# Classification automatique et données manquantes

Mohamed Nadif

UNIVERSITE DE METZ
1991
LABORATOIRE DE RECHERCHE EN INFORMATIQUE DE METZ



# THESE

PRESENTEE ET SOUTENUE PUBLIQUEMENT
LE 16 SEPTEMBRE 1991

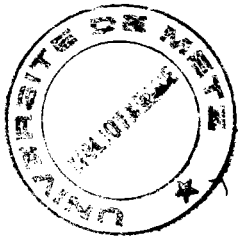
POUR OBTENIR LE TITRE DE

# DOCTEUR DE L'UNIVERSITE DE METZ

SPECIALITE : INFORMATIQUE
MENTION : ANALYSE DES DONNEES

Par

Mohamed NADIF



# SUJET

## CLASSIFICATION AUTOMATIQUE ET DONNEES MANQUANTES

MEMBRES DU JURY

| | |
|---|---|
| Président : | G. GOVAERT, Professeur (COMPIEGNE) |
| Rapporteurs : | G. CELEUX, Chargé de Recherche INRIA (PARIS)<br>J.M. PROTH, Directeur de Recherche INRIA (METZ) |
| Examinateurs : | Y. GARDAN, Professeur (METZ)<br>G. SALLET, Professeur (METZ) |

## *REMERCIEMENTS*

*Je suis heureux d'exprimer ma profonde reconnaissance à Monsieur le Professeur G. GOVAERT qui m'a accueilli dans son laboratoire, pour la confiance qu'il m'a accordée sur le plan de la recherche et le soutien qu'il m'a fourni pendant la durée de ce travail. Qu'il trouve ici l'expression de toute ma reconnaissance et mes remerciements car il me fait l'honneur de présider le jury de cette thèse.*

*Je désire vivement remercier Messieurs G. CELEUX et J.M PROTH d'avoir accepté d'être rapporteurs de cette thèse et de participer à ce jury ainsi que pour les propositions judicieuses qu'ils m'ont suggérées.*

*Je remercie Messieurs Y. GARDAN et G. SALLET d'avoir accepté de juger ce travail et de participer à ce jury.*

*Je remercie l'INRIA pour le soutien financier qu'elle m'a apportée.*

*Il m'est agréable de remercier mes collègues du département informatique de l'IUT de Metz pour leur disponibilité et l'équipe d'Analyse des données dont Monsieur Y. LEMOINE pour ses conseils et Mademoiselle Y. BENCHEIKH pour son amitié.*

*Enfin, j'ai une pensée toute particulière pour Monsieur F. MARCHETTI qui, par sa disponibilité et ses conseils, a contribué à l'élaboration de cette thèse.*

# TABLE DES MATIERES

CHAPITRE II



CHAPITRE III

## PARTIE B : DONNEES QUALITATIVES NOMINALES

CHAPITRE VII



CHAPITRE VIII

# INTRODUCTION

# INTRODUCTION

Le passage d'une analyse relativement simple en une analyse complexe par suite de l'absence de quelques informations est un problème auquel de nombreux statisticiens se consacrent activement depuis ces dernières années. Il est évident que la meilleure manière de traiter les problèmes en présence d'information manquante est encore de ne pas avoir de données manquantes. Cependant, des circonstances malheureuses se présentent parfois dans lesquelles l'information est manquante et son remplacement s'avère difficile. Comme les accidents, l'absence d'information ne peut être prévue mais doit être gérée quand elle est présente. Cette absence peut être imputable à diverses raisons : erreurs de saisie, erreurs d'expérimentation, choix de l'échantillon, impossibilités matérielles, refus de réponses, etc...

Devant ce problème, nous ne pouvons éviter de nous poser la question concernant les causes de l'apparition des données manquantes : sont-elles dues à un hasard ou, au contraire, à des raisons déterministes ? Little et Rubin (1987) ont distingué deux types de données manquantes dont ils donnent les définitions suivantes :
-Données Manquantes au Hasard (DMH) lorsqu'elles peuvent être considérées pour une même variable comme un sous-échantillon de l'échantillon initial.
-Données Manquantes Complètement au Hasard (DMCH) lorsqu'elles constituent un sous-échantillon aléatoire des valeurs prises par l'échantillon initial.

**Exemple.** Considérons une population décrite par deux variables, l'âge et le revenu. Si la non connaissance d'un revenu n'est pas liée à sa valeur, l'occurence de son absence est alors indépendante de sa valeur et la donnée manquante est de type DMH. Si de plus, l'absence de l'information n'est pas liée à l'âge de la personne interrogée, la donnée manquante est alors de type DMCH car l'occurence de son absence est indépendante de toutes les valeurs que prend l'individu qui présente cette donnée manquante.

Divers mécanismes peuvent être à l'origine des données manquantes. La connaissance ou l'ignorance de ces mécanismes est un élément de choix dans l'analyse et l'interprétation des résultats.

Ces mécanismes peuvent être sous le contrôle du statisticien et, dans ce cas, peuvent dépendre par exemple de la sélection de l'échantillon. Si l'échantillonnage est réalisé par probabilité, alors le mécanisme est sous le contrôle du modèle et est dit "ignoré". Les analyses de données sont alors dépendantes des présomptions du mécanisme qui se doit d'être expliqué. D'autre part, les données censurées dépendent d'un événement bien précis. Nous rencontrons par exemple ce type de données dans les analyses médicales, biologiques ou épidémiologiques. Même censurées, ces données sont prises en compte dans l'analyse afin d'éviter les résultats biaisés. Dans ce cas, le mécanisme n'est pas sous le contrôle du statisticien mais est connu. De nombreux ouvrages sont consacrés à l'analyse statistique à partir des données censurées tels que celui de Kalbfleish et Prentice (1980).

