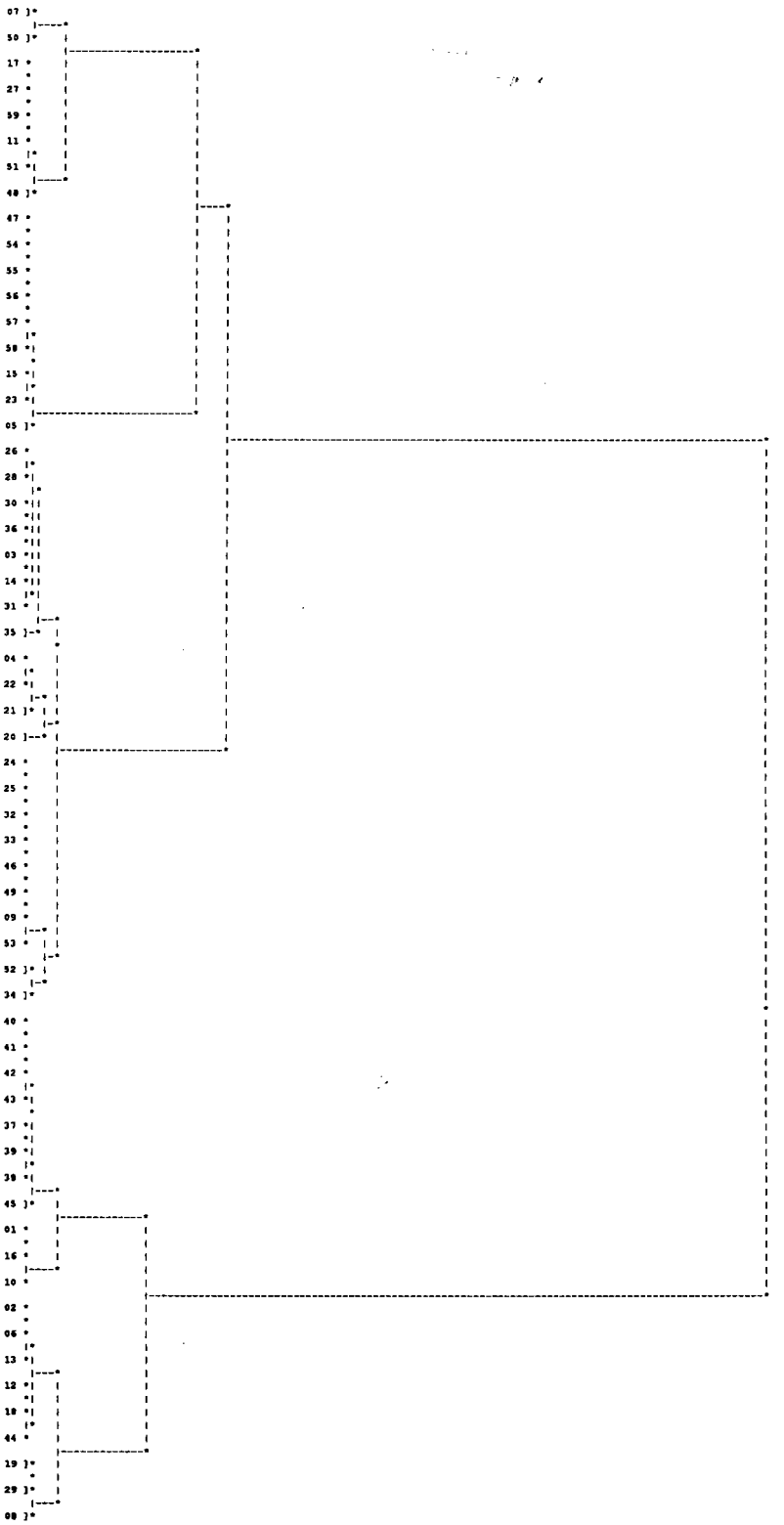
Les hiérarchies obtenues sont toutes différentes, mais on peut noter que la coupure la plus proche de la partition MNDBIN est celle de $\mathbf{H_1}$ (de façon analogue à $\mathbf{H_4}$ et MNDQAN, qui fournit, sur cet exemple, des résultats voisins de ceux de MNDBIN). La hiérarchie $\mathbf{H_2}$ se rappoche d'avantage de $\mathbf{H_1}$ que de $\mathbf{H_3}$, cette dernière se distingue d'ailleurs de toutes les autres. Pour $\mathbf{H_1}$, $\mathbf{H_2}$ et $\mathbf{H_4}$, les 5 classes sont agrégées de façon analogue dans le haut de ces hiérarchies (pour le groupe **A**, notons que $\mathbf{A_2}$ et $\mathbf{A_3}$ sont agrégés en premier lieu). Pour $\mathbf{H_3}$ les classes du goupe **A** sont agrégées de façon différente ($\mathbf{A_1}$ et $\mathbf{A_3}$ en premier lieu).

D'autres essais ont été effectués à partir de données simulées. Il apparait alors que les coupures de la hiérarchie utilisant le pseudo-indice de Ward (si des problèmes d'inversion ne se posent pas) et celles de la hiérarchie utilisant l'inertie binaire sont souvent très proches des partitions fournies par MNDBIN. Lorsque, pour un tableau de données, les applications des méthodes MNDBIN et MNDQAN ne diffèrent que très peu, la hiérarchie utilisant l'indice de Ward est voisine des deux précédentes. Enfin, la hiérarchie construite à partir du critère de la part de valeurs minoritaires se distingue

Après plusieurs autres essais, une remarque peut être faite au sujet du critère de la part de valeurs minoritaires : si le tableau initial est centré, c'est-à-dire si on place le centre médian à l'origine, l'indice correspond alors à la part de valeurs 1; la hiérarchie obtenue est différente du cas non centré, mais se rapproche de façon sensible de celle construite en utilisant l'indice de Ward usuel.

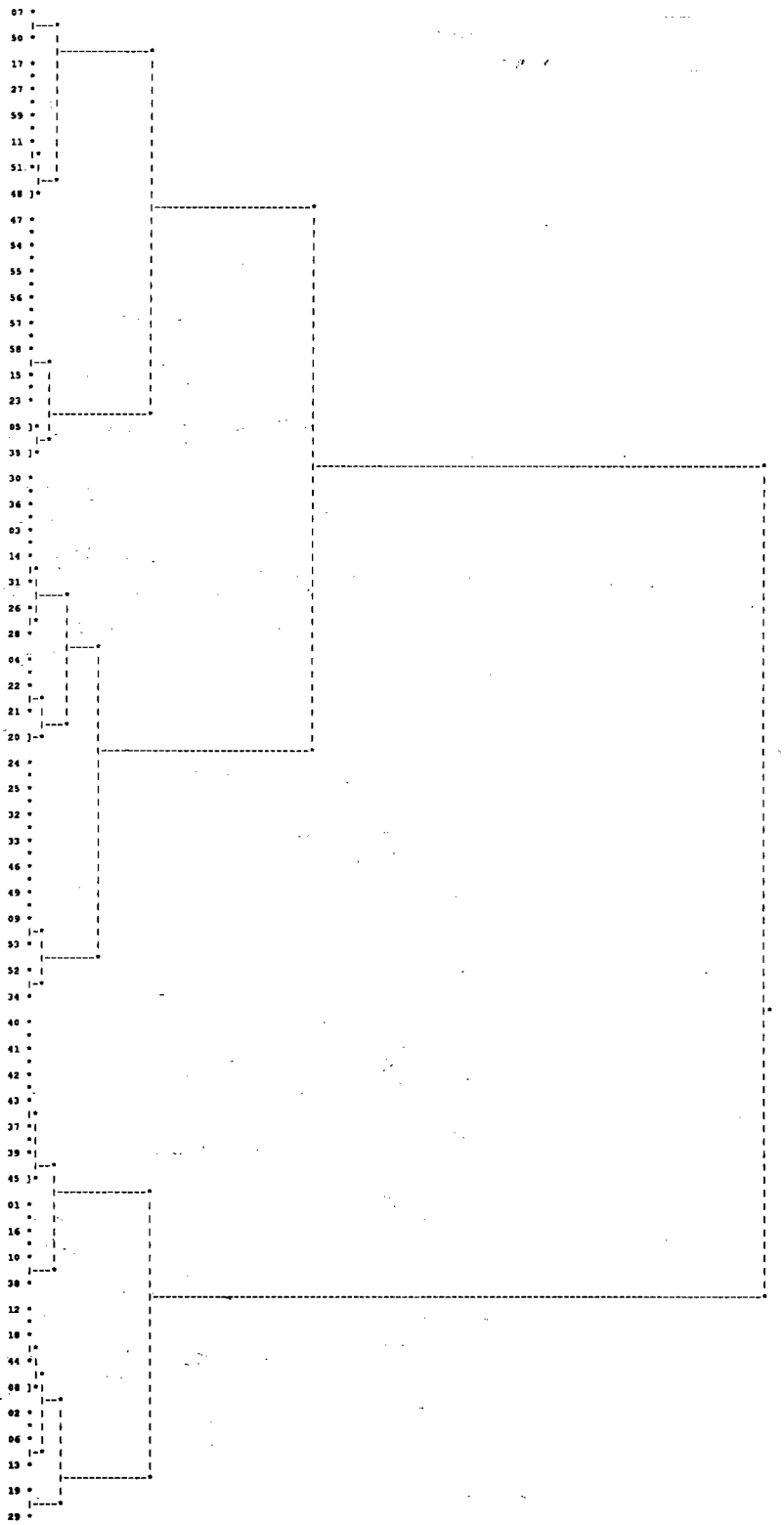
```
COMMANDE : CAHBIN <> classification hierarchique de donnees binaires
```

Arbre de la classification hierarchique utilisant le pseudo-indice de Ward :



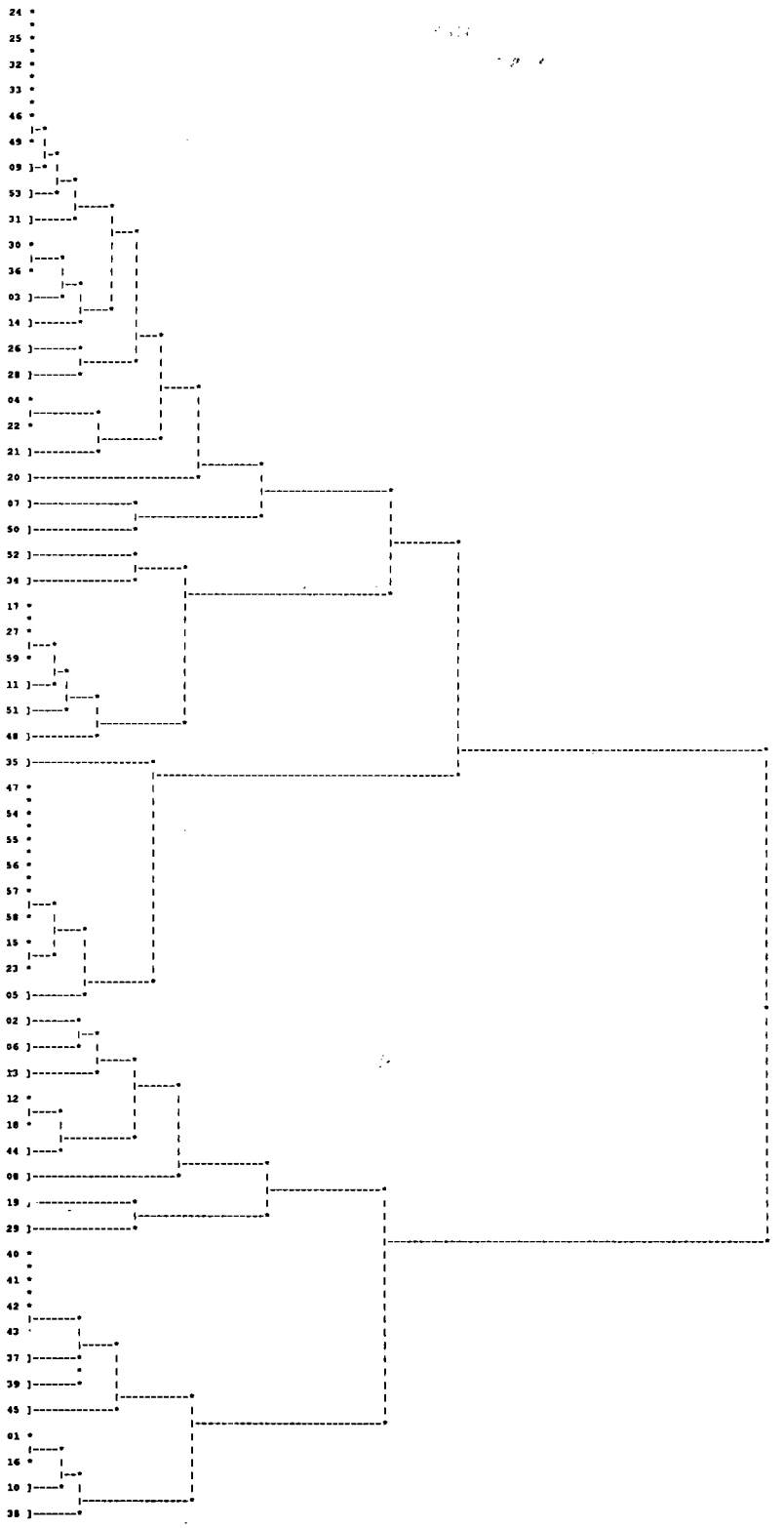
COMMANDE : CAHBIN <> classification hierarchique de donnees binaires

Arbre de la classification hierarchique utilisant l'indice de l'inertie binaire :



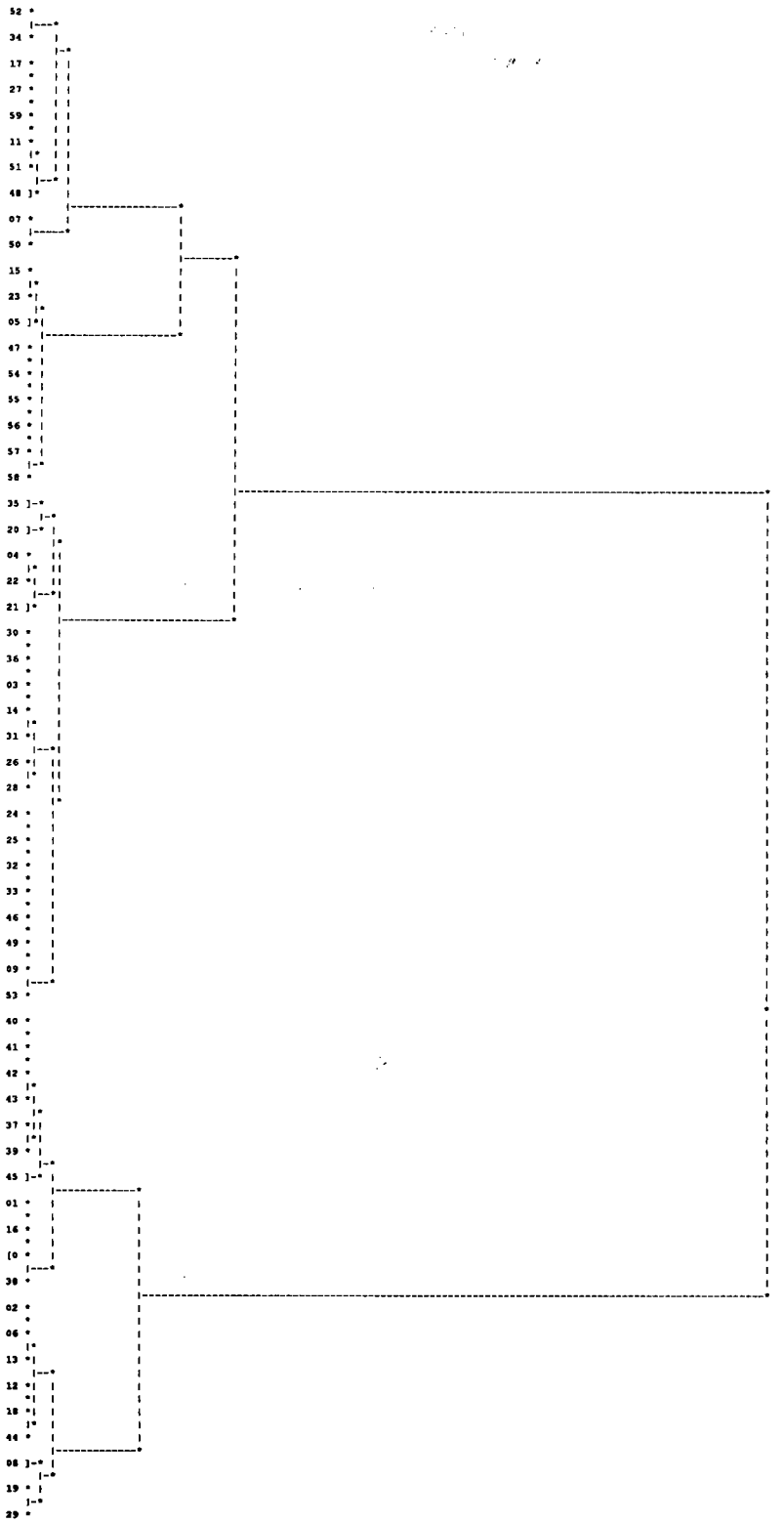
```
COMMANDE : CAHBIN <> classification hierarchique de donnees binaires
```

Arbre de la classification hierarchique utilisant l'indice de la part de valeurs minoritaires :



```
COMMANDE : WARD <> hierarchie avec l'indice de l'accroissement de l'inertie
```

Arbre de la classification hierarchique :



CHAPITRE 6

CLASSIFICATION ET ANALYSE EN COMPOSANTES PRINCIPALES POUR DONNÉES BINAIRES

1. INTRODUCTION

Lorsque les données sont quantitatives, on associe souvent à l'analyse factorielle le modèle général suivant :

$$X = L \times S + R$$

où X représente la variable multidimensionnelle observée, S la matrice des facteurs (appelée aussi matrice des "factor scores"), L la matrice des composantes des individus dans la base des facteurs (ou matrice des "factor loadings") et R la variable résiduelle. Le problème est alors d'estimer les paramètres.

Plusieurs approches permettant d'étendre l'analyse factorielle, terme ici employé au sens large (recouvrant à la fois l'analyse factorielle proprement dite et l'analyse en principales), ont été envisagées pour les données binaires.

Une première approche consiste à calculer une matrice de corrélations (appelées "tetrachoric correlations") et à appliquer ensuite l'analyse factorielle habituelle. Mais un problème peut se poser car la matrice ainsi calculée n'est pas nécessairement définie positive. Pour cette raison, d'autres alternatives ont été proposées (R.D. Bock et M. Liebermann 1970, A. Christofferson 1975, B. Muthen 1978, B.S. Everitt 1984). Ces derniers considèrent que la variable multidimensionnelle X (la variable observée) est issue d'une variable normale X* (variable latente) de moyenne nulle suivant un shéma précis. Ils supposent alors que cette variable multidimensionnelle X* suit le modèle habituel de l'analyse factorielle. Notons également les travaux de D.J. Bartholomew (1980 et 1984) considérant une approche plus générale pour les variables qualitatives qui inclut les modèles précédents.

M.R. Mickey, P. Mundel et L. Engelman (1984) ont une approche particulière du problème qui consite à rester "proche" des données binaires initiales. Ils reprennent le modèle habituel de l'analyse factorielle en considérant les facteurs comme des vecteurs binaires et les "factors loadings" comme des valeurs binaires. Les opérateurs arithmétiques utilisés sont les operateurs booléens "et" et "ou". Le problème consiste alors à déterminer les matrices L et S de sorte que l'écart entre le tableau X et le produit LxS soit minimum. Cette méthode est appelée analyse factorielle booléenne et optimise un critère équivalent à celui optimisé par la méthode MNDBIN. Dans ce chapitre, nous reviendrons sur cette méthode présentée dans le cadre du logiciel BMDP.

Une autre approche est de calculer une matrice de distance et à appliquer ensuite une méthode de type "multidimensionnal scaling" (L. Teledgi 1986).

Citons enfin B. Fichet et A. Gbégan (1985) qui utilisent une dissimilarité euclidienne déduite de la similarité d'Ochiaï. L'analyse proposée conduit à des formules de transition et de reconstitution des données.

En analyse en composantes principales, les données binaires peuvent être traitées comme des données quantitatives et, en analyse des correspondances, comme des données qualitatives. On se place alors dans un espace du type $\mathbf{R}^p$ que l'on munit d'une métrique euclidienne. Dans les deux cas, on ne tient pas compte de la forme particulière des données initiales. Les axes factoriels déterminés de cette façon ne sont pas directement interprétables par rapport aux données initiales.

Dans ce chapitre, nous présentons une méthode d'analyse en composantes principales spécifique à ce type de données. Pour cela, nous nous plaçons dans l'espace binaire $\mathbf{B}^p$ que l'on munit de la distance en valeurs absolues. Comme pour la méthode de classification MNDBIN, nous imposons à la future analyse de respecter la structure initiale des données.

Dans un premier temps, nous complétons l'étude de l'espace $\mathbf{B}^p$. Il nous reste à définir les notions nécessaires à la formulation du problème. Pour cela, nous suivons une démarche simple : nous reprenons, dans le cadre de l'espace $\mathbf{B}^p$ (muni de la distance en valeurs absolues), les notions habituellement définies sur l'espace $\mathbf{R}^p$ (muni d'une métrique euclidienne). Nous définissons les notions d'axe binaire, d'orthogonalité, de projection sur un système d'axes binaires et d'inertie d'un nuage par rapport au sous-espace engendré par un système d'axes binaires. Bien sûr, cette analogie va comporter des limites, limites qui vont influencer la formulation du problème d'analyse en composantes principales sur données binaires. Ces limites sont dues au fait que l'espace $\mathbf{B}^p$ n'a pas une structure d'espace vectoriel.

Après ce travail, nous proposons une méthode d'analyse en composantes principales sur tableau de variables binaires. Celle-ci est présentée sous une forme habituelle, le but étant de rechercher un ensemble restreint d'axes binaires ou axes factoriels binaires réprésentant le mieux possible le nuage des individus (ou des variables). Comme dans le cas quantitatif, l'ensemble recherché est solution d'un problème d'optimisation d'un critère : l'inertie du nuage des individus par rapport au sous-espace engendré par un système d'axes.

Cependant, il subsiste une différence fondamentale avec la méthode existante. Les propiétés de l'espace $\mathbf{B}^p$, mises en évidence dans ce chapitre, nous obligent à fixer, a priori, le nombre d'axes factoriels à rechercher. Sous cette hypothèse, deux approches sont envisagées. La première consiste à rechercher un ensemble d'axes orthogonaux (au sens où nous l'avons défini). Nous montrons alors le lien entre la recherche d'axes factoriels et la recherche d'une partition de l'ensemble des variables. La contrainte imposée au système d'axes facilite la résolution du problème de recherche. La seconde approche consiste à rechercher un ensemble d'axes, non plus orthogonaux, mais quelconques. Comme nous le verrons, la suppression de la contrainte permet d'améliorer la qualité des résultats. La résolution d'un tel problème s'avère cependant plus délicate. Nous avons construit un algorithme, appelé ACPBIN, qui fournit seulement une solution approximative à ce problème. Nous montrons ensuite que cette seconde approche est très voisine de la méthode "d'analyse factorielle booléenne" ou AFB (M.R. Mickey, P. Mundel, L. Engelman 1983).

Dans un nouveau paragraphe, nous reprenons toutes les méthodes (de classification et d'analyse factorielle) pour tableaux de variables binaires que nous associons à un modèle matriciel. Ce modèle permet d'illustrer les différences et les similitudes entre toutes ces méthodes. Il sera alors possible d'effectuer un bilan pour les méthodes pour tableaux de variables binaires proposées dans cette étude.

Enfin, une application sur un tableau binaire est proposée. Les résultats fournis par les méthodes MNDBIN et ACPBIN sont détaillés et interprétés.

2. NOTATIONS

2.1 LES DONNÉES

Les données étudiées sont celles d'un tableau **X**(I,J) croisant un ensemble I={1,...,n} de n individus et un ensemble J={1,...,p} de p variables binaires. On note :

$\mathbf{X}(I,J) = (x_i^j)$ le tableau de données binaires,

$\alpha(I,J) = (\alpha_i^j)$ le tableau de pondérations,

où α_i^j représente la pondération associée à x_i^j.

Nous nous plaçons ici dans le cas le plus général où les pondérations initiales sont quelconques.

A partir des lignes et des colonnes de ces deux tableaux, on définit le nuage des individus **N**(I) et le nuage **N**(J) des variables par :

$$\mathbf{N}(I) = \{ (x_i, \alpha_i), i \in I \} \qquad \mathbf{N}(J) = \{ (x^j, \alpha^j), j \in J \}$$

2.2 L'ESPACE

Les données sont plongées dans un espace du type $\mathbf{B}^p$ qui sera, ici, le seul espace de référence envisagé. D'un point de vue géométrique, nous pouvons considérer que cet espace est la restriction de $\mathbf{R}^p$ à "l'hypercube" dont les sommets sont les points de coordonnées toutes égales à 0 ou à 1. Il s'agit donc d'un espace fini contenant exactement 2^p points.

Chaque point du nuage des individus **N**(I) est alors représenté par l'un de ces sommets. Pour des valeurs élevées de p, il devient difficile de visualiser un tel nuage. Aussi, en suivant le principe de l'analyse en composantes principales, nous allons définir une représentation visuelle plus simple, en tenant compte de la forme initiale des données. Nous essayons de représenter le nuage des individus dans un espace du type $\mathbf{B}^M$ avec M<<p, en respectant le mieux possible les proximités entre individus, ces proximités étant, ici encore, mesurées par la distance en valeurs absolues.

En fait, nous essayons de construire un ensemble de M nouvelles variables binaires résumant le mieux possible les p variables initiales.

2.3 VECTEURS BINAIRES ET OPÉRATIONS

A chaque point de $\mathbf{B}^p$, nous associons un **vecteur binaire**. C'est à l'ensemble de ces vecteurs, au nombre de 2^p, que nous nous intéressons ici. Nous nous limitons aux vecteurs joignant l'origine (notée O) de $\mathbf{R}^p$ à un sommet ou point de l'espace. Nous reviendrons dans la suite sur cette restriction.

A chaque point de $\mathbf{B}^p$ correspond donc un vecteur binaire. Ces deux notions étant très liées, nous utilisons, dans la suite, une notation unique pour désigner à la fois le point et le vecteur correspondant. Il nous reste maintenant à définir des opérations sur ces vecteurs binaires, opérations qui doivent nous permettre de rester en accord avec la forme des données.

Somme de deux vecteurs binaires
La somme repose sur la fonction logique "ou" sur l'ensemble {0,1}. On utilise le symbole "+" pour représenter cette opération. Par exemple, dans $\mathbf{B}^4$, la somme des vecteurs u=(1,1,0,0) et v=(0,1,0,1) est le vecteur u+v=(1,1,0,1).

Produit par un scalaire binaire
Nous proposons d'utiliser la fonction logique "et" pour définir le produit d'un vecteur par un scalaire binaire. On note simplement a.v ou av le produit du vecteur u par le scalaire a. On obtient alors les résultats attendus : le produit d'un vecteur v par le scalaire 1 est v, le produit de v par le scalaire 0 est l'origine O de l'espace.

2.4 NOTION DE BASE POUR L'ESPACE BINAIRE

L'espace $\mathbf{B}^p$ muni des opérateurs "et" et "ou" n'a pas une structure d'espace vectoriel. Cependant, dans la suite, nous parlerons abusivement de "base" pour cet espace ou pour des "sous-espaces" de $\mathbf{B}^p$.

La base cononique usuelle $(e_1,\ldots,e_p)$ de $\mathbf{R}^p$ permet d'atteindre tous les points de $\mathbf{B}^p$. Nous dirons alors que le système de vecteurs binaires $(e_1,\ldots,e_p)$ constitue une "base" de l'espace $\mathbf{B}^p$: tout vecteur binaire peut s'écrire comme une combinaison linéaire, à coefficients dans {0,1}, des vecteurs de bases. En utilisant les opérateurs logiques présentés dans le paragraphe précédent, on peut écrire :

$$\forall\, u \in \mathbf{B}^p \quad u = u^1 e_1 + u^2 e_2 + \cdots + u^p e_p = \sum_{j \in J} u^j e_j$$

$$\text{où} \quad \forall\, j \in J \quad u^j \in \{0,1\}$$

Remarquons qu'une combinaison linéaire de vecteurs binaires est également un vecteur binaire.

3. VECTEURS BINAIRES ET SOUS-ESPACES BINAIRES

3.1 VECTEUR BINAIRE ET SOUS-ENSEMBLE ASSOCIÉ

Soit J={1,...,p} l'ensemble des indices des vecteurs de la base $(e_1,\ldots,e_p)$ de $\mathbf{B}^p$.

Définition
Soit A une partie de J. Nous définissons le vecteur u_A associé à la partie A par :

$$\forall j \in A \quad u_A^j = \begin{cases} 1 \text{ si } j \in A \\ 0 \text{ sinon} \end{cases}$$

De la même manière, à un vecteur de $\mathbf{B}^p$ on peut associer une partie A de J. Si J représente l'ensemble des variables, la partie A associée au vecteur réponse d'un individu contient les variables auxquelles l'individu a répondu par 1. Compte tenu de

cette définition, à une combinaison linéaire des caractères initiaux va donc correspondre un sous-ensemble de variables.

3.2 AXE BINAIRE

Définition
Soit Δ un axe de $\mathbf{R}^p$ passant par l'origine. Nous appelons ***axe binaire*** *l'ensemble des points intersection de Δ et de $\mathbf{B}^p$.*

Dans la suite, nous nous intéressons uniquement aux axes binaires contenant tous un point commun, par exemple l'origine du repère (on peut toujours se ramener à ce cas de figure par l'intermédiaire d'un changement d'origine). Cette restriction va s'avérer suffisante lorsque nous aborderons l'analyse en composantes principales. D'ailleurs, la méthode habituelle fournit, dans $\mathbf{R}^p$, un système d'axes factoriels sécants en un même point (l'origine du repère ou le centre de gravité du nuage des individus).

En suivant cette définition, un axe binaire apparait alors comme un ensemble de deux points : l'origine et un point de $\mathbf{B}^p$. En conséquence, à chaque partie A de J correspond un vecteur u_A engendrant l'axe Δ_A défini par :

$$\Delta_A = \{O, u_A\}$$

La notation u_A représente à la fois le vecteur de base et le point de l'axe (ils sont tous deux associés à la partie A de J).

3.3 AXES ORTHOGONAUX

Définition
Soient A et B deux sous-ensemble de J et $\Delta_A=\{O,u_A\}$, $\Delta_B=\{O,u_B\}$ les axes associés. Nous dirons que ces deux axes sont orthogonaux sur $\mathbf{B}^p$ si et seulement si ils le sont sur $\mathbf{R}^p$.

Par conséquent, deux axes Δ_A et Δ_B sont donc orthogonaux si et seulement si les deux parties associées sont disjointes :

$$\Delta_A \text{et } \Delta_B \text{ orthogonaux} \quad \Leftrightarrow \quad A \cap B = \emptyset$$

3.4 SYSTEME D'AXES BINAIRES ET SOUS-ESPACE ENGENDRÉ

Considérons maintenant un ensemble $Q=(Q^1,\ldots,Q^M)$ de M parties de J. A chacune de ces parties est associé un axe binaire défini par :

$$\forall\, m=1,\ldots,M \quad \Delta_m = \{O,u_m\}$$

Définition
*Soit **E** l'ensemble des points s'exprimant comme une combinaison linéaire des M vecteurs $(u_1,\ldots,u_M)$. Nous dirons que **E** est le sous-espace engendré par le système $(u_1,\ldots,u_M)$ ou, plus simplement, par Q.*

Le système $(e_1,\ldots,e_p)$ est le seul système permettant d'atteindre, de façon unique, tous les points de $\mathbf{B}^p$. Nous dirons alors abusivement que ce système est une base de $\mathbf{B}^p$ (qui n'a pas de structure d'espace vectoriel).

De même, nous dirons que le système $(u_1,\ldots,u_M)$ engendrant le sous-espace **E** est une base si et seulement si ce sytème permet de définir 2^M points distincts. Ainsi, tout point de **E** s'exprime comme un M-uplets à valeurs dans {0,1}. Le sous-espace **E** apparait alors comme un espace du type $\mathbf{B}^M$. Si le système d'axes est orthogonal, il constitue évidemment une base de **E**. Si le système est quelconque, cela n'est pas toujours vrai.

D'après les hypothèse faites jusqu'ici, il apparait que la notions de sous-espaces supplémentaires n'existe pas sur $\mathbf{B}^p$. Soit, par exemple, le sous-espace **E** engendré par les vecteurs $u_1=(1,1,0)$ et $u_2=(0,0,1)$. Il est impossible de trouver un troisième vecteur u_3 tel que le système (u_1,u_2,u_3) soit une base de $\mathbf{B}^3$.

Dans la suite, nous allons définir les notions de projection et d'inertie d'un nuage de points par rapport à un sous-espace **E** de $\mathbf{B}^p$. Deux études sont proposées. La première consiste à rechercher un système d'axes orthogonaux dans lequel le nuage soit le mieux possible représenté. La seconde approche consiste, cette fois, à rechercher un système d'axes quelconques.

4. PROJECTION SUR UN SOUS-ESPACE BINAIRE

4.1 PROJECTION SUR UN AXE BINAIRE

Soient A une partie de J, u_A l'axe associé et $\Delta_A=\{O,u_A\}$ l'axe engendré. Pour mesurer la distance d'un point x de $\mathbf{B}^p$ à un point a de l'axe Δ_A, nous proposons d'utiliser une forme pondérée de la distance en valeurs absolues. Elle est notée **D** et définie par :

$$\mathbf{D}(x,a) = \Im_a(\{x\}) = \sum_{j\in J} \alpha^j(x) \mid x^j - a^j \mid$$

où $\alpha(x)$ est le vecteur de pondérations associé à x. Pour des pondérations toutes égales à 1, on retrouve simplement la distance en valeurs absolues.

Définition
*Nous définissons la projection d'un point x sur l'axe Δ_A comme étant le point O ou u_A de l'axe le plus proche de x au sens de la distance **D**, soit :*

$$pr(x/\Delta_A) = \begin{cases} O \text{ si } D(x,O) \le D(x,u_A) \\ u_A \text{ sinon} \end{cases}$$

Le projeté de x n'a qu'une seule coordonnée si on l'exprime dans le repère (O,u_A) de l'axe. La projection apparait alors comme une application de l'ensemble $\mathbf{B}^p$ vers l'ensemble {0,1}. A tout point de $\mathbf{B}^p$, elle associe l'une des valeurs 0 ou 1. Une interprétation très simple peut être donnée à cette valeur binaire.

Propriété
Soit x un point de $\mathbf{B}^p$ et $\alpha(x)$ le vecteur de pondérations associé. La projection de x sur l'axe Δ_A ne dépend que des composantes j appartenant à A et :

$$pr(x/\Delta_A) = au_A$$

$$ou \quad a = médiane\ binaire\ de\ \{ (x^j, \alpha^j(x)), j\in A \}$$

Preuve
Le projeté de x étant un point de l'axe, on peut écrire :

$$\mathbf{pr}(x/\Delta_A) = au_A$$

$$\text{avec} \quad a \in \{0,1\}$$

Le problème est alors de rechercher la valeur a minimisant la distance $\mathbf{D}(x,au_A)$ définie par :

$$\mathbf{D}(x,au_A) = \sum_{j\in J} \alpha^j(x)|x^j - au_A|$$

$$\Leftrightarrow \quad \mathbf{D}(x,au_A) = \sum_{j\in A} \alpha^j(x)|x^j - a| + \sum_{j\in J-A} \alpha^j(x)x^j$$

Il suffit alors de trouver la valeur a minimisant la quantité :

$$\sum_{j\in A} \alpha^j(x)|x^j - a|$$

La solution est de choisir pour a la valeur médiane de l'ensemble $\{(x^j, \alpha^j(x)), j\in A\}$.

Pondération associée au projeté d'un point
Un point x de $\mathbf{B}^p$ est muni d'un vecteur de pondérations de $\mathbf{R}^{p+}$. Soit a l'image de x par la projection, a est alors une simple valeur binaire. Le dernier problème à résoudre est donc celui de la pondération (appartenant à $\mathbf{R}^+$) à associer à l'image a de x. En suivant notre approche, il apparait naturel de retenir, pour a, la pondération associée à la valeur médiane qu'elle représente. La valeur a et sa pondération se déduisent alors du point x et de son vecteur de pondérations et :

$$a = \text{médiane de } \{ (x^j, \alpha^j(x)), j\in A \}$$

$$\alpha(a) = \left| \sum_{j\in A} \alpha^j(x)x^j - \sum_{j\in A} \alpha^j(x)(1-x^j) \right|$$

La projection apparait alors comme une application qui à tout couple $(x,\alpha(x))$ de $(\mathbf{B}^p,\mathbf{R}^{p+})$ associe un couple $(a,\alpha(a))$ de $(\{0,1\},\mathbf{R}^+)$.

La pondération permet en outre de mesurer la qualité de l'image d'un point. Elle varie de 0 (l'individu est mal représenté sur l'axe) à Card(A) (l'individu est bien représenté sur l'axe).

Exemples
Considérons l'axe de $\mathbf{B}^7$ passant par u=(1,0,1,1,1,0,1). En considérant des points munis de pondérations toutes égales à 1, on a :

- l'image de (1,1,1,1,0,0,1) est la valeur 1 de pondération 3,
- l'image de (1,1,1,1,0,0,0) est la valeur 1 de pondération 1,
- l'image de (0,1,0,0,0,0,0) est la valeur 0 de pondération 5.

4.2 IMAGE D'UN NUAGE PAR LA PROJECTION SUR UN AXE

Nous nous intéressons maintenant à l'image du nuage des individus $\mathbf{N}(I)$ par la projection sur un axe Δ_A. Notons $\mathbf{N}(I/\Delta_A)$ le nuage projeté sur l'axe. Dans $\mathbf{B}^p$, ce nuage est constitué de n points dont une partie est égale à l'origine O et l'autre au point u_A (puisque le projeté ne peut être que l'un de ces deux points).

Si on se place sur l'axe Δ_A de vecteur de base u_A, le nuage $\mathbf{N}(I/\Delta_A)$ apparait alors comme un ensemble de n valeurs binaires pondérées définies par :

$$\mathbf{N}(I/\Delta_A) = \{ (a_i, \alpha(a_i)), i \in I \}$$

où, pour tout i, on a :

$$\mathbf{pr}(x_i/\Delta_A) = a_i u_A$$

$$\text{avec} \quad a_i = \text{médiane de } \{ (x_i^j, \alpha_i^j), j \in A \}$$

$$\text{et} \quad \alpha(a_i) = \left| \sum_{j \in A} \alpha_i^j x_i^j - \sum_{j \in A} \alpha_i^j (1-x_i^j) \right|$$

Propriété
Si on note

λ le centre médian de N(I),

λ_A l'image λ de par la projection sur l'axe Δ_A,

alors

λ_A est la valeur médiane du nuage projeté $N(I/\Delta_A)$.

Preuve
Appelons a la médiane de l'ensemble des valeurs binaires contenues dans le nuage $\mathbf{N}(I/\Delta_A)$. Celle-ci est définie par :

$$a = \text{médiane de } \{ (a_i, \alpha(a_i)), i \in I \}$$

$$\alpha(a) = \left| \sum_{i \in I} \alpha(a_i) a_i - \sum_{i \in I} \alpha(a_i)(1-a_i) \right|$$

D'après la propriété de conservation de la médiane, on a :

$$a = \text{médiane de } \{ (x_i^j, \alpha_i^j), j \in A \text{ et } i \in I \}$$

$$\alpha(a) = \left| \sum_{i \in I} \sum_{j \in A} \alpha_i^j x_i^j - \sum_{i \in I} \sum_{j \in A} \alpha_i^j (1-x_i^j) \right|$$

Le centre médian λ de $\mathbf{N}(I)$ est défini par :

$$\forall j \in J \quad \lambda^j = \text{médiane de } \{ (x_i^j, \alpha_i^j), i \in I \}$$

$$\text{et} \quad \alpha^j(\lambda) = \left| \sum_{i \in I} \alpha_i^j x_i^j - \sum_{i \in I} \alpha_i^j (1-x_i^j) \right|$$

La projection de λ sur Δ_A est la valeur λ_A définie par :

$$\lambda_A = \text{médiane de } \{ (\lambda^j, \alpha^j(\lambda)), j \in A \}$$

$$\alpha(\lambda_A) = | \sum_{j \in J} \alpha^j(\lambda)\lambda^j - \sum_{j \in J} \alpha^j(\lambda)(1-\lambda^j) |$$

Et, d'après la propriété de conservation de la médiane, on a :

$$\lambda_A = \text{médiane de } \{ (x_i^j, \alpha_i^j), j \in A \text{ et } i \in I \}$$

$$\alpha(\lambda_A) = | \sum_{i \in I} \sum_{j \in A} \alpha_i^j x_i^j - \sum_{i \in I} \sum_{j \in A} \alpha_i^j (1-x_i^j) |$$

Cela montre l'égalité entre λ_A (le projeté du centre médian du nuage **N**(I)) et a (la valeur médiane du nuage projeté **N**(I/Δ_A)).

Remarque
La projection sur l'axe Δ_A permet de définir une dichotomie de l'ensemble I des individus. Celle-ci est définie par :

$$I = I_0 \cup I_A$$

où I_0 est l'ensemble des individus se projetant sur l'origine,

et I_A est l'ensemble des individus se projetant sur le point u_A.

Le nuage **N**(I) de $\mathbf{B}^p$ est défini à partir du tableau à n lignes et p colonnes **X**(I,J). Après projection sur l'axe associé à une partie A de J, nous obtenons un nuage projeté contenant n valeurs binaires. Nous pouvons alors construire le tableau **X**(I,A) à n lignes et une colonne résumant ce nouveau nuage.

Le tableau **X**(I,A) n'est autre que celui construit à partir du couple (I,A) (comme indiqué dans le chapitre 5, paragraphe 4.4). Il en va de même pour le tableau de pondérations associé.

Si on réordonne les lignes de ce tableau en respectant la dichotomie de I, on obtient simplement :

	A
I_0	0
I_A	1

où la valeur 0 représente exactement la médiane de l'ensemble des valeurs du sous-tableau **X**(I_0,A), et la valeur 1 la médiane des valeurs du sous-tableau **X**(I_A,A).

4.3 PROJECTION SUR UN SYSTEME D'AXES

Soit $Q=(Q^1,\ldots,Q^M)$ un ensemble de M parties de J et **E** l'espace engendré par les M axes associés définis, pour tout m, par $\Delta_m=\{O,u_m\}$.

Définition
*Nous définissons la projection d'un point x sur **E** comme étant le point de **E** le plus proche de x au sens de la distance **D**.*

La projection apparait alors comme une application de $\mathbf{B}^P$ vers **E** : à tout point x de $\mathbf{B}^P$, elle associe un point x_E de **E** ou $\mathbf{B}^M$.

Deux cas se présentent alors : soit les M parties de Q sont disjointes, soit elles sont quelconques. Dans les paragraphes qui suivent, nous étudions en détail ces deux approches.

4.4 CAS D'UN SYSTEME D'AXES ORTHOGONAUX

Lorsque les M parties de Q sont deux à deux disjointes, le système d'axes associé est orthogonal. Comme nous l'avons dit, la notion de sous-espaces supplémentaires n'existe pas sur $\mathbf{B}^P$. Pour cette raison, nous proposons de choisir les parties de Q de sorte qu'elles forment une partition de l'ensemble J. Ainsi, chaque vecteur de la base canonique est pris en compte par l'un des vecteurs de base $(u_1,\ldots,u_M)$ de **E**. C'est la position que nous adoptons dans ce paragraphe.

4.4.1 Propriété

Le projeté d'un point s'exprime comme un M-uplets par rapport au système $(u_1,\ldots,u_M)$. La propriété ci-dessous nous montre comment le calculer.

Propriété
*Soient x un point de $\mathbf{B}^P$ et $\alpha(x)$ son vecteur de pondérations. Le projeté de x sur **E** est le point x_E de **E** défini par :*

$$pr(x/E) = x_E$$

$$où \quad x_E = \sum_{m=1}^{M} x_E^m u_m = \sum_{m=1}^{M} pr(x/\Delta_m)$$

$$et \quad \forall m=1,\ldots,M \quad x_E^m = médiane\ binaire\ de\ \{ (x^j, \alpha^j(x)), j \in Q^m \}$$

Preuve
Le projeté x_E de x s'écrit comme une combinaison linéaire à coefficients dans {0,1} des vecteurs $(u_1,\ldots,u_M)$, de sorte que :

$$x_E = \sum_{m=1}^{M} a^m u_m$$

Le problème est alors de rechercher les M coordonnées $(a^1,\ldots,a^M)$ de x_E minimisant la distance $D(x,x_E)$ s'exprimant ici par :

$$D(x,x_E) = \sum_{j \in J} \alpha^j(x) |x^j - (a^1 u_1^j + a^2 u_2^j + \cdots + a^M u_M^j)|$$

$$\Leftrightarrow \quad D(x,x_E) = \sum_{m=1}^{M} \sum_{j \in Q^m} \alpha^j(x) |x^j - a^m|$$

Pour déterminer le point x_E, il suffit donc de trouver, pour tout m, la valeur a^m minimisant :

$$\sum_{j \in Q^m} \alpha^j(x) |x^j - a^m|$$

La solution évidente est de choisir pour a^m la valeur médiane de $\{(x^j, \alpha^j(x)), j \in Q^m\}$.

4.4.2 Vecteur de pondérations associé au projeté d'un point

Un point x de $\mathbf{B}^p$ est muni d'un vecteur de pondérations de $\mathbf{R}^{p+}$. Soit x_E appartenant à **E** l'image de x par la projection. Le point x_E appartenant à l'espace **E** ou $\mathbf{B}^M$, il sera muni d'un vecteur de pondérations appartenant à $\mathbf{R}^{M+}$.

Chaque composante m de x_E représentant la valeur médiane d'un ensemble (propriété précédente), il est naturel de choisir pour celle-ci la pondération correspondant à cette médiane. On obtient ainsi un vecteur de pondérations appartenant à $\mathbf{R}^{M+}$.

Le point x_E et son vecteur de pondérations se déduisent alors du point x et du vecteur $\alpha(x)$ de la façon suivante :

$$x_E = (x_E^1, x_E^2, \ldots, x_E^M)$$

$$\forall\, m=1,\ldots,M \quad x_E^m = \text{médiane de } \{(x^j, \alpha^j(x)), j \in Q^m\}$$

$$\text{et} \quad \alpha^m(x_E) = \left| \sum_{j \in Q^m} \alpha^j(x) x^j - \sum_{j \in Q^m} \alpha^j(x)(1-x^j) \right|$$

La projection apparait ainsi comme une application qui à tout couple $(x,\alpha(x))$ de $(\mathbf{B}^p,\mathbf{R}^{p+})$ associe le couple $(x_E,\alpha(x_E))$ de $(\mathbf{B}^M,\mathbf{R}^{M+})$.

De nouveau, le vecteur de pondérations x_E fournit une indication sur la qualité de représentation de x (appartenant à $\mathbf{B}^p$) dans un espace de dimension inférieure ($\mathbf{B}^M$ avec $M<<p$). Chaque composante m (variant de 0 à $Card(Q^m)$) de ce vecteur est une mesure de la qualité de représentation de x sur l'axe Δ_m correspondant.

Dans un prochain paragraphe, nous proposons de nouveaux indices à partir de la notion d'inertie (qui reste à définir) du nuage par rapport au sous-espace **E**.

4.4.3 Projection du nuage des individus

Pour tout i, on note a_i l'image du point x_i de **N(I)** par la projection sur **E** et :

$$\forall m=1,\ldots, M \quad a_i^m = \text{médiane de } \{ (x_i^j, \alpha_i^j), j \in Q^m\}$$

Le projeté a_i peut s'exprimer comme un M-uplet de $\mathbf{B}^M$. Si on note $\alpha(a_i)$ le vecteur de pondération associé, il appartient à $\mathbf{R}^{M+}$ et :

$$\forall m=1,\ldots, M \quad \alpha^m(a_i) = \left| \sum_{j \in Q^m} \alpha_i^j x_i^j - \sum_{j \in Q^m} \alpha_i^j (1-x_i^j) \right|$$

Nous pouvons alors définir le nuage projeté **N(I/E)** qui, exprimé dans l'espace $\mathbf{B}^M$, s'écrit :

$$\mathbf{N(I/E)} = \{ (a_i, \alpha(a_i), i \in I \}$$

Propriété
Si on note

λ, inclus dans $\mathbf{B}^p$, le centre médian du nuage N(I),

λ_E, inclus dans $\mathbf{B}^M$, le centre médian du nuage projeté N(I/E),

alors

λ_E est l'image de λ par la projection.

Preuve

D'après la propriété précédente, le point λ_E s'obtient en faisant la somme des projections du centre médian de N(I) sur les différents axes. Pour démontrer la propriété, il suffit alors de reprendre et de compléter la démonstration du paragraphe 4.2 concernant la projection du centre médian du nuage sur un seul axe.

4.4.4 Remarques

Le nuage N(I) est défini à partir du tableau initial **X**(I,J) à n lignes et p colonnes. Si Q est une partition de J, nous savons construire le tableau **X**(I,Q) associé au couple (I,Q) (comme indiqué dans le chapitre 5, paragraphe 4.4). Le nuage **N**(I/**E**) apparait alors comme le nuage défini à partir des lignes de **X**(I,Q). Notons également que le nuage défini à partir des colonnes de **X**(I,Q) n'est autre que le nuage des centres médians des classes de la partition Q.

La projection permet de représenter le nuage des individus dans un espace $\mathbf{B}^M$ de dimension inférieure à $\mathbf{B}^p$. Tout ce passe comme si nous avions défini M nouvelles variables binaires, chacune étant la variable résumée d'un sous-ensemble de variables initiales, chaque variable initiale n'étant présente que dans une seule classe.

4.5 CAS D'UN SYSTEME D'AXES QUELCONQUES

Nous considérons ici que les M parties de Q sont quelconques. Cela revient à considérer un système d'axes sans contrainte d'orthogonalité. Pour les mêmes raisons que précédemment, nous supposons que tout élément de J est contenu dans au moins une partie de Q. La réunion de ces parties est donc égale à J et Q apparait alors comme un recouvrement de J. C'est dans ce contexte que nous allons étudier la projection d'un point sur l'espace **E** engendré par Q.

4.5.1 Le problème de la projection

La recherche de l'image d'un point x de $\mathbf{B}^p$ par la projection sur **E** s'avère plus délicate que précédemment. En effet, la propriété permettant de construire cette image dans le cas d'axes orthogonaux n'est plus vérifiée ici. Considérons, par exemple, les trois axes binaires de $\mathbf{B}^5$ engendrés par les vecteurs :

$$u_1 = (1, 1, 1, 0) \quad u_2 = (1, 0, 0, 1) \quad u_2 = (0, 1, 0, 1)$$

Soit le point x=(1,1,0,1) muni de pondérations toutes égales à 1. Ce point appartient à l'espace engendré par les trois axes puisque $x=u_2+u_3$. Quant à la propriété permettant de

calculer le projeté lorsque le système est orthogonal, elle fournit ici le point (1,1,1,1). En effet, si on considère les projections sur les différents axes, le point x se projette sur u_1, u_2 et u_3. D'après la propriété, le projeté est alors le point $u_1+u_2+u_3$.

Il se pose donc le problème de la recherche du point de **E** le plus proche de x au sens de la distance **D**. Le problème du choix du vecteur de pondérations va de pair, d'autant plus que la notion de vecteur de pondérations, utilisée jusqu'ici, ne peut plus s'appliquer (même si nous savons calculer exactement l'image d'un point par la projection). En effet, dans l'exemple ci-dessus, le projeté a pour composantes (1,1,0) dans (u_1,u_2,u_3) et nous ne savons pas quelle pondération associer à la valeur 0.

L'espace **E** étant fini, il est possible de calculer les distances entre le point x et tous les points de **E**. Cependant, ce procédé devient coûteux dès que le nombre d'axes est important.

4.5.2 Un algorithme

Nous proposons ici un algorithme fournissant, de façon approchée, le projeté x_E d'un point x sur **E**. Celui-ci recherche, à chaque étape, l'une des M coordonnées de x_E. Plus exactement, il recherche successivement toutes les coordonnées égales à 1 de ce point. Nous décrivons tout d'abord cet algorithme sur lequel nous reviendrons dans la suite. Il procède de la façon suivante :

a) initialisation :

- on calcule $d^0 = \mathbf{D}(x,O)$, la distance entre x et l'origine de $\mathbf{B}^P$.
- on pose $P = \emptyset$, l'ensemble des indices des coordonnées de x_E égales à 1.

b) étape r :

- on cherche l'indice j n'appartenant pas à P et tel que :

$$d^r = \mathbf{D}(x, u_j + \sum_{k \in P} u_k) \text{ minimum}$$

- si $d^0 \leq d^1$ alors $x_E = O$.

  Lors de la première étape, si le point x est plus proche de l'origine que de tout autre point $(u_1,\ldots,u_L)$ de **E**, il est inutile de poursuivre l'algorithme, la projection de x est l'origine O de l'espace.

- si $d^r \geq d^{r-1}$ alors $x_E = \sum_{k \in P} u_k$ sinon $P = P \cup \{j\}$.

  Lors d'une étape r, soit il n'y a aucune amélioration, on arrête alors l'algorithme en gardant la solution de l'étape précédente, soit on trouve une amélioration, on ajoute alors l'indice j à l'ensemble P.

Pour un système $(u_1,\ldots,u_M)$ constituant une base du sous-espace de projection, l'algorithme fournit la bonne solution. Dans ce cas, il n'est pas nécessaire d'utiliser cet algorithme puisque le nuage projeté est simplement défini à partir du tableau associé au couple (I,Q) où Q est une partition de J engendrant le système d'axes. Par contre, pour des systèmes d'axes quelconques, la solution fournie n'est qu'approximative. Nous n'avons pu construire un algorithme fournissant le projeté exact d'un point.

Dans la pratique, les résulats sont acceptables si le système d'axes n'est pas engendré par des parties incluses les unes dans les autres. Dans la suite, nous utiliserons tout de même celui-ci pour construire un algorithme de recherche d'un système d'axes quelconques minimisant l'inertie (qui reste à définir) d'un nuage par rapport au sous-espace engendré par ce sytème. Ci-dessous, nous donnons des exemples montrant les possibilités et les limites de l'algorithme de recherche du projeté d'un point.

Exemple 1
Nous présentons ici un exemple où le système d'axes n'est pas orthogonal et où le point, dont on cherche le projeté, appartient déja au sous-espace de projection. L'algorithme fournit ici la bonne solution.

Soit l'espace **E** engendré par les vecteurs :

$$u_1 = (1, 1, 0, 1, 0) \quad u_2 = (1, 0, 0, 0, 1) \quad u_3 = (0, 1, 0, 0, 1)$$

On cherche le projeté x_E du point x=(1,1,0,0,1). On suppose que toutes les composantes du vecteur de pondérations de x sont égales. L'algorithme fournit alors les résultats suivants :

- étape 1 : $x_E = u_2$
- étape 2 : $x_E = u_2 + u_3 = x$

Exemple 2
Ici, on présente un exemple montrant les limites de l'algorithme. Bien que le point considéré appartienne déjà au sous-espace de projection, l'algorithme ne permet pas de le retrouver.

Soit **E** l'espace engendré par les vecteurs :

$$u_1 = (1, 1, 0, 0, 0) \quad u_2 = (0, 0, 0, 1, 1) \quad u_3 = (1, 1, 1, 1, 1)$$

On cherche le projeté x_E du point x=(1,1,0,1,1). On suppose encore que toutes les composantes du vecteur de pondérations de x sont égales. De manière évidente, le point x appartient à l'espace **E** (puisque $x=u_1+u_2$), mais l'algorithme fournit, après 2 étapes, le point projeté $x_E=u_3+u_1=(1,1,1,1,1)$.

4.5.3 Remarque

Considérons le nuage **N**(I) défini à partir du tableau **X**(I,J). Il appartient à $\mathbf{B}^p$ et le nuage projeté sur le sous-espace (engendré par un recouvrement Q en M classes de) est inclus dans $\mathbf{B}^M$. Tout se passe comme si nous avions défini M nouvelles variables binaires, chacune étant associée à une partie de Q. Mais cette fois, une variable de J peut appartenir à plusieurs parties du recouvrement Q.

5. INERTIE PAR RAPPORT À UN SOUS-ESPACE BINAIRE

5.1 INERTIE D'UN NUAGE PAR RAPPORT À UN AXE

5.1.1 Définition

Soit $\Delta_A=\{0,u_A\}$ l'axe associé à la partie A de J. Nous avons vu que la projection sur cet axe permet de définir le nuage projeté $N(I/\Delta_A)$ comme un ensemble de n valeurs

$$N(I/\Delta_A) = \{ (a_i, \alpha(a_i)), i\in I \}$$

où

$$\mathbf{pr}\ (x_i/\Delta_A) = a_i u_A,$$

a_i médiane binaire de $\{ (x_i^j, \alpha_i^j), j\in A \}$.

Définition
Nous définissons l'inertie du nuage N(I) par rapport à l'axe Δ_A comme la somme des distances (dans B^p) des points du nuage à leur projeté :

$$\mathcal{I}_{\Delta_A}(N(I)) = \sum_{i\in I} D(x_i,a_iu_A) = \sum_{i\in I}\sum_{j\in J} \alpha_i^j\, |x_i^j - a_iu_A^j|$$

Lorsque les pondérations initiales sont toutes égales à un, l'inertie du nuage par rapport à l'axe représente le nombre de composantes différentes entre les points de N(I) et les points correspondants du nuage projeté exprimés dans l'espace B^p.

5.1.2 Propriétés

Soit a le point de B^n défini par :

$$a = (a_1, a_2, \dots, a_n)$$

$$\text{où} \quad \forall i\in I, \quad \mathbf{pr}(x_i/\Delta_A) = a_iu_A$$

Ce point n'est autre que le centre médian du nuage N(A), inclus dans le nuage des variables N(J), et défini par :

$$N(A) = \{ (x^j, \alpha^j), j\in A \}$$

Le vecteur de pondérations associé à a appartient à R^{n+} et :

$$\forall\, i\in I \quad \alpha_i(a) = \alpha(a_i) = \Big| \sum_{j\in A} \alpha_i^j x_i^j - \sum_{j\in A} \alpha_i^j(1-x_i^j) \Big|$$

Il existe donc un lien entre le nuage projeté $N(I/\Delta_A)$ et le nuage des variables N(A) : les valeurs binaires constituant le nuage projeté sont égales aux composantes du centre médian du nuage N(A).

Ci-dessous, nous énonçons une propriété précisant ce lien en terme d'inertie. Cette propriété va également permettre de formuler un problème équivalent au problème de recherche de l'axe d'inertie minimale.

Propriétés
Soit $N(I/\Delta_A)$ le nuage projeté sur l'axe Δ_A, on a :

$$(i) \qquad \Im_0(N(I/\Delta_A)) = \Im_0(\{a\}) = D(a,O)$$

L'inertie du nuage projeté $N(I/\Delta_A)$ par rapport à l'origine et l'inertie du nuage $N(I)$ par rapport à l'axe Δ_A sont liées par la relation :

$$(ii) \qquad \Im_0(N(I)) = \Im_{\Delta_A}(N(I)) + \Im_0(N(I/\Delta_A))$$

Preuve
La relation (i) se déduit de l'expression de l'inertie du nuage projeté :

$$\Im_0(N(I/\Delta_A)) = \sum_{i\in I} \alpha(a_i)a_i = \sum_{i\in I} \alpha_i(a)a_i = \Im_0(\{a\})$$

Si on décompose l'expression de l'inertie du nuage par rapport à l'axe, on a :

$$\Im_{\Delta_A}(N(I)) = \sum_{i\in I}\sum_{j\in J} \alpha_i^j \mid x_i^j - a_i x^j \mid$$

$$\Leftrightarrow \quad \Im_{\Delta_A}(N(I)) = \sum_{i\in I}\sum_{j\in A} \alpha_i^j \mid x_i^j - a_i \mid + \sum_{i\in I}\sum_{j\in J-A} \alpha_i^j x_i^j$$

$$\Leftrightarrow \quad \Im_{\Delta_A}(N(I)) = \sum_{j\in J}\sum_{i\in I} \alpha_i^j \mid x_i^j - a_i \mid + \sum_{i\in I}\sum_{j\in J} \alpha_i^j x_i^j - \sum_{j\in A}\sum_{i\in I} \alpha_i^j x_i^j$$

$$\Leftrightarrow \quad \Im_{\Delta_A}(N(I)) = \Im(N(A)) + \Im_0(N(I)) - \Im_0(N(A))$$

or :

$$\Im_0(N(A)) = \Im(N(A)) + \Im_0(\{a\})$$

d'où :

$$\Im_{\Delta_A}(N(I)) = \Im_0(N(J)) - \Im_0(\{a\})$$

Ce qui démontre la relation (ii) puisque $\Im_0(N(J)) = \Im_0(N(I))$.

5.1.3 Axe d'inertie minimale

Des deux relations (i) et (ii), nous pouvons déduire les équivalences entre les deux problèmes suivants :

- rechercher un axe tel que l'inertie du nuage N(I) par rapport à cet axe soit minimale (ou tel que l'inertie du nuage projeté soit maximale),
- rechercher une partie A de J telle que l'inertie $\Im_0(\{a\})$, où a est le centre médian de N(A), soit maximale.

Resoudre un tel problème revient à donc à rechercher la partie A de J telle que la somme des pondérations des valeurs 1 du centre médian de N(A) soit la plus élevée possible.

Remarques
Il est possible de construire un algorithme itératif analogue à celui du paragraphe 4.5 : à chaque itération, on ajoute une variable à la partie A (initialement vide) si et seulement si l'inertie du nuage par rapport à l'axe associé à cette partie est améliorée. Cependant, cet algorithme ne fournit pas toujours l'axe d'inertie minimale. L'algorithme fournissant la solution optimale reste à construire. Dans la suite, nous nous attachons à la recherche d'un système d'axes d'inertie minimale et, comme nous le verrons, il sera inutile de connaître le meilleur axe puisque la propriété d'inclusion de sous-espaces optimaux n'est pas vérifiée sur $\mathbf{B}^p$.

Si on se place dans l'espace $\mathbf{R}^p$ muni d'une métrique euclidienne, le centre de gravité du nuage appartient à l'axe d'inertie minimale. Le résultat analogue n'est pas vrai sur $\mathbf{B}^p$ muni de la distance en valeurs absolues : l'axe binaire d'inertie minimale ne passe pas toujours par le centre médian du nuage. L'exemple ci-dessous en est une illustration.

Considérons, par exemple, le nuage composé des 5 points suivants :

1	0	0	1
2	0	0	1
3	1	1	0
4	1	1	0
5	1	1	1

Supposons également que les pondérations initiales soient toutes égales à 1. Le centre médian est alors le point (1,1,1). L'inertie du nuage par rapport à l'axe passant par ce point est égale à 4. Pour l'axe passant par le point (1,1,0), on obtient une inertie de 3.

5.1.4 Sous-tableau homogène

Dans le paragraphe 4.2, nous avons mis en évidence la dichotomie suivante :

$$I = I_0 \cup I_A$$

où

$$I_0 = \{ i \in I / pr(x_i/\Delta_A)=O \} = \{ i \in I / a_i=0 \},$$

$$I_A = \{ i \in I / pr(x_i/\Delta_A)=u_A \} = \{ i \in I / a_i=1 \}.$$

On a alors l'égalité suivante :

$$\Im_0(\{a\}) = \sum_{i \in I} \alpha_i(a) a_i = \sum_{i \in I_A} \alpha_i(a)$$

Rechercher l'axe d'inertie minimale revient alors à rechercher une partie A de J telle que le sous-tableau $\mathbf{X}(I_A,A)$ soit le plus homogène possible en valeurs 1. Une mesure de cette homogénéité est fournie par l'inertie ci-dessus, qui peut encore s'exprimer par

On se rapproche ici de la définition intuitive d'une classe polythétique formulée par B. Lefebvre et J. Losfeld (1979) : "une classe polythétique est intuitivement considérée comme un sous-tableau qui possède une proportion et une répartition acceptables de valeurs 1". Pour rechercher ce type de classe, les auteurs ont proposé un algorithme reposant sur la théorie des sous-ensembles flous.

5.1.5 Inertie et mesure d'information

Le nuage projeté sur l'axe associé à une partie A de J est celui défini à partir du tableau **X**(I,A) associé au couple (I,A). Dans le chapitre précédent, nous avons défini une mesure d'information (notée **Info**) d'un tableau binaire. Pour le tableau **X**(I,A), on

$$\mathbf{Info}(\mathbf{X}(I,A)) = \Im_0(\mathbf{N}(I/\Delta_A)) = \Im_0(\{a\}) = \mathbf{D}(a,O)$$

Nous pouvons alors réécrire la relation (ii) de la façon suivante :

$$\mathbf{Info}(\mathbf{X}(I,J)) = \Im_{\Delta_A}(\mathbf{N}(I)) + \mathbf{Info}(\mathbf{X}(I,A))$$

Rechercher l'axe d'inertie minimale revient alors à rechercher la partie A de J telle que la mesure d'information du tableau **X**(I,A) soit maximale.

5.2 INERTIE D'UN NUAGE PAR RAPPORT À UN SYSTEME D'AXES ORTHOGONAUX

5.2.1 Définition

Soient $Q=(Q^1,\ldots,Q^M)$ une partition de J en M classes et **E** l'espace engendré. Pour tout m, on note $\Delta_m=\{O,u_m\}$ l'axe binaire associé à la classe de Q^m. Nous avons vu que le nuage projeté **N**(I/**E**) pouvait s'écrire comme un ensemble de n points de $\mathbf{B}^M$:

$$\mathbf{N}(I/\mathbf{E}) = \{ (a_i, \alpha(a_i)), i \in I \}$$

$$\text{avec} \quad a_i = \mathbf{pr}(x_i/\mathbf{E}) = \sum_{m=1}^{M} a_i^m u_m$$

$$\text{et} \quad \forall m=1 \ldots, M \quad a_i^m = \text{médiane de } \{ (x_i^j, \alpha_i^j), j \in Q^m \}$$

Définition
*Nous définissons l'inertie du nuage N(I) par rapport au sous-espace **E** comme la somme des distances (dans $\mathbf{B}^p$) des points du nuage à leur projeté :*

$$\Im_E(N(I)) = \sum_{i \in I} D(x_i,a_i) = \sum_{i \in I} \sum_{m=1}^{M} \sum_{j \in Q^m} \alpha_i^j \, / \, x_i^j - a_i^m \, /$$

Si toutes les pondérations initiales sont égales à 1, l'inertie du nuage par rapport à **E** représente le nombre de composantes différentes entre les points du nuage initial et les points du nuage projeté exprimés dans $\mathbf{B}^p$.

5.2.2 Propriétés

Pour tout m, on définit le point a^m de $\mathbf{B}^n$ par :

$$a^m = (a_1^m, \ldots, a_n^m) \quad \text{où} \quad \forall i \in I, \quad a_i^m = \mathbf{pr}(x_i/\mathbf{E})$$

Ce point est exactement le centre médian de la classe Q^m. Le vecteur de pondérations associé est le point de $\mathbf{R}^{n+}$ défini par :

$$\forall i \in I \quad \alpha_i(a^m) = \alpha^m(a_i) = | \sum_{j \in Q^m} \alpha_i^j x_i^j - \sum_{j \in Q^m} \alpha_i^j (1-x_i^j) |$$

Soit **X**(I,Q) le tableau binaire à n lignes et M colonnes construit à partir du couple (I,Q). Le nuage projeté correspond au nuage des lignes de ce tableau. Le nuage des centres médians des classes de la partition Q de J est celui associé aux colonnes de ce même tableau. Cela montre le lien entre le nuage projeté et la partition Q. Les propriétés ci-dessous précisent ce lien et permettent une interprétation simple de l'inertie du nuage par rapport au sous-espace.

Propriétés
*Soient **B**(Q) l'inertie interclasse, exprimée par rapport à l'origine, de la partition Q :*

$$B(Q) = \sum_{m=1}^{M} \mathfrak{I}_0(\{a^m\})$$

*et **W**(Q) l'inertie intraclasse :*

$$W(Q) = \sum_{m=1}^{M} \mathfrak{I}(Q^m)$$

On a :

(i) $\mathfrak{I}_0(N(I/\Delta_m)) = \mathfrak{I}_0(\{a^m\})$

(ii) $\mathfrak{I}_0(N(I/E)) = B(Q) = \sum_{m=1}^{M} \mathfrak{I}_0(N(I/\Delta_m))$

(iii) $\mathfrak{I}_E(N(I)) = W(Q)$

(iv) $\mathfrak{I}_0(N(I)) = \mathfrak{I}_E(N(I)) + \mathfrak{I}_0(N(I/E))$

Preuve
La relation (i) est évidente puisque, par définition, on a :

$$\mathfrak{I}_0(N(I/\Delta_m)) = D(a^m, O) = \mathfrak{I}_0(\{a^m\})$$

Lorsqu'elles sont calculées par rapport aux origines des deux espaces respectifs, l'inertie du nuage des lignes du tableau **X**(I,Q) est égale à l'inertie du nuage des colonnes de ce même tableau. Cela démontre la relation (ii).

La relation (iii) se déduit de l'inertie du nuage par rapport au sous-espace **E** :

$$\mathfrak{I}_E(N(I)) = \sum_{i \in I} \sum_{m=1}^{M} \sum_{j \in Q^M} \alpha_i^j | x_i^j - a_i^m |$$

$$\Leftrightarrow \quad \mathfrak{I}_E(N(I)) = \sum_{m=1}^{M} \mathfrak{I}(Q^M)$$

$$\Leftrightarrow \quad \mathfrak{I}_E(N(I)) = W(Q)$$

La relation de décomposition de l'inertie du nuage **N**(J) nous donne :

$$\mathfrak{I}_0(\mathbf{N}(J)) = \mathbf{W}(Q) + \mathbf{B}(Q) \quad \text{et} \quad \mathfrak{I}_0(\mathbf{N}(J)) = \mathfrak{I}_0(\mathbf{N}(I))$$

Des relations (ii) et (iii), on deduit finalement la relation (iv) :

$$\mathfrak{I}_0(\mathbf{N}(I)) = \mathfrak{I}_E(\mathbf{N}(I)) + \mathfrak{I}_0(\mathbf{N}(I/E))$$

5.2.3 Sous-espace d'inertie minimale

Les trois propriétes ci-dessus montrent l'équivalence entre les deux problèmes d'optimisation suivants :

- rechercher un système de M axes tel que l'inertie du nuage par rapport au sous-espace engendré soit minimale (ou telle que l'inertie du nuage projeté soit maximale),
- rechercher une partition Q de J en M classes telle que l'inertie intraclasse de la partition soit minimale (ou telle que l'inertie interclasse soit maximale).

Une solution à ce problème est fournie par la méthode MNDBIN appliquée au nuage des variables. Les classes obtenues permettent de définir les axes engendrant le sous-espace. Les noyaux des classes permettent de définir le nuage projeté sur ce sous-espace. Cependant, dans cette approche, il est nécessaire de fixer, a priori, le nombre d'axes à rechercher.

Dans $\mathbf{R}^p$ muni de la métrique euclidienne usuelle, on dispose du résultat suivant : le sous-espace de dimension M minimisant l'inertie contient le sous-espace de dimension M-1 minimisant également l'inertie. Il est alors alors possible de rechercher la suite des sous-espaces optimaux.

La propriété d'inclusion des sous-espaces optimaux n'est plus vraie sur $\mathbf{B}^p$. Nous avons déjà vu que le meilleur axe ne contient pas forcément le centre médian du nuage. Plus généralement, le meilleur sous-espace engendré par M axes binaires ne contient pas obligatoirement le meilleur sous-espace engendré par M-1 axes binaires. Rappelons également que la notion de sous-espaces supplémentaires n'existe pas ici. La solution que nous proposons est donc de fixer, a priori, le nombre d'axes à rechercher. La méthode de classification MNDBIN fournit alors une solution au problème posé.

5.2.4 Sous-tableaux homogènes

Dans le paragraphe 5.1.4, nous avons montré comment un axe binaire induisait une dichotomie de l'ensemble des individus. Cette dichotomie permettait alors de définir un sous-tableau homogène en valeurs 1. Nous avons ici les mêmes résultats pour chacun des axes engendrant le sous-espace de projection **E**. Ceci est possible car la projection sur **E** s'obtient à partir des projections sur les différents axes.

Rechercher un sous-espace **E** d'inertie minimale revient à rechercher une partition Q de J maximisant :

$$\mathbf{B}(Q) = \mathfrak{I}_0(\mathbf{N}(I/E)) = \sum_{m=1}^{M} \mathfrak{I}_0(\{a^m\})$$

Si on note I_m l'ensemble des individus se projetant sur le point u_m de l'axe $\Delta_m=\{O,u_m\}$ associé à Q^m, on obtient :

$$\Im_0(\mathbf{N}(I/\mathbf{E})) = \sum_{m=1}^{M} \sum_{i\in I} \alpha_i(a^m)a_i^m = \sum_{m=1}^{M} \sum_{i\in I_m} \alpha_i(a^m)$$

En maximisant cette quantité, la méthode fournit aussi un ensemble de M sous-tableaux homogènes en valeurs 1. La mesure de l'homogénéité d'un sous-tableau $\mathbf{X}(I_m,Q^m)$ est fournit par $\Im_0(\{a^m\})=\mathbf{D}(a^m,O)$.

5.2.5 Inertie et mesure d'information

Le nuage projeté sur le sous-espace engendré par Q est celui défini à partir du tableau X(I,Q). De nouveau, nous pouvons poser le problème en terme d'optimisation d'une mesure d'information. En effet, si on reprend la définition de la mesure **Info**, la relation (iv) se réécrit :

$$\mathbf{Info}(\mathbf{X}(I,J)) = \Im_E(\mathbf{N}(I)) + \mathbf{Info}(\mathbf{X}(I,Q))$$

Les deux problèmes précédent sont alors équivalents au problème de la recherche d'une partition Q de J en M classes telle que la mesure d'information du tableau X(I,Q) soit maximale. Cette mesure étant égale à la somme des pondérations des valeurs 1 du tableau X(I,Q), le meilleur sous-espace engendré par une partition Q de J est donc celui représentant le mieux possible les valeurs 1 du tableau initial. Pour un tableau centré, le meilleur sous-espace est celui représentant le mieux possible les valeurs minoritaires.

5.3 INERTIE D'UN NUAGE PAR RAPPORT À UN SYSTEME D'AXES QUELCONQUES

Soit $Q=(Q^1,\ldots,Q^M)$ un recouvrement de M parties de J. On note $(u_1,\ldots,u_M)$ les vecteurs binaires associés aux parties de Q et **E** l'espace engendré.

Comme dans le paragraphe 5.2, nous définissons l'inertie du nuage **N**(I) par rapport au sous-espace **E** par :

$$\Im_E(\mathbf{N}(I)) = \sum_{i\in I} \mathbf{D}(x_i,a_i)$$

$$\text{où} \quad a_i = \mathbf{pr}(x_i/\mathbf{E}) = \sum_{m=1}^{M} a_i^m u_m$$

Nous nous sommes déja heurtés au problème de la recherche du projeté d'un point et du vecteur de pondérations à lui associer. Nous avons mis en évidence le fait que, même s'il est possible de calculer le projeté, il nous est impossible de lui associer un vecteur de pondérations. La conséquence est que nous ne retrouvons pas, ici, des propriétés analogues à celles démontrées lorsque le système d'axes est orthogonal.

Si on cherche à optimiser l'inertie, il est nécessaire de rechercher à la fois les M axes, M étant fixé a priori, et les projetés des points sur l'espace engendré. La résolution d'un tel problème est assez délicate. Avant de présenter l'algorithme que nous avons construit, nous allons étudier la différence entre cette approche (axes quelconques) et l'approche précédente (axes orthogonaux).

5.3.1 Suppression de la contrainte

Si Q est une partition, une variable n'est prise en compte que par un seul axe. Dans ce cas, les variables sont traitées de la même manière. Ainsi, on considère de façon identique une variable à forte proportion de 1 et une variable à faible proportion de 1. Si Q n'est pas une partition, une variable à forte proportion de 1 peut intervenir dans la définition de plusieurs axes. La conséquence est que l'inertie par rapport au sous-espace engendré par un recouvrement en M parties sera meilleure que celle calculée par rapport au sous-espace engendré par une partition en M classes.

D'autre part, si Q est une partition, les classes obtenues regroupent des variables ayant un comportement analogue sur l'ensemble des individus. Pour les parties d'un recouvrement Q, nous n'avons plus le même type d'interprétation. Nous reviendrons sur ce point dans la suite.

Pour illustrer ces remarques, considérons les ensembles I={**1**,**2**,**3**,**4**,**5**,**6**,**7**} et J={**a**,**b**,**c**,**d**,**e**} des variables. Le tableau croisant ces deux ensembles est représenté en figure 1. On suppose, ici, que les composantes des vecteurs de pondérations de ces 7 points sont toutes égales à 1. Dans ce cas, l'inertie représente exactement le nombre de valeurs différentes entre le tableau initial et le tableau associé au nuage projeté.

On cherche alors un système de trois axes. La méthode MNDBIN appliquée aux variables fournie la partition ($\mathbf{A_1}$,$\mathbf{B_1}$,$\mathbf{C_1}$)=({**a**,**b**},{**c**,**e**},{**d**}). On obtient une inertie optimale égale à 3 par rapport au sous-espace engendré par les vecteurs orthogonaux $\mathbf{u_1}$=(1,1,0,0,0), $\mathbf{v_1}$=(0,0,1,0,1) et $\mathbf{w_1}$=(0,0,0,1,0) associés aux classes de la partition. Le tableau associé au nuage projeté (figure 2) diffère de trois valeurs avec le tableau initial.

Soit le sous-espace engendré par les vecteurs $\mathbf{u_2}$=(1,1,0,1,0), $\mathbf{v_2}$=(0,0,0,1,1) et $\mathbf{w_2}$=(0,0,1,0,1) associés aux parties ($\mathbf{A_2}$,$\mathbf{B_2}$,$\mathbf{C_2}$)=({**a**,**b**,**d**),{**d**,**e**},{**c**,**e**}) constituant un recouvrement de l'ensemble des variables. On obtient alors une inertie égale à 1, le tableau associé au nuage projeté (figure 3) diffère d'une seule valeur avec la tableau

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	1
2	1	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 1
tableau initial

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	0
2	1	1	0	1	0
3	0	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	0
6	0	0	1	1	1
7	0	0	1	1	1

figure 2
partition

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	1
2	1	1	0	1	0
3	1	1	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 3
recouvrement

La contrainte implique le regroupement de variables ayant un comportement analogue sur l'ensemble des individus. C'est le cas, par exemple de la classe $\mathbf{A_1}$={**a**,**b**}. Pour la partie $\mathbf{A_2}$={**a**,**b**,**d**) du recouvrement, cela n'est plus vrai : les trois variables n'ont un comportement semblable que sur le sous-ènsemble (**1**,**2**,**3**).

Nous allons maintenant nous intéresser à la recherche d'un recouvrement optimal et étudier, ensuite, le type d'interprétation que l'on peut faire de cette structure.

5.3.2 Le problème d'optimisation

Le problème d'optimisation de l'inertie du nuage par rapport à sous-espace est double :

il s'agit de trouver un ensemble de M axes binaires $(u_1,\ldots,u_M)$ et les projetés $(a_1,\ldots,a_n)$ tels que l'inertie :

$$\Im_E(N(I)) = \sum_{i\in I} D(x_i,a_i) = \sum_{i\in I}\sum_{j\in J} \alpha_i^j \,|x_i^j - (a_i^1 u_1^j + \cdots + a_i^M u_M^j)|$$

soit minimale.

Supposons que **E** soit le sous-espace minimisant ce critère. A partir des projetés $(a_1,\ldots,a_n)$, nous pouvons alors construire les M points $(a^1,\ldots,a^M)$ de $\mathbf{B}^n$ définis par :

$$\forall m=1,\ldots,M \quad a^m = (a_1^m, a_2^m, \ldots, a_n^m)$$

Le système $(a^1,\ldots,a^M)$ engendre alors un sous-espace **F** de $\mathbf{B}^n$.

A partir des M vecteurs $(u_1,\ldots,u_M)$, nous pouvons définir le système de p vecteurs $(u^1,\ldots,u^p)$ définis par :

$$\forall j\in J \quad u^j = (u_1^j, u_2^j, \ldots, u_M^j)$$

On a alors :

$$\Im_E(\mathbf{N}(I)) = \sum_{j\in J}\sum_{i\in I} \alpha_i^j \,|x_i^j - (u_1^j a_i^1 + \cdots + u_M^j a_i^M)|$$

$$\Leftrightarrow \quad \Im_E(\mathbf{N}(I)) = \sum_{j\in J} \mathbf{D}(x^j,u^j) \quad \text{où } u^j = \sum_{m=1}^{M} u_m^j\, a^m$$

$$\Leftrightarrow \quad \Im_E(\mathbf{N}(I)) = \Im_F(\mathbf{N}(J))$$

Finalement, résoudre le problème unique, initialement posé, revient à résoudre les trois problèmes suivants :

- rechercher un sous-espace **E** de $\mathbf{B}^p$ tel que le critère d'inertie $\Im_E(\mathbf{N}(I))$ soit minimum,
- rechercher un sous-espace **F** de $\mathbf{B}^n$ tel que le critère d'inertie $\Im_F(\mathbf{N}(J))$ soit minimum,
- rechercher simultanément le sous-espace **E** de $\mathbf{B}^p$ et le sous-espace **F** de $\mathbf{B}^n$ tels que le critère d'inertie $\Im_E(\mathbf{N}(I))=\Im_F(\mathbf{N}(J))$ soit minimum.

5.3.3 L'algorithme

Nous proposons ici un algorithme, que nous appelons ACPBIN, fournissant une solution au problème d'optimisation. Celui-ci utilise l'algorithme présenté dans le paragraphe 4.5.2 qui fourni le projeté d'un point dans un sous-espace.

ACPBIN est un algorithme itératif qui, à partir d'un recouvrement initial Q de M parties de J, M fixé a priori, optimise tour à tour l'inertie du nuage **N**(I) par rapport à un

sous-espace **E** et l'inertie du nuage **N**(J) par rapport à un sous-espace **F**. A l'itération r, les deux étapes permettant de minimiser le critère sont les suivantes :

(i) Soit **E** le sous-espace de $\mathbf{B}^p$ engendré par le système $(u_1,\ldots,u_M)$ défini à partir des points $(u^1,\ldots,u^p)$ calculés dans l'étape (ii) de l'itération r-1. On calcule, en utilisant l'algorithme du paragraphe 4.5, les projetés $(a_1,\ldots,a_n)$ des points du nuage des individus **N**(I).

(ii) Soit **F** le sous-espace $\mathbf{B}^n$ engendré par le système $(a^1,\ldots,a^M)$ défini à partir des points $(a_1,\ldots,a_n)$ calculés dans l'étape (i) précédente. On calcule, en utilisant l'algorithme du paragraphe 4.5, les projetés $(u^1,\ldots,u^p)$ des points du nuage des variables **N**(J).

A chaque fois, lors les étapes (i) et (ii), on minimise un même critère d'inertie puisque $\Im_E(\mathbf{N}(I))=\Im_F(\mathbf{N}(J))$.

A la convergence, l'agorithme ACPBIN fournit une solution à notre problème, solution qui dépend évidemment du choix du recouvrement initial.

5.3.4 Remarques

Lorsque Q est une partition, l'inertie du nuage par rapport à l'espace engendré n'est autre que l'inertie intraclasse de Q. Elle s'interprète alors comme une mesure de l'écart entre la situation initiale et la situation idéale (correspondant au cas où les variables sont idéntiques aux noyaux des classes auxquelles elles appartiennent). Cette situation idéale correspond également au cas où le nuage des individus est inclus dans le sous-espace de projection : les vecteurs réponses des individus s'écrivent comme des combinaisons linéaires des vecteurs associés aux classes de la partition. Ces combinaisons linéaires se déduisent des projections sur les différents axes et des sous-tableaux homogènes en valeurs 1 peuvent alors être construit.

Lorsque Q est un recouvrement, la notion de situation idéale a un sens plus large contenant la précédente. Elle correspond au cas où les vecteurs réponses des individus s'écrivent comme des combinaisons linéaires des vecteurs associés, cette fois, aux parties du recouvrement. La différence majeure avec le cas précédent est que les projections sur les différents axes ne permettent pas, sauf cas très particulier, de construire le nuage projeté.

Nous pouvons cependant construire les sous-tableaux $\mathbf{X}(I_m,Q^m)$ induits par les dichotomies du type :

$$I = I_0 \cup I_m$$

$$\text{où } I_0 = \{\, i\in I \,/\, a_i^m=0 \,\} \text{ et } I_m) = \{\, i\in I \,/\, a_i^m=1 \,\}$$

Lorsque Q est une partition, I_m est l'ensemble des individus se projetant sur l'axe Δ_m associé à Q^m. Dans le cas d'un recouvrement, cela n'est pas toujours vrai. Les deux cas suivants peuvent se produire :

(a) $i\in I_m$ alors que la médiane de $\{\, (x_i^j, \alpha_i^j), j\in Q^m \,\}$ n'est pas $a_i^m=1$,

(b) $i\in I_0$ alors que la médiane de $\{\, (x_i^j, \alpha_i^j), j\in Q^m \,\}$ n'est pas $a_i^m=0$.

Dans la pratique, le cas (a) n'apparait que très rarement. Le cas (b) est beaucoup plus fréquents mais ne nuit pas à la qualité des sous-tableaux obtenus comme indiqué

ci-dessus. Les variables d'une partie Q^m ont un comportement analogue sur l'ensemble I_m (et non sur tous les individus comme dans le cas des classes d'une partition); les sous-tableaux induits sont pratiquement toujours homogènes en valeurs 1.

Nous n'avons pu ni établir de conditions ni caractériser précisemment les recouvrements menant aux cas (a) et (b). Noton enfin, que ceux-ci disparaissent si on augmente le nombre d'axes ou de parties du recouvrement à rechercher.

Exemple
Soit le tableau croisant 9 individus et 7 variables binaires (figure 1). A partir du recouvrement (**A**,**B**,**C**)=({**a**,**b**,**d**},{**c**,**e**},{**d**,**e**}), on définit le sous-espace de projection **E** engendré par les vecteurs $\mathbf{u_A}$=(1,1,0,1,0), $\mathbf{u_B}$=(0,0,1,0,1) et $\mathbf{u_C}$=(0,0,0,1,1). L'inertie du nuage des individus par rapport à **E** est égale à 2. Le nuage projeté est indiqué en figure 2.

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	1
2	1	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1
8	1	1	1	1	1
9	1	1	0	0	1

figure 1
tableau initial

	$\mathbf{u_A}$	$\mathbf{u_B}$	$\mathbf{u_C}$
1	1	0	1
2	1	0	0
3	1	0	0
4	0	1	0
5	0	0	1
6	0	1	1
7	0	1	1
8	1	0	1
9	0	1	1

figure 2
nuage projeté

On déduit alors les sous-tableaux homogènes en valeurs 1 :

- {**1**, **2**, **3**, **8**} et {**a**, **b**, **d**}
- {**4**, **6**, **7**, **9**} et {**c**, **e**}
- {**1**, **5**, **6**, **7**, **8**, **9**} et {**d**, **e**}

Sur cet exemple, l'indivdu **9** correspond au cas (a) : la médiane des valeurs prises par **9** sur {**d**,**e**} peut être soit 0, soit 1. L'individu **8** correspond au cas (b), il peut également s'écrire $\mathbf{u_A}+\mathbf{u_B}+\mathbf{u_C}$ et être alors ajouté au second bloc ci-dessus.

6. ANALYSE EN COMPOSANTES PRINCIPALES AVEC CONTRAINTE D'AXES ORTHOGONAUX

Habituellement, l'A.C.P fournit un ensemble d'axes factoriels orthogonaux, afin de construire des représentations visuelles des individus sur les différents plans factoriels. Nous proposons ici une méthode d'analyse en composantes principales, spécifique aux données binaires, recherchant un ensemble d'axes factoriels binaires et respectant la forme initiale des données.

6.1 ANALYSE EN COMPOSANTES PRINCIPALES ET CLASSIFICATION

Dans le paragraphe 5, nous avons démontré qu'une telle méthode était équivalente à la méthode de classification MNDBIN appliquée au nuage des variables. Nous avons

démontré l'égalité entre l'inertie du nuage des individus **N**(I) par rapport à l'espace **E** engendré par une partition Q et l'inertie intraclasse de la partition Q.

La méthode MNDBIN fournit donc une solution au problème de l'ACP sur données binaires. A partir de la partition Q en M classes de J fournie par la méthode, nous pouvons construire les différents axes associés et en déduire le sous-espace **E** engendré. Les noyaux des classes permettent, eux, de définir le nuage projeté, c'est à dire les composantes des individus dans la base de **E**. Celui-ci peut d'ailleurs s'interpréter comme un espace de type $\mathbf{B}^M$. Ainsi, nous avons représenter le nuage initial dans un espace $\mathbf{B}^M$ de dimension inférieure à $\mathbf{B}^p$.

Réciproquement, effectuer une ACP binaire sur le nuage des variables revient à effectuer une classification des individus. La méthode MNDBIN appliquée à **N**(I) fournit une partition P engendrant un sous-espace **F** optimal.

On ne retrouve pas ici le shéma de dualité habituel entre l'analyse des individus et l'analyse des variables. En effet, les composantes des n individus exprimés dans le sous-espace de projection ne permettent pas, sauf cas très particulier, de définir un système de M axes orthogonaux de $\mathbf{B}^n$.

Enfin, la méthode CROBIN fournit un couple (P,Q) de partitions en K et M classes de (I,J) maximisant la mesure d'information du tableau **X**(P,Q) associé. Cette méthode apparait alors comme une méthode recherchant simultanément les deux sous-espaces **E** (engendré par Q) et **F** (engendré par P) optimaux.

6.2 REMARQUES SUR LA MÉTHODE

L'application de MNDBIN sur J met en évidence une partition Q et un ensemble de M axes minimisant un critère d'inertie. On réalise ainsi une ACP sous contrainte. Plusieurs remarques à ce sujet peuvent être faites.

C'est à l'utilisateur que revient le choix du nombre M d'axes ou de classes à rechercher par la méthode MNDBIN. Les représentations des individus dans le sous-espace de projection **E** engendré par les M axes sont alors fournies par les noyaux. Puisque **E** est fini, nous pouvons associer, à chaque élément de **E**, l'ensemble des points du nuage des individus qui se projettent sur cet élément. On dispose ainsi d'un maximum de 2^M classes d'individus qui forment une partition de I (pour M=2 axes, on a une partition de I en 4 classes au plus). Si M est grand, il peut être intéressant d'appliquer ensuite la méthode MNDBIN au nuage projeté. Le résulat est alors une partition de I où les classes regroupent des individus ayant des projections voisines.

La projection du nuage des individus sur l'espace **E** induit un recouvrement de I où les parties sont composées d'individus se projettant sur les points non nuls des différents axes (on peut rajouter la partie contenant les points se projetant sur l'origine si le cas se présente). La méthode apparait alors comme recherchant un recouvrement de I minimisant un critère d'inertie. En croisant ce recouvrement et la partition de J, on obtient des sous-tableaux homogènes en valeurs 1. En effet, chaque axe induit une dichotomie de l'ensemble des individus. L'application fait alors ressortir un ensemble de M sous-tableaux homogènes en valeurs 1 du type $\mathbf{X}(I_m,Q^m)$ où I_m est l'ensemble des individus se projetant sur le point associé à Q^m.

Un autre point de vue peut être abordé ici, celui de la réduction du nombre de variables. La méthode fournit un ensemble de M nouvelles variables binaires résumant le mieux possible les p variables initiales. On peut alors appliquer la méthode MNDBIN aux lignes du tableau du nuage projeté. D'autre part, puisque les variables initiales sont binaires, chacune d'elle induit une dichotomie de I. Il est alors possible d'appliquer à ces p dichotomies un algorithme de recherche des formes fortes. Si p est grand, cela devient

coûteux. On peut alors appliquer cet algorithme à partir des M dichotomies, en plus petit nombre, induites par les nouvelles variables.

6.3 INTERPRÉTATION DES RESULTATS

La valeur du critère à la convergence fournie par la méthode MNDBIN est l'inertie du nuage des individus par rapport à **E**. Il représente simplement le nombre de différences entre le tableau initial et le tableau du nuage projeté (si toutes les pondérations initiales sont égales à 1).

Les centres médians des classes et leurs pondérations permettent de tirer les premières conclusions sur la qualité de représentation des individus. Par exemple, si la pondération de la composante m du projeté d'un individu i est élevée (proche du cardinal de la classe Q^m), cela signifie que l'individu i est bien représenté sur l'axe Δ_m.

Nous pouvons également reprendre les indices d'aide à l'interprétation d'une partition (chapitre 5) et leurs donner une interprétation en matière d'ACP pour données binaires. Cette fois, il s'agit de décrire une partition Q de J en M classes. Pour obtenir des indices qui soient ici interprétables, il suffit de les redéfinir à partir de la relation de décomposition de l'inertie du nuage des variables par rapport à l'origine. La relation de décomposition est la suivante :

$$\mathfrak{I}_0(\mathbf{N}(I)) = \mathfrak{I}_E(\mathbf{N}(I)) + \mathfrak{I}_0(\mathbf{N}(I/E))$$

$$\Leftrightarrow \quad \mathfrak{I}_0(\mathbf{N}(J)) = \mathbf{W}(Q) + \mathbf{B}(Q)$$

Nous reprenons ci-dessous les indices les plus représentatifs, en donnant leur nouvelle expression et leur interprétation.

L'indice général
L'indice **R** représente la part d'inertie ou encore la part de la mesure d'information conservée :

$$\mathbf{R} = \frac{\mathfrak{I}_0(\mathbf{N}(I/E))}{\mathfrak{I}_0(\mathbf{N}(I))} = \frac{\mathbf{Info}(\mathbf{X}(I,Q))}{\mathbf{Info}(\mathbf{X}(I,J))}$$

Un indice **R** grand (proche de 1) signifie que le nuage projeté représente correctememt le nuage des individus.

Qualité de représentation des individus
Soient x_i le point représentatif d'un individu i et a_i l'image de x_i par la projection. La qualité de représentation de l'individu i dans **E** est fournie par l'indice **COR**(i) défini par :

$$\mathbf{COR}(i) = \frac{\mathfrak{I}_0(\{a_i\})}{\mathfrak{I}_0(\{x_i\})} = \frac{\mathbf{D}(a_i,O)}{\mathbf{D}(x_i,O)}$$

L'indice varie de 0 (l'individu est mal représenté) à 1 (x_i appartient au sous-espace de projection).

Qualité de représentation des individus sur les axes
L'indice permettant de mesurer la qualité de représentation d'un individu i sur un axe Δ_m est l'indice **COR**(i,m) défini par :

$$\mathbf{COR}(i,m) = \frac{\alpha^m(a_i)\, a_i^m}{\Im_0(\{a_i\})}$$

où

a_i^m est la coordonnées du projeté de x_i sur l'axe Δ_m,

$\alpha^m(a_i)$ est la pondération associée.

Ces indices permettent de distinguer les axes sur lesquels les individus sont le mieux représentés.

Description des axes
Nous reprenons ici les indices de description des classes d'une partition. Ils sont au nombre de 3 et leur interprétation, pour tout axe Δ_m, est la suivante :

T(m) représente la part de valeurs 1 extraite par l'axe Δ_m. Il s'agit également de la part d'information extraite par l'axe Δ_m. Un axe est d'autant plus important qu'il lui correspond une part de valeurs 1 élevée.

B(m) représente la part d'inertie du nuage projeté prise en compte par l'axe Δ_m. Il s'agit de la part d'information prise en compte par cet axe.

W(m) représente la contribution relative de l'axe Δ_m à l'inertie du nuage des individus par rapport au sous-espace.

Dans le paragraphe suivant, nous proposons une application sur un tableau élémentaire. Les indices définis dans ce paragraphe sont calculés et interprétés.

6.4 EXEMPLE D'APPLICATION

Considérons le tableau croisant l'ensemble I={**1**,**2**,**3**,**4**,**5**,**6**,**7**} des individus et l'ensemble J={**a**,**b**,**c**,**d**,**e**} des variables (représenté dans la figure 1). On suppose également que les pondérations initiales sont toutes égales à 1.

On applique la méthode MNDBIN sur l'ensemble J en demandant 3 classes. A la convergence, la méthode fournit la partition Q=(**A**,**B**,**C**)=({**a**,**b**},{**c**, **e**},{**d**}). De celle-ci, nous déduisons les trois vecteurs :

$$\mathbf{u_A}= (1, 1, 0, 0, 0) \quad \mathbf{u_B}= (0, 0, 1, 0, 1) \quad \mathbf{u_C}= (0, 0, 0, 1, 0)$$

engendrant le sous-espace de projection **E**.

Le critère égal à 3 représente le nombre de valeurs différentes entre le tableau initial et le tableau du nuage projeté (indiqué en figure 2) exprimé dans l'espace $\mathbf{B}^5$. Les noyaux nous donnent les coordonnées des individus dans le sous-espace de projection **E** (tableau de la figure 3). De ce tableau se dégage la partition suivante des individus :

$$(\{\mathbf{1}, \mathbf{2}\}, \{\mathbf{3}, \mathbf{5}\}, \{\mathbf{4}\}, \{\mathbf{6}, \mathbf{7}\})$$

Notons que l'indice **R**=0.842 signifie que le sous-espace prend en compte 84.2% de l'information initiale.

Nous précisons enfin la qualité de représentation des individus dans l'espace **E** (figure 4) et sur chacun des axes Δ_A, Δ_B et Δ_C (figure 5). Par exemple, on peut constater que les individus **2,4,6** et **7** appartiennent au sous espace de projection **E**. L'individu **1** est correctement représenté, tandis que les individus **3** et **5** ne le sont que moyennement.

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	1
2	1	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 1
tableau initial

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	0
2	1	1	0	1	0
3	0	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	0
6	0	0	1	1	1
7	0	0	1	1	1

figure 2
nuage projeté
exprimer dans $\mathbf{B}^5$

	$\mathbf{u_A}$	$\mathbf{u_B}$	$\mathbf{u_C}$
1	1	0	1
2	1	0	1
3	0	0	1
4	0	1	0
5	0	0	1
6	0	1	1
7	0	1	1

figure 3
nuage projeté

	E
1	$\frac{3}{4}$
2	1
3	$\frac{1}{2}$
4	1
5	$\frac{1}{2}$
6	1
7	1

figure 4
indices **COR**(i)

	Δ_A	Δ_B	Δ_C
1	$\frac{2}{3}$	0	$\frac{1}{3}$
2	$\frac{2}{3}$	0	$\frac{1}{3}$
3	0	0	1
4	0	1	0
5	0	0	1
6	0	$\frac{2}{3}$	$\frac{1}{3}$
7	0	$\frac{2}{3}$	$\frac{1}{3}$

figure 5
indices **COR**(i,Q)

En outre, le tableau du nuage projeté permet de construire trois sous-tableaux homogènes en valeurs 1. Il s'agit de ceux croisant :

- **{1, 2}** et **{a, b}**
- **{4, 6, 7}** et **{c, e}**
- **{1, 2, 3, 5, 6, 7}** et **{d}**

6.5 INFLUENCE DE LA TRANSFORMATION DU TABLEAU INITIAL

La transformation étudiée ici est celle due à un changemant d'origine de l'espace binaire $\mathbf{B}^p$. Considérons le tableau initial **X**(I,J) et un point b de $\mathbf{B}^p$. Prendre b pour origine de l'espace revient à transformer le tableau **X**(I,J) en un tableau **X**(I(b),J(b)) défini par :

$$\forall i \in I, \forall j \in J \quad x_i^j(b) = | x_i^j - b^j |$$

Comme nous l'avons vu dans le paragraphe 3.5 du chapitre 5, cette transformation n'a aucune influence sur la méthode MNDBIN appliquée au nuage des lignes. En effet, le changement d'origine n'influe pas sur les proximités entres individus.

Par contre, les résultats change si on applique la méthode sur le nuage des colonnes puisque certaines d'entre-elles se trouvent inversées. Cela revient à appliquer la méthode à l'ensemble de p variables J(b) défini par :

$$J(b) = \{j / b^j=0\} \cup \{j' / b^j=1\}$$

où j' est la variable inverse de j : si j prend la valeur 1 (resp. 0) j' prend la valeur 0 (resp. 1). La transformation influe donc sur les résultats de l'ACP binaire. Nous donnons ci-dessous des exemples illustratifs.

Exemple1
Reprenons le tableau du paragraphe 6.3 et représenté en figure 1. Par rapport à l'espace engendré par la partition ({**a**,**b**},{**c**,**e**},{**d**}), nous avions obtenu une inertie égale à 3. Si nous prenons comme origine le centre médian de l'espace, l'ensemble à classifier devient {**a**,**b**,**c**,**d**,**e**} où les variables soulignées représentent les variables inversées. On obtient alors le tableau indiqué en figure 2. Si on applique la méthode MNDBIN en demandant trois classes, elle fournit un critère égal à 2 pour la partition ({**a**,**b**,**e**},{**c**},{**d**}). Le tableau du nuage projeté est indiqué en figure 3.

	a	**b**	**c**	**d**	**e**
1	1	1	0	1	1
2	1	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 1
tableau initial

	a	**b**	**c**	**d**	**e**
1	1	1	0	0	0
2	1	1	0	0	1
3	1	0	0	0	1
4	0	0	1	1	0
5	0	0	0	0	0
6	0	0	1	0	0
7	0	0	1	0	0

figure 2
tableau initial
transformé

	a	**b**	**c**	**d**	**e**
1	1	1	0	0	1
2	1	1	0	0	1
3	1	1	0	0	1
4	0	0	1	1	0
5	0	0	0	0	0
6	0	0	1	0	0
7	0	0	1	0	0

figure 3
tableau du nuage
projeté

Dans cet exemple, on obtient un meilleur résultat en choisissant le centre médian pour origine. Mais le meilleur résultat possible n'est pas toujours celui obtenu pour cette transformation. Dans l'exemple qui suit, nous montrons qu'il existe une transformation, autre que celle par rapport au centre médian, et aboutissant au meilleur résultat.

Exemple2
Considérons le tableau croisant I={**1**,**2**,**3**,**4**,**5**,**6**,**7**} et J={**a**,**b**,**c**} représenté en figure 4 (page suivante). Si on applique la méthode en demandant 2 classes on obtient un critère égal à 3 pour la partition ({**a**,**c**},{**b**}).

Le centre médian du nuage des individus est le point (0,1,1). Le tableau transformé croisant I et {**a**,**b**,**c**} est indiqué en figure 5. Cette fois la méthode fournit la partition ({**a**,**b**},{**c**}) et un critère égal à 2.

Enfin, considérons le point (0,1,0) et le tableau transformé croisant I et {**a**,**b**,**c**} indiqué en figure 6. La méthode fournit alors la partition ({**a**},{**b**,**c**}) et un critère égal à 1.

	a	b	c
1	1	0	1
2	1	0	1
3	0	0	1
4	0	1	1
5	0	1	0
6	1	1	0
7	0	1	0

figure 4

	a	b	c
1	1	1	0
2	1	1	0
3	0	1	0
4	0	0	0
5	0	0	1
6	1	0	1
7	0	0	1

figure 5

	a	b	c
1	1	1	1
2	1	1	1
3	0	1	1
4	0	0	1
5	0	0	0
6	1	0	0
7	0	0	0

figure 6

Le problème
Les résultats ci-dessus nous amènent à énoncer le problème suivant :

> *trouver un point b de* B^p *et un ensemble de M axes orthogonaux engendrant un sous-espace* **E** *tels que le critère de l'inertie du nuage des individus N(I(b)) par rapport à* **E** *soit minimale.*

Le changement d'origine n'influe pas sur les proximités entre individus mais sur celles des variables. Ici, nous proposons de rechercher le point de B^p qui, lorsqu'il est choisi comme origine, conduit à un critère minimum. Nous n'avons pas encore résolu ce problème.

Si nous dédoublons le tableau initial en ajoutant les p variables inversées, nous aboutissons encore à un problème équivalent au précedent.

6.6 LES LIMITES DE CETTE APPROCHE

Le fait de rechercher des axes orthogonaux nuit à la qualité des résultats fournis. En effet, une variable n'intervient que dans la définition d'un seul axe. Ainsi, une variable ayant une forte proportion de 1 est traitée de la même manière qu'une variable ayant une faible proportion de 1. Une première solution consiste alors à centrer le tableau, c'est-à-dire à choisir le centre médian comme origine. Ainsi, toutes les colonnes du tableau transformé auront une médiane égale à 0. On réduit les disparités entres variables et, ainsi, les résultats sont améliorés. Mais cela ne résoud pas le problème de la transformation optimale.

Une autre solution consiste à supprimer la contrainte d'orthogonalité (voir exemple du paragraphe 5.3.1). Ainsi, une variable peut intervenir dans la définition de plusieurs axes binaires. Ce sera notamment le cas des variables à forte proportion de valeurs 1. Les résultats n'en peuvent être qu'améliorés. En procédant ainsi, on se rapproche d'avantage de l'ACP habituelle, où les axes recherchés sont des combinaisons linéaires des caractères initiaux (un caractère initial contribue, de façon plus ou moins importante, à la définition de plusieurs axes factoriels). Dans le paragraphe qui suit, nous nous intéressons à la recherche d'un système d'axes quelconques engendrant un sous-espace

7. ANALYSE EN COMPOSANTES PRINCIPALES SANS CONTRAINTE D'AXES ORTHOGONAUX

La suppression de la contrainte donne un nouveau sens à l'analyse. On recherche maintenant un recouvrement en M parties de J engendrant un sous-espace **E**. Chacune de ces parties contient un ensemble de variables au comportement analogue sur un sous-ensemble seulement d'individus.

Après avois rappeler le principe de la méthode ACPBIN, nous nous intéressons au problème du choix d'un point initial. Nous montrons ensuite le lien entre cette méthode et la méthode d'Analyse Factorielle Booléenne AFB présentée dans le cadre du logiciel BMDP (M.R. Mickey, P. Mundel et L. Engelman 1983). Cette méthode pose le problème sous une forme matricielle. Elle présente la particularité de rechercher un système de M axes, M fixé a priori, à partir d'un système de m axes, m fixé a priori et m inférieur à M. Pour cela, elle utilise un algorithme intermédiaire qui correspond à

7.1 LA MÉTHODE

Dans le paragraphe 5.3, nous avons décrit l'algorithme ACPBIN fournissant un sous-espace **E** engendré par M vecteurs binaires et minimisant l'inertie du nuage des individus par rapport à cet espace. Nous avons également démontré que cette méthode fournit par la même occasion un sous-espace **F** minimisant l'inertie du nuage des variables. En fait, la méthode fournit simultanément les deux sous-espaces **E** et **F** optimaux.

Les résultats fournis par ACPBIN dépendent, évidemment, du choix d'un point initial. Une solution simple consiste à choisir comme point initial une partition en M classes de J. Il suffit alors de faire plusieurs essais, à partir de différentes partitions initiales de J, et de retenir le meilleur résultat.

Une solution plus juste (se rapprochant d'avantage de la notion de recouvrement de J) consiste à appliquer, tout d'abord, la méthode MNDBIN au nuage des individus en demandant M classes. Nous utilisons les M noyaux obtenus pour définir un système de M axes engendrant un sous-espace **E**. Nous proposons de choisir **E** comme point initial de la méthode ACPBIN. La valeur initiale du critère optimisé par ACPBIN est alors égale à celle obtenue à la convergence par la méthode MNDBIN (un individu se projette sur le centre médian de la classe à laquelle il appartient). Ce choix induit une convergence plus rapide de l'algorithme ACPBIN. De façon symétrique, nous pouvons choisir comme point initial un sous-espace **F** engendré par une partition ou un

Les deux choix initiaux finalement retenus sont :

- une partition de I fournie par MNDBIN,
- une partition quelconque de I.

Nous proposons, dans la suite, un exemple simple d'application, sur lequel nous appliquons les différentes méthodes d'analyse en composantes principales étudiées ici.

Dans le dernier paragraphe de ce chapitre, nous appliquons toutes ces méthodes sur des tableaux plus conséquents. Il sera alors possible de comparer les résultats obtenus à partir des différents points initiaux indiqués ci-dessus.

7.2 REMARQUE SUR LA MÉTHODE

La méthode ACPBIN fournit simultanément les deux sous-espaces de projection. Elle apparait également comme une méthode permettant de rechercher simultanément un recouvrement de I et un recouvrement de J par minimisation du critère d'inertie.

Les axes engendrant **E** induisent un recouvrement de J : à chaque axe de vecteur u_m est associé la partie Q^m de J. On a un résultat symétrique pour les axes engendrant l'espace **F**. En croisant ceraines parties des recouvrements de I et de J, on peut définir des sous-tableaux. Chaque axe u_m de l'espace **E** est associé à une partie Q^m de J et induit la dichotomie suivante de I :

$$I = I_0 \cup I_m$$

où I_0 est l'ensemble des individus se projetant sur l'origine,

et I_m est l'ensemble des individus se projetant sur le point u_m.

Dans le paragraphe 5.3.4, nous avons souligné les limites de l'interprétation des sous-tableaux du type $X(I_m,Q^m)$. Cependant, lors des nombreux essais effectués, ceux-ci sont pratiquement toujours homogènes en valeurs 1. On obtient évidemment les mêmes sous-tableaux en considérant les axes engendrant **F**.

7.3 L'ANALYSE FACTORIELLE BOOLÉENNE

Nous montrons ici que l'algorithme AFB utilise un algorithme d'amélioration qui n'est autre que l'algorithme ACPBIN. Tous deux nécessitent le choix d'un point initial. Pour celui-ci, différentes possiblités sont également envisagées.

7.3.1 Principe de la méthode

Soit **X**(I,J) d'ordre (n,p) le tableau initial muni de pondérations toutes égales à 1 (la méthode AFB s'applique à des données sans pondérations particulières). Le but de la méthode est de représenter les p variables initiales $\mathbf{X}=(x^1,\ldots,x^p)$ par un ensemble $\mathbf{S}=(s^1,\ldots,s^M)$ de M "factors" ou facteurs (M << p) .

Les facteurs, éléments de $\mathbf{B}^n$, sont des combinaisons linéaires des variables initiales à coefficients dans {0,1}. Les opérateurs utilisées pour évaluer ces combinaisons sont toujours les oprérateurs logiques "et" et "ou".

Cette approche est associée au modèle matriciel :

$$\mathbf{X} = \mathbf{S} \times \mathbf{L}$$

où **L** d'ordre (M,p) est la matrice des "factors loading" ou matrice des coefficients des combinaisons linaires des facteurs, **S** d'ordre (n,M) est la matrice des "factors scores".

Les opérations matricielles reposent également sur les opérateurs logiques. Le problème consiste alors à déterminer les matrices **S** et **L** de sorte que le nombre de différences entre le tableau initial et le tableau résultant du produit **SxL** soit minimum. Si on note **Y=SxL**, le critère **W** à minimiser est alors le suivant :

$$\mathbf{W} = \sum_{i\in I} \sum_{j\in J} | x_i^j - y_i^j | \quad \text{où} \quad y_i^j = \sum_{m=1}^{M} s_i^m l_m^j$$

Si on note :

$$\mathbf{R} = (r_i^j)$$

la matrice (n,p) d'ajustement entre les matrices **X** et **Y=SxL**, on a :

$$\mathbf{X} = \mathbf{Y} + \mathbf{R}$$

Les éléments de **R** sont alors définis par :

$$r_i^j = 0 \quad \text{si} \quad x_i^j = y_i^j$$

$$r_i^j = 1 \quad \text{si} \quad x_i^j = 1 \text{ et } y_i^j = 0$$

$$r_i^j = -1 \quad \text{si} \quad x_i^j = 0 \text{ et } y_i^j = 1$$

Le critère à minimiser s'écrit alors :

$$\mathbf{W} = \sum_{i \in I} \sum_{j \in J} | r_i^j |$$

$$\Leftrightarrow \quad \mathbf{W} = \sum_{i \in I} \sum_{j \in J} x_i^j(1 - y_i^j) + (1 - x_i^j)y_i^j$$

$$\Leftrightarrow \quad \mathbf{W} = \sum_{i \in I} \sum_{j \in J} x_i^j - \sum_{i \in I} \sum_{j \in J} x_i^j y_i^j - (1 - x_i^j)y_i^j$$

Il est donc équivalent de maximiser le nouveau critère **W** défini cette fois par :

$$\mathbf{W} = \sum_{i \in I} \sum_{j \in J} x_i^j y_i^j - (1 - x_i^j)y_i^j$$

Ce critère représente alors la différence entre le nombre de valeurs 1 communes aux tableaux **X** et **Y** et le nombre de valeurs 0 du tableau **X** représentées par des valeurs 1 dans le tableau **Y**.

7.3.2 La méthode

L'utilisateur choisit le nombre m de facteurs initiaux et le nombre M de facteurs à recherchés. La méthode se déroule ensuite en deux étapes consistant à ajouter et supprimer des facteurs.

Tout d'abord, l'ensemble des m facteurs initiaux est déterminé soit par lecture, soit par calcul. Puis, les M-m autres facteurs sont calculés de la façon suivante : l'algorithme AFB ajoute un $(m+1)^{ème}$ facteur et un $(m+2)^{ème}$ facteurs, puis élimine le $m^{ème}$ facteur. A chaque ajout ou élimination d'un facteur, l'algorithme améliore le critère de la façon suivante : pour une matrice **S** fixée il recherche une matrice **L** (et, réciproquement, pour une matrice **L** fixée il recherche une matrice **S**) de sorte que, à chaque fois, le critère **W** soit optimisé. Ces opérations sont répétées jusqu'à l'obtention du $M^{ème}$ facteur pour la deuxième fois.

Par exemple, si m=2 et M=5, l'algorithme va rechercher 2, 3, 4, 3, 4, 5, 4 et 5 facteurs.

L'amélioration du critère est réalisée par un algorithme appelé "boolean regression" ou régression booléenne. Celui-ci permet de rechercher, à **S** fixé (resp. à **L** fixé), une colonne de la matrice **L** (resp. une ligne de la matrice **S**) suivant le modèle :

$$x^j = y^j + r^j \quad \text{avec} \quad y^j = S x L^j \quad (\text{resp. } x_i = s_i x \mathbf{L} + r_i)$$

Le critère optimisé s'écrit :

$$\mathbf{W}^j = \sum_{i \in I} |r_i^j| = \sum_{i \in I} x_i^j y_i^j - (1 - x_i^j) y_i^j \quad (\text{resp. } \mathbf{W}_i = \sum_{j \in J} x_i^j y_i^j - (1 - x_i^j) y_i^j)$$

7.4 LIEN ENTRE LES DEUX MÉTHODES

L'algorithme ACPBIN que nous avons proposé est identique à l'algorithme d'amélioration utilisé par l'AFB. La matrice **S** apparait comme la matrice des vecteurs engendrant un sous-espace **F** de $\mathbf{B}^n$ et la matrice **L** comme la matrice des coordonnées des variables initiales dans ce sous-espace. De façon symétrique, la matrice **L** apparait comme la matrice des vecteurs engendrant un sous-espace **E** de $\mathbf{B}^p$ et **S** comme la matrice des coordonnées des individus dans ce sous-espace. Dans les deux cas, le critère optimisé représente le nombre de différences entre le tableau initial et le tableau **S**x**L**, ce dernier n'étant autre que le tableau du nuage projeté. Le critère optimisé est donc l'inertie du nuage des individus (resp. des variables) par rapport au sous-espace **E** (resp. **F**).

Comme nous l'avons vu dans ce chapitre, nous ne pouvons pas construire un système optimal en m axes à partir du système optimal en m-1 axes. L'AFB s'inspire pourtant de ce principe, puisqu'il rajoute deux axes avant d'en supprimer un en améliorant à chaque fois le critère. Il utilise ainsi un système initial de m axes (m fixé a priori) pour aboutir à un système de M axes.

Le problème du point initial de la méthode ACPBIN se pose donc également pour la méthode AFB. Actuellement, le point initial de la méthode AFB peut soit être fourni par l'utilisateur, soit être déterminé par un algorithme recherchant des relations d'inclusions entre les colonnes du tableau initial (nous n'avons pas d'informations plus précises sur ce sujet). Par exemple, pour le tableau initial indiqué en figure 1, cet algorithme conduit au point initial indiqué en figure 2.

1	0	1
1	0	1
0	1	1
0	1	1

figure 1

1	0	1
0	1	1

figure 2

Une solution est de démarrer l'AFB en partant d'un seul facteur, c'est-à-dire d'un seul axe binaire, par exemple celui minimisant l'inertie. Nous avons proposé un algorithme approximant un tel axe. Nous pouvons donc l'utiliser ici pour obtenir une matrice **L** initiale et un critère initial de bonne qualité.

Puisque l'algorithme améliore le choix initial, une autre solution consiste à démarrer la méthode à partir d'un axe quelconque. Si on choisit celui passant par le centre médian, le point initial est alors optimal, mais il est possible d'obtenir un résultat final de meilleure qualité en partant d'un point initial quelconque (c'est une conséquence de la propriété d'inclusion des sous-espaces optimaux qui n'est pas vraie sur $\mathbf{B}^p$). Nous avons des remarques analogues pour le choix d'une matrice **S** initiale.

Une autre solution consiste à choisir une matrice **L** initiale ne contenant qu'une seule ligne dont tous les éléments sont choisis égaux à 0. De cette façon, on ne suppose aucune inclusion entre les colonnes de la matrice initiale. Cette ligne **L** de 0 induit une colonne **S** de 0 et, après ajout de deux facteurs, toutes ces valeurs nulles seront éliminées par l'étape de suppression. En fait, en procédant ainsi, le résultat ne va dépendre que de la manière dont on ajoute et supprime des facteurs et de l'algorithme d'amélioration. Ce choix permet alors de tester l'algorithme AFB.

Enfin, nous pouvons également procéder comme nous l'avons fait pour la méthode ACPBIN : on détermine le point initial soit à partir d'une partition quelconque de I, soit à partir de la meilleure partition de I obtenue par MNDBIN (de façon symétrique, nous pouvons faire le même type de choix pour une partition de J).

En résumé, les différents choix initiaux envisagés pour de futures applications sont les suivant :

- une partition initiale de I fournie par MNDBIN,
- une partition initiale quelconque de I,
- l'axe passant par le centre médian,
- aucun point initial, c'est-à-dire une matrice ligne **L** identiquement nulle,
- un axe quelconque.

7.5 UN EXEMPLE SIMPLE D'APPLICATION

Considérons le tableau croisant l'ensemble I={**1**,**2**,**3**,**4**,**5**,**6**,**7**} des individus et l'ensemble J={**a**,**b**,**c**,**d**,**e**} des variables (figure 1 de la page suivante). Nous appliquons tout d'abord la méthode ACPBIN puis la méthode AFB en demandant 3 axes.

Application de ACPBIN
Le premier point initial envisagé est la meilleure partition de I fournie par MNDBIN. On obtient une valeur égale à 3 pour l'inertie du nuage par rapport au système fourni par ACPBIN. Cependant, cela ne constitue pas la meilleure solution, comme le montre l'application suivante.

Nous appliquons ACPBIN en partant d'une partition initiale quelconque. Après plusieurs essais, on obtient comme meilleur résultat une inertie égale à 1 pour le sous-espace E engendré par le système ($\mathbf{u}_A$,$\mathbf{u}_B$,$\mathbf{u}_C$) indiqué en figure 4. Ces vecteurs sont ceux associés au recouvrement :

$$(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (\{\mathbf{a}, \mathbf{b}, \mathbf{d}\}, \{\mathbf{c}, \mathbf{e}\}, \{\mathbf{d}, \mathbf{e}\})$$

fourni à la convergence de l'algorithme. Paralléllement, ACPBIN fourni également le recouvrement de l'ensemble des individus :

$$(\mathbf{I}_A, \mathbf{I}_B, \mathbf{I}_C) = (\{\mathbf{1}, \mathbf{2}, \mathbf{3}\}, \{\mathbf{4}, \mathbf{6}, \mathbf{7}\}, \{\mathbf{1}, \mathbf{5}, \mathbf{6}, \mathbf{7}\})$$

que l'on obtient simplement à partir du tableau du nuage projeté exprimé dans **E** (figure 3). Le nuage projeté exprimé dans $\mathbf{B}^5$ est représenté en figure 2.

A partir de ces deux recouvrements, nous déduisons les tableaux homogènes en valeurs 1 correspondant aux couples ($\mathbf{A}$,$\mathbf{I}_A$), ($\mathbf{B}$,$\mathbf{I}_B$), ($\mathbf{C}$,$\mathbf{I}_C$). Ceux-ci sont respectivement représentés dans les figure 5, 6 et 7.

Application de AFB
Parmi les choix initiaux, trois seulement ont permis d'aboutir à une inertie de 1 :

(i) une partition quelconque en deux classes de I,

(ii) une matrice ligne **L** identiquement nulle,

(iii) un axe initial quelconque.

On retrouve alors des résultats identiques à ceux fournis par ACPBIN. Encore une fois, le choix d'un point initial quelconque (une partition de I en deux classes ou un axe quelconque) permet d'aboutir, après plusieurs essais, au meilleur résultat. Le choix (ii) montre que les algorithmes d'ajout, de suppression et d'amélioration utilisés par l'AFB fournisssent des bons résultats sur ce tableau élémentaire. Sur des tableaux plus importants, cela n'est pas toujours vrai comme nous le verrons dans le dernier paragraphe.

	a	b	c	d	e
1	1	1	0	1	1
2	1	1	0	1	0
3	1	0	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 1
tableau initial

	a	b	c	d	e
1	1	1	0	1	1
2	1	1	0	1	0
3	1	1	0	1	0
4	0	0	1	0	1
5	0	0	0	1	1
6	0	0	1	1	1
7	0	0	1	1	1

figure 2
nuage projeté
exprimé dans $\mathbf{B}^5$

	u_A	u_B	u_C
1	1	0	1
2	1	0	0
3	1	0	0
4	0	1	0
5	0	0	1
6	0	1	1
7	0	1	1

figure 3
nuage projeté
dans **E**

	a	b	c	d	e
u_A	1	1	0	1	0
u_B	0	0	1	0	1
u_C	0	0	0	1	1

figure 4
système
engendrant **E**

	a	b	d
1	1	1	1
2	1	1	1
3	1	0	1

figure 5
sous-tableau
$(\mathbf{A},\mathbf{I}_A)$

	c	e
4	1	1
6	1	1
7	1	1

figure 6
sous-tableau
$(\mathbf{B},\mathbf{I}_B)$

	d	e
1	1	1
5	1	1
6	1	1
7	1	1

figure 7
sous-tableau
$(\mathbf{C},\mathbf{I}_C)$

8. UNE CONCLUSION SUR LES MÉTHODES POUR TABLEAU DE VARIABLES BINAIRES

Nous disposons avec la méthode de classification MNDBIN, la méthode de classification croisée CROBIN et la méthode d'analyse factorielle ACPBIN ou AFB de 3 méthodes pour analyser des tableaux de variables binaires. Toutes ces méthodes ont un même but : rechercher une structure satisfaisante, la plus proche possible du tableau initial, afin d'en extraire plus facilement un certain nombre d'informations.

A chaque fois, le critère associé à chaque méthode représente une mesure de l'écart entre les deux tableaux. Si les pondérations initiales sont toutes égales à 1, le critère représente exactemant le nombre d'éléments différents entre le tableau initial et le tableau recherché. En résumé :

- la méthode MNDBIN permet de représenter les lignes (ou les colonnes) du tableau initial par un ensemble restreint de noyaux ou de vecteurs
- la méthode CROBIN fourni un tableau résumé du tableau initial.
- les méthodes ACPBIN et AFB fournissent des axes binaires et des recouvrements de I et J.

Dans un premier temps, nous proposons un modèle matriciel (non probabiliste) résumant toutes ces méthodes. Puis, nous étudions dans quelles mesures celles-ci peuvent être appliquées.

8.1 MODELE MATRICIEL ASSOCIÉ AU MÉTHODES

Nous considérons ici un modèle matriciel général, résumant toutes les méthodes et utilisant les opérations "et" et "ou" de l'algèbre booléenne.

Le modèle général est le suivant :

$$\mathbf{X} = \mathbf{Y} + \mathbf{R}$$

où :

X est le tableau initial (n,p),

Y le tableau (n,p) fourni par les méthodes,

R la matrice (n,p) d'ajustement ou matrice des erreurs.

Toutes les méthodes recherchent un tableau **Y** tel que la quantité suivante soit

$$\mathbf{W} = \sum_{i\in I}\sum_{j\in J} |x_i^j - y_i^j| = \sum_{i\in I}\sum_{j\in J} |r_i^j|$$

où :

r_i^j est le terme de la $i^{ème}$ ligne et $j^{ème}$ colonne de **R**,

x_i^j est le terme de la $i^{ème}$ ligne et $j^{ème}$ colonne de **X**,

y_i^j est le terme de la $i^{ème}$ ligne et $j^{ème}$ colonne de **Y**.

Les méthodes diffèrent simplement par la nature de la matrice **Y**.

Pour la méthode MNDBIN recherchant une partition P de I en K classes, le tableau **Y** peut se décomposer de la façon suivante :

$$\mathbf{Y} = \mathbf{S} \times \mathbf{L}$$

où :

S est la matrice (n,K) disjonctive associée à la partition P,

L est la matrice (K,p) des noyaux.

Si la méthode recherche une partition Q de J en K classes, **S** correspond à la matrice des noyaux des classes et **L** à la matrice disjonctive associée à la partition Q. Dans les deux cas, l'application de cette méthode aboutit à un tableau **Y** de structure parfaite (correspondant au cas où les éléments de l'ensemble à classifier sont identiques aux noyaux des classes auxquelles ils appartiennent).

Pour la méthode ACPBIN ou AFB recherchant un système de M axes maximisant l'inertie du nuage projeté des individus, le tableau **Y** peut se décomposer comme ci-dessus, la différence étant que la matrice **S** n'est pas nécessairement disjonctive :

$$\mathbf{Y} = \mathbf{S} \times \mathbf{L}$$

où :

S est la matrice (n,M) du nuage projeté des individus,

L est la matrice (M,p) des axes du sous-espace de projection.

On a un shéma symétrique pour l'application au nuage des variables. La matrice **S** n'étant pas disjonctive, le critère associé à cette méthode est nécessairement inférieur à celui fournit par MNDBIN (dans la mesure où le nombre de classes est égal au nombre d'axes). Le tableau **Y** n'a pas la même forme que dans le cas précédent, la structure parfaite correspond ici au cas où les vecteurs réponses des individus appartiennent au sous-espace de projection. Finalement, **Y** se raproche d'avantage du tableau initial tout en conservant une structure débouchant sur des interprétations simples.

Pour la méthode CROBIN recherchant simultanément une partition P de I en K classes et une partition Q de J en M classes, le tableau **Y** peut se décomposer comme suit :

$$\mathbf{Y} = \mathbf{S} \times \mathbf{L} \times \mathbf{T}$$

où :

S est la matrice (n,K) disjonctive associée à la partition P,

T est la matrice (M,p) disjonctive associée à la partition Q,

L est la matrice (K,M) des noyaux.

Le tableau **Y** a ici une structure parfaite menant à des interprétations simples.

Le problème qui se pose est de savoir quelle méthode appliquée. Cela dépend de l'analyse souhaitée par l'utilisateur mais également de la nature du tableau initial.

Dans le paragraphe suivant, nous revenons en détail sur ces points.

8.2 UTILISATION DES MÉTHODES

Au problème :

trouver le tableau structuré Y le plus proche possible du tableau initial X,

les méthodes étudiées ici fournissent une solution suivant le type d'analyse souhaitée. Nous proposons une stratégie d'application de ces méthodes. Cette stratégie ne découle pas d'une étude théorique mais résulte des nombreuses applications effectueés sur différents tableaux de données binaires. Le but de ces applications est d'approcher et d'avoir ainsi une idée de la structure du tableau initial.

Si le tableau initial a une structure voisine d'une structure parfaite (que ce soit en lignes ou en colonnes), la méthode MNDBIN la retrouve. Le critère indique directement la différence entre le tableau initial **X** et le tableau parfait obtenu **Y**. L'application des autres méthodes n'aboutit à aucune amélioration notable du critère. C'est le cas des méthodes ACPBIN et ACP qui fournissent non plus des partitions mais des recouvrements. Un tel tableau initial se résume plus simplement par une partition plutôt que par un recouvrement. La méthode CROBIN permet de rechercher un tableau **Y** qui posséde une structure parfaite à la fois sur les lignes et les colonnes. Toutefois, si telle est la structure initiale, il suffit d'appliquer la méthode MNDBIN sur les lignes puis sur les colonnes du tableau initial et de construire ensuite le tableau associé à ce couple de

Si la structure du tableau initial s'éloigne d'une structure parfaite, il devient intéressant d'appliquer toutes ces méthodes. La méthode MNDBIN fournit une première indication en recherchant des partitions de I ou de J. L'analyse du critère permet de situer, dans un premier temps, l'éloignement entre le tableau initial et le tableau correspondant à la situation idéale. En outre, la méthode fournit des recouvrements de I ou de J et plus généralement des sous-tableaux homogènes en valeurs 1. La méthode CROBIN peut compléter cette étude en recherchant simultanément des partitions des lignes et des colonnes et donc des sous-tableaux homogènes en valeurs 0 ou 1. Enfin, les méthodes ACPBIN ou AFB permettent d'approfondir la connaissance de cette structure et fournissent des recouvrements et des sous-tableaux de façon plus précise (le critère est meilleur).

En résumé, toutes ces méthodes permettent de rechercher une structure interprétable, la plus proche possible du tableau initial. Dans un premier temps, on applique la méthode MNDBIN. Puis, suivant la valeur du critère, on applique ou non les autres méthodes pour rechercher des recouvrements et des sous-tableaux homogènes. Les structures mises en évidence sont les suivantes :

- une partition de I (resp. de J), un recouvrement de J (resp. de I) menant à des sous-tableaux homogènes en valeurs 1 en croisant la partition et le recouvrement (MNDBIN),

- un couple de partitions de I et de J et des sous-tableaux homogènes en croisant ces deux partitions (CROBIN),

- un couple de recouvrements de I et de J menant à des sous-tableaux homogènes en valeurs 1 en croisant ces deux recouvrements (ACPBIN ou AFB).

9. PROGRAMMES ET APPLICATIONS

Comme dans les cas précédents, le programme ACPBIN (et une version de l'AFB, utilisant ACPBIN et les points initiaux que nous avons proposés) a été intégré au logiciel SICLA.

En utilisant la commande HASBIN, nous avons généré plusieurs tableaux de données binaires afin de tester nos programmes et notamment les différentes possibilités pour les points initiaux. Il ressort que le choix d'un point initial (une partition ou un axe) engendré de façon aléatoire conduit toujours, après plusieurs tentatives, au meilleur résultat. Le choix initial d'une matrice ligne identiquement nulle ne conduit que dans certains cas au meilleur critère. Les algorithmes d'ajout, de suppression et d'amélioration (ACPBIN) utilisés par AFB n'aboutissent pas toujours à des résultats satisfaisants.

Nous proposons d'appliquer nos méthodes aux données MERO (présentées dans le chapitre 2). La meilleure partition en 5 classes des individus est associée à un critère de **100**. La méthode ACPBIN n'améliore pas ce critère (5 axes ont été demandés), au mieux on retrouve la valeur **100**. Cela montre que pour 5 classes, la partition des individus suffit pour dégager une structure de ces données.

Nous allons donc étudier la partition en 7 classes (mise en évidence dans le chapitre 2) qui se composait comme suit :

A_1 **4, 9 , 14, 21, 22, 24, 25, 26, 28, 30, 31, 32, 33, 36, 46, 49, 52, 53**

A_2 **15, 23, 47, 54, 55, 56, 57, 58**

A_3 **11, 17, 27, 34, 48, 51, 59**

B_1 **2, 6, 8, 12, 13, 18, 19, 29, 44, 45**

B_2 **1, 10, 16, 37, 38, 39, 40, 41, 42, 43**

Les deux autres classes avaient des effectifs faibles et regroupaient les éléments (**7,50**) et (**3,5,20,35**). Le critère obtenu était de **82** différences entre données initiales et noyaux. Les caractéristiques de chacun de ces groupes ont été détaillées dans le chapitre 2, paragraphe 6.2.

Nous allons maintenant appliquer la méthode MNDBIN à l'ensemble des 29 variables en demandant 7 classes. La meilleur essai donne un critère de **99**. En pages 200 et 201, nous indiquons les résultats suivants :

- la partition obtenue dont les classes permettent de définir les 7 vecteurs binaires engendrant le sous-espace de projection,
- la tableau des noyaux qui est aussi le tableau du nuage projeté,
- le tableau des homogénéités par classe et par variable montrant également la qualité de représentation des individus sur les axes,
- le tableau initial réordonné en respectant les classes de la partition,
- la partition du nuage projeté construite à partir du tableau des noyaux,
- les 7 sous-tableaux homogènes en valeurs 1 qui se déduisent également du tableau des valeurs idéales.

A chaque classe de la partition on peut associer un sous-ensemble d'individus (ceux se projettant sur l'axe défini à partir de la classe). En croisant cette classe et ce sous-ensemble, on obtient alors un sous-tableau homogène en valeurs 1. Ici, on constate que les classes 5 et 6 conduisent à une partition en deux classes des individus (d'après le tableau des valeurs idéales). On retrouve exactement la décomposition en deux groupes **A** et **B** que nous avions obtenue par l'application de MNDBIN (en demandant 2 classes d'individus). On retrouve aussi les mêmes caractéristiques :

sous-tableau 5 **A** est caractérisé par **C15, C22** et **C24,**

sous-tableau 6 **B** est caractérisé par **C17, C29, C33, C38** et **C39.**

En plus de ce résultat, l'application de MNDBIN aux variables permet aussi de caractériser certains sous-groupes de **A** et de **B**. On retrouve cette fois les principaux résultats de l'application de MNDBIN aux individus (en demandant 7 classes). En effet, en examinant les sous-tableaux homogènes, on constate que :

sous-tableau 2 A_3 est caractérisé par **C26, C30** et **C35,**

sous-tableau 3 B_1, auquel sont ajoutés **7** et **50**, est caractérisé par **C31, C34** et **C41,**

sous-tableau 4 A_1, auquel est ajouté **3**, est caractérisé par **C17, C29, C33, C38** et **C39,**

sous-tableau 7 A_2 est caractérisé par **C01, C19** et **C28.**

Le sous-tableau 1 ne permet pas de caractériser le groupe B_2, les individus **40, 41, 42** et **43** de ce groupe ne se projettant que sur l'axe associé à **B**. Parmi les individus ne figurant pas dans les 5 principaux groupes, **7** et **50** ont été affecté à B_1, **3** à A_1.

Le nuage projeté est inclus dans B^7, mais se résume en 10 points différents : on obtient ainsi une partition où toute classe est constituée des individus se projetant sur un même point. Cette partition permet pratiquement de retrouver les groupes A_1, A_2, A_3 et B_1; le groupe B_2 est décomposé est 2 parties (dont l'une contient **40, 41, 42** et **43**). Les éléments **5, 7, 20, 34** (qui appartient à A_1), **35** et **50** ont des projections particulières.

Dans le chapitre 2, deux applications de MNDBIN ont abouti aux deux groupes **A** et **B** d'une part, et à des sous-groupes d'autres part. La seule application de MNDBIN au variables permet de retrouver pratiquement les mêmes résultats (d'après l'étude des sous-tableaux homogènes en valeurs 1. De plus, la partition du nuage projeté permet de distinguer des plaques-boucles particulières. Si nous avions effectué cette analyse en premier lieu, les résultats obtenus nous aurait alors permis d'avoir une idée du nombre de classes d'individus à rechercher. Par exemple, pour les données MERO, si on demande 5 axes binaires, le recouvrement des individus est paratiquement une partition.

Nous avons ensuite appliqué la méthode ACPBIN en demandant 7 axes. Le critère obtenu est de **77**, c'est-à-dire inférieur à celui des autres applications. En page 202, on

- le détail des 7 axes obtenus,
- la partition du nuage projeté,
- les sous-tableaux homogènes en valeurs 1.

Ces sous-tableaux sont obtenus en croisant le recouvrement des variables et celui des individus, tous deux fournis par la méthode. Une partie regroupe des variables n'ayant un comportement homogène que sur un sous-ensemble d'individus. Cela a alors permis d'améliorer le critère. Par exemple, le groupe $\mathbf{A_1}$ est décomposé et se retrouve dans la définition de deux sous-tableaux (1 et 4). Cela est une conséquence de la faible homogénéité de la variable **C17** dans ce groupe (**61%** des individus de $\mathbf{A_1}$ possédent ce caractère). Les autres sous-tableaux permettent de retrouver les groupes **B** (sous-tableau 5), $\mathbf{B_1}$ (6), $\mathbf{B_2}$ (7), $\mathbf{A_2}$ (2) et $\mathbf{A_3}$ (3). Par rapport à la situation précédentes (où $\mathbf{B_2}$ de cardinal 10 était divisé), les sous-tableaux obtenus ici ont un pourcentage de valeurs 1 plus important ($\mathbf{A_1}$ de cardinal 18 est divisé). On retrouve le même type de résultats en consultant la partition du nuage projeté (contenant aussi 10 classes).

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires


          valeur du critere a chaque essai :  99


          partition en ligne

          classe numero  1  :  effectif 3
            C14   C32   C40

          classe numero  2  :  effectif 3
            C26   C30   C35

          classe numero  3  :  effectif 3
            C31   C34   C41

          classe numero  4  :  effectif 5
            C17   C29   C33   C38   C39

          classe numero  5  :  effectif 3
            C15   C22   C24

          classe numero  6  :  effectif 6
            C16   C23   C25   C36   C37   C42

          classe numero  7  :  effectif 3
            C01   C19   C28


          tableau des valeurs ideales

          000000000011111111112222222222333333333344444444445555555555
          123456789012345678901234567890123456789012345678901234567890123456789
      1   1         1     1                        111
      2         1   1     1          1      1              1  1        1
      3    1   111    11     11          1                 1     1
      4      11     1    1       11 111 1 11111 1           1  1  11
      5     111 1 1 1  11 1  111111111 1111111          11111111111111
      6   11   1 1 1 11  1 11          1        111111111
      7                  1       1                        1      11111


          tableau initial reordonne

     000000000011111111112222222222233333333334444444444455555555555
     123456789012345678901234567890123456789012345678901234567890123456789
-----------------------------------------------------------------------------
C14  1         1     1    1    1 1         111
C32  1         1     1                     1111111
C40  11   1   1 11  1 1
-----------------------------------------------------------------------------
C26         1   1      1          1         1 11111    1 11        1
C30         1   1      1          1       1            1  11       1
C35             1      1          1       1            1  1        1
-----------------------------------------------------------------------------
C31   1   1 1   11    11          1              1     1
C34   1    11             1        11     1            1 1
C41   1   1111  11     11          1              1     1
-----------------------------------------------------------------------------
C17   1 1 11 1 1 1          1 11   1  111            1  11 11
C29    11    1            111 111 1 1111 11         11  1   1
C33    11    1    1        11 111 1 1111  1          1  1   1
C38     1  1 1 1  1  1    11 11111 11111 1           1  11111      1
C39    111    1    1     1111 111 111111111         1  1  11
-----------------------------------------------------------------------------
C15    1 1 1 1 1  11 1      111111 1111111          111111111111111
C22    111 1 1 1  11 1  111111111 1111111          11111 11111111
C24    111 1 1 1  11 1  111111111 1111111          111111111111111
-----------------------------------------------------------------------------
C16  11 1 1   1 11  1 11111       1        111111111
C23  11   1 1 1 11  1 11          1        111111111
C25  11   1 1 1 11  1 11          1        111111111
C36  11   1   1 11  1 111                  111111111
C37  11   1 1 1 11  1 11          1      1 111111111
C42  11   111   1   1 1 1           1      11111111  1 1
-----------------------------------------------------------------------------
C01               1       1                        1      11111
C19                                                1     111111
C28      1        1       1                        1      11111
-----------------------------------------------------------------------------
```

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires
```

tableau des homogeneites

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15
1	100	67	100	100	100	67	100	100	100	100	100	67	67	100	100
2	100	100	100	100	100	100	67	100	100	100	100	100	100	100	100
3	100	100	100	100	100	67	67	100	67	100	100	67	67	100	100
4	100	80	60	100	80	80	60	100	100	100	60	100	80	60	100
5	100	100	100	67	100	100	100	100	100	100	100	100	100	100	100
6	100	100	100	83	100	100	83	67	100	83	100	100	83	100	100
7	100	100	100	100	67	100	100	100	100	100	100	100	100	100	67

	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
1	100	100	67	100	100	67	100	100	100	100	67	100	67	100	100
2	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
3	100	100	67	100	100	100	100	100	100	100	100	100	67	100	100
4	100	80	100	80	60	80	100	100	100	100	80	80	80	60	80
5	100	100	100	100	67	67	67	100	100	100	100	100	100	100	100
6	100	100	100	83	50	83	83	100	100	100	100	100	100	67	100
7	100	100	100	100	100	100	100	67	100	100	100	100	100	100	100

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
1	100	100	100	100	100	100	67	67	67	67	67	67	67	100	100
2	100	100	100	67	100	100	67	100	67	67	67	67	67	100	100
3	100	100	100	100	67	100	100	100	100	100	100	100	100	67	100
4	80	100	100	60	60	80	100	100	100	100	100	100	100	100	80
5	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
6	83	100	100	100	83	100	83	100	100	100	100	100	100	100	100
7	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100

	46	47	48	49	50	51	52	53	54	55	56	57	58	59
1	100	100	100	100	100	100	100	100	100	100	100	100	100	100
2	100	100	100	100	67	100	67	100	100	100	100	100	100	100
3	100	100	100	100	100	100	67	100	100	100	100	100	100	100
4	100	100	100	100	60	80	60	100	100	100	100	100	100	80
5	100	100	100	100	100	67	100	100	100	100	100	100	100	100
6	100	100	83	100	83	100	100	100	100	100	100	100	100	100
7	100	100	100	100	100	100	100	67	100	100	100	100	100	100

```
COMMANDE : MNDBIN <> nuees dynamiques sur variables binaires


        partition du nuage projeté


     1-  projete  0000010  :  40  41  42  43  45

     2-  projete  0000100  :   5  20  35

     3-  projete  0000101  :  15  23  47  54  55  56  57  58

     4-  projete  0001100  :   3   4   9  14  21  22  24  25  26  28  30  31  32  33  36
                               46  49  52  53

     5-  projete  0010010  :   2   6   8  12  13  18  19  29  44

     6-  projete  0010100  :  50

     7-  projete  0100100  :  11  17  27  59

     8-  projete  0101100  :  34

     9-  projete  0110100  :   7

    10-  projete  1000010  :   1  10  16  37  38  39


        sous-tableaux homogenes en valeurs 1

        sous-tableau 1 :
             classe 1  :   C14   C32   C40
            individus  :    1  10  16  21 26  28  37  38  39

        sous-tableau 2 :
             classe 2  :   C26   C30   C35
            individus  :    7  11  17  27  34  48  51  59

        sous-tableau 3 :
             classe 3  :   C31   C34   C41
            individus  :    2   6   7   8  12  13  18  19  29  44  50

        sous-tableau 4 :
             classe 4  :   C17   C29   C33   C38   C39
            individus  :    3   4   9  14  21  22  24  25  26  28  30  31  32  33  34
                           36  46  49  52  52

        sous-tableau 5 :
             classe 5  :   C15   C22   C24
            individux  :    3   4   5   7   9  11  14  15  17  20  21  22  23  24  25
                           26  27  28  30  31  32  33  34  35  36  46  47  48  49  50
                           51  52  53  54  55  56  57  58  59

        sous-tableau 6 :
             classe 6  :   C16   C23   C25   C36   C37   C42
            indivisus  :    1   2   6   8  10  12  13  16  18  19  29  37  38  39  40
                           41  42  43  44  45

        sous-tableau 7 :
             classe 7  :   C01   C19   C28
            individus  :   15  23  47  54  55  56  57  58
```

```
COMMANDE : ACPBIN  <> analyse en composantes principales sur variables binaires


        le critere obtenu est :   77


        axes binaires


        CCCCCCCCCCCCCCCCCCCCCCCCC
        01111122222223333333333444
        14567923456890123456789012

    1   00100010100010001000011000
    2   10100110100100000000000000
    3   00100010101001000010010000
    4   00101010100000000000010000
    5   00010001010000000001100001
    6   00000000000000100100000011
    7   01000000001000010000000000


        nuage projete :


        00000000011111111112222222222333333333344444444445555555555
        12345678901234567890123456789012345678901234567890123456789

    1   00110000100001000001110111010111101100000000010010001000000
    2   00001000000000100000001000000000000000000000001000000111110
    3   00000010001000001000000000100000010000000000000100100000001
    4   00010010101000000000010110000001110000000000010011011000000
    5   11000101010110010110000000001000000011111111100000000000000
    6   01000111000110000110000000001000000000000000100000100000000
    7   10000000010000010000000000000000000001111111000000000000000


        partition du nuage projeté

   1-  projete  0000100  :  45  52

   2-  projete  0000101  :   1  10  16  37  38  39  40  41  42  43

   3-  projete  0000110  :   2   6   8  12  13  18  19  29  44

   4-  projete  0001010  :  50

   5-  projete  0010000  :  17  27  48  51  59

   6-  projete  0011000  :  11  34

   7-  projete  0011010  :   7

   8-  projete  0100000  :   5  15  23  47  54  55  56  57  58

   9-  projete  1000000  :   3  14  20  21  26  28  30  31  35  36

  10-  projete  1001000  :   4   9  22  24  25  32  33  46  49  53

        sous-tableaux homogenes en valeurs 1


        sous-tableau 1 :
               axe 1  :   C15   C22   C24   C29   C33   C38   C39
           individus  :     3   4   9  14  20  21  22  24  25  26  28  30  31  32  33
                            35  36  46  49  53

        sous-tableau 2 :
               axe 2  :   C01   C13   C19   C22   C24   C28
           individus  :     5  15  23  47  54  55  56  57  58

        sous-tableau 3 :
               axe 3  :   C15   C22   C24   C26   C30   C35   C38
           individus  :     7  11  17  27  34  48  51  59

        sous-tableau 4 :
               axe 4  :   C15   C17   C22   C24   C38
           individus  :     4   7   9  11  22  24  25  32  33  34  46  49  50  52  53

        sous-tableau 5 :
               axe 5  :   C16   C23   C25   C36   C37   C42
           individus  :     1   2   6   8  10  12  13  16  18  19  29  37  38  39  40
                            41  42  43  44  45

        sous-tableau 6 :
               axe 6  :   C31   C34   C41   C42
           individus  :     2   6   8  12  13  18  19  29  44  50

        sous-tableau 7 :
               axe 7  :   C15   C26   C32
           individus  :     1  10  16  37  38  39  40  41  42  43
```

CONCLUSION

Les méthodes de classification que nous avons proposées dans cette étude reposent toutes sur un principe simple d'homogénéité, consistant à ne pas dénaturer les données. Ainsi, que les variables soient binaires ou qualitatives, les résultats obtenus sont directement interprétables par rapport au données initiales. Le critère associé à la partition ainsi que les indices d'aides à l'interprétation constituent autant d'éléments descriptifs simples, facilement compréhensibles, et permettant de juger rapidement de l'adéquation entre la structure initiale (qui est celle à caractériser) et la structure "idéale" la plus proche (qui est celle fournie par nos méthodes). Ceci a été possible en définissant un critère à partir de la distance L_1. Toutes ces considérations permettent de distinguer les algorithmes contruits ici (MNDBIN, MNDDIJ, MNDORD) de ceux utilisés habituellement (MNDQAN, MNDQAL)

Nous avons ensuite repris, sous de nouvelles hypothèses, les méthodes de classification automatique pour tableau de variables binaires (MNDBIN et CROBIN). L'utililisation de vecteurs de pondérations et la définition d'une inertie binaire aboutissent à une relation de type Huyghens. Celle-ci nous a alors permis de replacer nos méthodes dans un contexte plus habituel. On retrouve ainsi des propriétés analogues à celles formulées habituellement dans l'approche euclidienne (conservation du centre médian, relation de décomposition de l'inertie binaire, indices d'aides à l'interprétation, mesure d'information pour données binaires). Une limite a été entrevue en ce qui concerne l'extension à la classification ascendante hiérarchique et le pseudo-indice de Ward (l'indice de l'augmentation de l'inertie binaire).

Notons que ce type d'approche a également été envisagée pour les données qualitatives. Il se trouve alors que, dans le cas ordinal, les propriétés précédentes sont toujours vraies lorsque les données sont transformées par le codage binaire additif (d'ailleurs, la méthode utilisée dans le cas binaire s'applique aussi ici). Il peut être intéressant d'approfondir ces résultats et d'étudier le type d'inteprétation qui s'en déduit. Par contre, dans le cas nominal, nous n'obtenons aucun résultat intéressant.

Pour terminer et compléter cette étude, nous avons envisagé d'étudier une méthode d'analyse en composantes principales spécifiques aux données binaires. Notre approche s'inspire de l'approche habituelle pour données quantitatives. Un certain nombre de résultats sont mis en évidence. Tout d'abord, l'analyse sous contrainte est équivalente à la méthode de classification MNDBIN sur l'ensemble des variables (nous avons un résultat analogue pour les données quantitatives et la méthode MNDQAN). Cela a permis de mettre en évidence des résultats qui n'apparaissaient pas directement lors de l'application de MNDBIN à l'ensemble des individus. D'une part, la recherche d'une partition de l'un des deux ensembles revient à rechercher un recouvrement de l'autre ensemble. D'autre part, la partition et le recouvrement permettent de définir des sous-tableaux homogènes en valeurs 1. La suppression de la contrainte aboutit à une méthode déjà proposée (ACPBIN), mais formulée en des termes très différents. Nous lui donnons ici une interprétation en terme d'inertie binaire. Pour minimiser cette inertie, la méthode recherche simultanément un recouvrement de l'ensemble des variables et un

recouvrement de l'ensemble des individus. Il nous semble alors intéressant d'étudier de plus près les techniques de recherche de recouvrements dans le cas de données binaires.

Des limites sont apparues dont la cause essentielle est l'abscence d'une structure d'espace vectoriel pour l'espace binaire. En fait, nous n'avons pu dégager une approche métrique sans reproche pour cet espace (cela expliqe les limites de l'algorithme ACPBIN). Cela nous amène alors à regarder du coté des techniques reposant sur les mathématiques des structures finies (théorie des graphes, treillis,...), d'autant plus que certaines d'entre elles ont pour but de rechercher des sous-tableaux homogènes en valeurs 1.

BIBLIOGRAPHIE

AMSTRONG R.D., FROME E.L. et KUNG D.S. (1979), "A Revised Simplex Algorithm for the Absolute Deviation Curve Fitting Problem". Commun. Statist. B8, pages 175-190.

AMSTRONG R.D. et KUNG D.S. (1978), "Least Absolute Value Estimates for a Simple Linear Regression Model. Applied Statistics 27, pages 325-328.

ANDERBERG M.R. (1973), "Cluster Analysis for Application". Academic Press, New-York.

ANDERSON J.A. (1983), "Robust Inference Using Logistic Models". Bull. Int. Statist. Inst., 48, 35-53.

ANDERSON J.A. (1984), "Regession and Ordered Categorical Variables". Journal of the Royal Statistical Society, B, 46, pages 149-192.

BARLOW W.E. and FEIGL P. (1985), "Analysis Binomial Data with a Nonzero Baseline using G.L.I.M.". Computationnal Statistics and Data Analysis, Vol 3, nº 3.

BARTHOLOMEW D.J. (1980), "Factors Analysis for Categorical Data". Journal of the Royal Statistical Society, B, 42, nº 3, pages 293-321.

BARTHOLOMEW D.J. (1984), "Scaling Binary Data Using a Factor Model". Journal of the Royal Statistical Society, B, 46, nº 1, pages 120-123.

BEDALL F.K. et ZIMMERMANN H. (1979), "The Mediacentre". Applied Statistics, pages 325-328.

BENZECRI J.P. (1973), "L'Analyse des Données (1. la Taxinomie, 2. l'Analyse des Correspondances)". Dunod.

BENZECRI J.P. (1977), "Sur l'Analyse des Tableaux Binaires Associés à une Correspondance Multiple". Les Cahiers de l'Analyse des Données, Vol 2 nº 1, pages 55-71.

BERTIN J. (1977), "La Graphique et le Traitement Graphique de l'Information". Flammarion, Paris.

BISHOP Y.M.M., FIENBERG S.E. and HOLLAND P.W. (1980), "Discrete Multivariate Analysis". MIT Press.

BLOOMFIELD P. et STEIGER W. (1980), "Least Absolute Deviations Curve-Fitting". SIAM J. Sci. Statis. Comput. 1, pages 290-300.

BOCK H.H. (1986), "Loglinear Models and Entropy Clustering Methods for Qualitative Data". Classification as a Tool of Research, W. Gaul and M. Schader (editors).

BOCK R.D. et LIEBERMANN M. (1970), "Fitting a Response Model for N Dichotomously Scored Items". Psychometrika 35, pages 179-197.

BROWN B.M. (1983), "Statistical Uses of the Spatial Median". Journal of the Royal Statistical Society, B, pages 25-30.

BUSER M.W. and BARONI-URBANI C. (1982), "A Direct Nondimensional Clustering Methode for Binary Data". Biometrics 38, pages 351-360.

CAILLEZ F. et PAGES J.P. (1976), "Introduction à l'Analyse des Données". SMASH.

CARAUX G. (1984), "Réorganisation et représentation Visuelle d'une Matrice de Données Numériques - Un Algorithme Itératif". Revue de Statistique Appliquée, 32, 4, pages 5-23.

CAZES P., BAUMERDER A., BONNEFOUS S. et PAGES J.P. (1977), "Codage et Analyse des Tableaux Logiques - Introduction à la Pratique des Variables Qualitatives". Extrait des Cahiers du Buro, 27.

CELEUX G. (1988), "Classification et Modèle". RSA vol 36, nº 3.

CELEUX G., DIDAY E., GOVAERT G., LECHEVALLIER Y., RALAMBONDRAINY H. (1989), "Classification Automatique des Données". Dunod.

CELEUX G. et GOVAERT G. (1989), "Clustering Criteria for Discrete Data and Latent Class Models". Rapport INRIA.

CHRISTOFFERSSON A. (1975), "Factor Analyses of Dichotomized Variables". Psychometrika, vol 40, nº 1.

CORMAK R.M. (1971), "A Review of Classification". Journal of the Royal Statistical Society, A, nº 134.

COX D.R. (1969), "Analyse de Données Binaires". Dunod, Paris (1972 pour la traduction).

DIDAY E. ET AL. (1980), "Optimisation en Classification Automatique". INRIA, Rocquencourt.

DIDAY E., LEMAIRE J., POUGET J. et TESTU F. (1982), "Eléments d'Analyse de Données". Dunod.

DIDAY E. et SIMON J.C. (1980), "Cluster Analysis. Digital Pattern Recognition". Springer-Verlag, pages 47-94.

DUPOND-GATELMAND C., (1978), "Deux Méthodes d'Analyse des Données Qualitatives. Typologie avec Codage Adaptatif des Préférences". Thèse de 3ème cycle, Paris IX.

DURAN B.S. et ODELL P.L. (1974), "Cluster Analysis". Springer-Varlag.

EVERITT B.S. (1984), "A Introduction to Latent Variable Models". Chapman and Hall, London.

FICHET B. et GBEGAN A. (1985), "Analyse Factorielle des Correspondances sur Signes de Présence-Absence". IV^{éme} Journées Internationales d'Analyse des Données et Informatique, pages 209-238. Versailles.

FICHET B. et LE CALVÉ C. (1984), "Structure Géométrique des Principaux Indices de Dissimilarité sur Signes de Présence-Absence". Statistiques et Analyse des Données, Vol 9 nº 3, pages 11-44.

FINNEY D.J. (1978), "Statistical Method in Biological Assay". Hafner, New York.

FISHER R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems". Eugenics, vol 7.

FLAMENT Cl. (1976), "L'Analyse Booléenne de Questionnaire". Mouton, Paris.

GENTLE J.E., SPOSITO V.A. and NARULA S.C. (1988), "Algorithmes for Unconstrained L_1 Simple Linear Regression". Computational Statistics and Data Analysis 6, pages 335-339.

GUENOCHE A. (1985), "Classification Using Dilemna Functions". Computational Statistics Quarterly, 2, 1, pages 103-108.

GUENOCHE A. et MONJARDET B. (1987), "Méthodes Ordinales et Combinatoires en Analyses des Données". Mathématique et Sciences Humaines, 25ème années, nº 100, pages 5-47.

GORDON A.D. (1981), "Classification". Chapmann and Hall.

GOVAERT G. (1983), "Classification Croisée". Thèse de Doctorat d'État, Université Pierre et Marie Curie, Paris VI.

GOVAERT G (1988), "Classification Binaire et Modèle". Rapport de Recherche INRIA, nº 949.

GOWER J.C. (1974), "Maximal Predictive Classification". Biometrics 30, 643-654.

GOWER J.C. (1974b), "Algorithm AS78 : The mediacentre". Applied Statistics, pages 466-470.

GOWER J.C. and LEGENDRE P. (1986), "Metric and Euclidean Properties of Dissimilarity Coefficients". Journal of Classification 3, pages 5-48.

HABERMAN S.J. (1978), "Analysis of Qualitative Data". Academic Press, New York.

JAMBU M. (1972), "Techniques de Classification Automatique Appliqués à des Données de Sciences Humaines". Thèse de Doctorat de 3ème cycle, Paris.

JAMBU M. (1978), "Classification Automatique pour l'Analyse des Données". Dunod.

JOSVANGER L.A. et SPOSITO V.A. (1983), "L_1-Norm Estimates for the Simple Regression Problem". Commun.Statist. B12, pages 215-221.

KOSFELD R. (1986), "Computational Aspects and Small Sample Properties of Multi-Dimensional Medians". Computational Statistics Quarterly 1, pages 21-36.

LAFAYE J.Y. (1979), "Une Méthode de Discrétisation de Données Continues". RSA Vol. 27, nº 2.

LEBART L., MORINEAU A. et TABARD N. (1977), "Techniques de la Description Statistique". Dunod.

LEFEBVFRE B. (1977), "Construction de Sous-Tableaux Homogènes dans un Tableau de Données Binaires". Colloque IRIA, Analyse de Données et Informatique, tome 1 IRIA, pages 93-97.

LEFEBVRE B. et LOSFELD J. (1979), "Formalisation Constructive de la Notion de Classe Polythétique pour un Tableau de Données Binaires". IIème Journées de Versailles.

LEREDDE H. et PERIN P. (1980), "Les Plaques-Boucles Mérovingiennes". Dossiers de l'Archéologie, nº 42.

LERMAN I.C. (1973), "Etude Distributionnelle de Statistiques de Proximité entre Structures Finies de même Type - Application à la Classification Automatique". ISUP, Cah. Bur. Universit. Rech. Oper. nº 19.

MICKEY M.R., MUNDEL P. and ENGELMAN L. (1983), "Boolean Factor Analysis". Manuel BMDP, pages 538-692.

MKHADRI A. (1989), "Pondération des Variables pour la Classification Binaire". Rapport de Recherche I.N.R.I.A., nº 1079.

MONJARDET B. (1980), "Théorie des Graphes et Taxonomie Mathématiques". Regards sur la Théorie des Graphes, Hansen P. et al., Eds Presses Polytechniques Romandes, pages 111-125.

MUTHEN B. (1978), "Contributions to Factor Analysis of Dichotomous Variables". Psychometrika, vol 43, nº 4.

PLACKETT P.L. (1981), "The Analysis of Categorical Data". 2ème ed. Griffin, London.

RALAMBONDRAINY H. (1988), "Etude des Données Qualitatives par les Méthodes Typologiques". Actes au Congrès de l'Association Française de Marketing. Montpelier.

REINERT A. (1983), "Une Méthode de Classification Descendante Hiérarchique : Application à l'Analyse Lexicale Par Contexte". Les Cahiers de l'Analyse des Données, Vol 8 nº 2, pages 187-198.

ROUX M. (1985), "Algorithmes de Classification". Masson.

SAPORTA G. (1968), "Liaisons entre plusieurs Ensembles de Variables et Codage de Variables Qualitatives". Thèse de Doctorat de 3ième cycle, Université de Paris VI.

SCHROEDER A. (1976), "Analyse d'un Mélange de Distribution de Probabilité de même Type". RSA vol 24, nº 1.

SCOTT A. et SYMONS M. (1971), "Clustering Methods Based on Likelihood Ratio Criteria". Biometrics 27.

SNEATH P.H.A et SOKAL R.R. (1963), "Principles of Numerical Taxonomy". Freeman.

SOKAL R.R. et MICHENER C.D. (1958), "A Statistical Method for Evaluating Systematic Relationships". University of Kansas Sciences Bulletin, 38, pages 1409-1438.

TALENG F. (1980), "Selection Typologique de Paramètres de Différents Types". Thèse de 3ème cycle, Paris IX.

TELEGDI L. (1986), "Multidimensional Scaling of Dichotomous Variables". Computer and Automation Institute, Hungarian academy of sciences, MS/22.

TUBBS J.D. (1989), "A Note on Binary Template Matching". Pattern Recognition, 22 (4), pages 359-367.

WARD J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function". Journal of the American Statistical Association 58, pages 238-244.