Dans de nombreuses analyses, le mécanisme n'est malheureusement pas introduit clairement. Ainsi, nous présumons que ce mécanisme est "ignoré". Il est cependant possible d'inclure le mécanisme dans un modèle statistique en introduisant une distribution R de variables d'indicateurs de réponse qui prennent la valeur 1 si l'item est observé et 0 sinon. Notons que le processus de la création des données manquantes ne peut généralement pas être ignoré. Ainsi, l'absence de réponse dans une enquête sur les revenus est souvent volontaire car leur montant n'est pas ignoré. L'expérience montre que ces données peuvent apporter beaucoup d'informations si nous les considérons comme des modalités de réponses au même titre que les autres modalités (Lebart, Morineau et Tabard 1977, Van der Heijden et Escofier 1988). Après classification des données qualitatives, Facy et Lechevalier (1978) réalisent des tests statistiques de manière à déterminer l'éventualité qu'une donnée manquante soit considérée comme une modalité absente ou comme une nouvelle modalité.
Soit Y un échantillon de taille n d'une variable reélle à valeurs dans $R^p$ et indexé par des paramètres inconnus $\theta$. Notons $\psi$ les paramètres inconnus qui indexent la distribution d'indicateurs R. Le modèle complet (pour les données et le mécanisme des données manquantes) spécifie une distribution $f(Y ; \theta)$ et une distribution $f(R ; Y, \psi)$ appelée distribution des données manquantes où Y est connu. En posant $Y = (Y_o, Y_m)$ où $Y_o$ représente les valeurs observées et où $Y_m$ représente les valeurs manquantes, Rubin (1976) définit les données manquantes comme DMCH si $f(R ; Y_o, Y_m, \psi) = f(R ; \psi)$, c'est-à-dire si l'absence de données ne dépend ni des valeurs observées ni des valeurs manquantes de Y. De plus cet auteur définit une condition plus faible concernant le mécanisme des données manquantes et reconnaît ces données comme étant de type DMH si $f(R ; Y_o, Y_m, \psi) = f(R ; Y_o, \psi)$, c'est-à-dire si l'absence de données ne dépend pas des valeurs manquantes $Y_m$ de Y mais peut dépendre des valeurs observées. Rubin montre également que, si les données manquantes sont de

type DMH et si $\theta$ et $\psi$ sont distincts, les inférences sur $\theta$ peuvent être basées sur $f(Y_o ; \theta)$ qui ignorent le mécanisme des données manquantes. La vraisemblance est alors obtenue comme suit :

$$f(Y_o ; \theta) = \int f(Y_o, Y_m ; \theta)\, dY_m$$

Les performances des méthodes traitant des données manquantes dépendent fortement des mécanismes qui régissent l'apparition de ces données. Les méthodes dites rapides et fréquemment utilisées dans les logiciels sont seulement appropriées aux présomptions fortes, c'est-à-dire lorsque les données sont de type DMCH (Little et Rubin, 1987). Dans ce cas, Dixon (1983) estime le vecteur moyenne et la matrice de covariance et Little (1988) propose pour ce type de données manquantes, un test statistique pour des données multivariables. En général, ces méthodes ne sont donc pas appropriées lorsque les données ne sont pas de type DMCH mais sont de type DMH. Au contraire, les méthodes basées sur la vraisemblance sont appropriées aux présomptions plus faibles, c'est-à-dire lorsque les données manquantes sont de type DMH. Cet avantage est d'une réelle importance dans de nombreux cas pratiques.

Plusieurs stratégies d'approche sont utilisées pour le traitement des données manquantes. Le plus souvent, les individus sur lesquels certaines variables n'ont pas été relevées, sont supprimés, risquant ainsi de réduire considérablement la taille de l'échantillon et d'introduire certains biais si les observations ne sont pas de type DMH. Une autre méthode, fréquemment employée dans le cas quantitatif, consiste à remplacer toute observation manquante par la valeur moyenne des observations disponibles sur la variable concernée. Là encore, l'échantillon sur lequel se pratique l'analyse peut être biaisé.

Parmi les nombreux articles consacrés à ce sujet, nous pouvons distinguer deux sortes d'approche :

La première approche consiste à reconstituer les données manquantes sans se préoccuper de l'usage qui en sera fait ultérieurement. Nous pouvons ainsi citer les méthodes utilisant une formule de reconstitution après une analyse factorielle des correspondances (Nora-Chouteau 1974) ou après une première analyse en composantes principales (Dear 1959, Gleason et Stealin 1975). Pour reconstituer une valeur manquante sur un objet, Buck (1960) et Frane (1976) utilisent, quant à eux, des régressions de la variable manquante sur une ou plusieurs variables disponibles.

La deuxième approche consiste à estimer les paramètres, ou plus généralement les coefficients du modèle utilisé, par une technique d'analyse de données ou par la méthode du maximum de vraisemblance.
A partir des données incomplètes dont ils disposaient, Afifi-Elashoff (1966), Anderson (1975), Dear (1959), Frane (1975) et Haitovsky (1968) ont estimé les paramètres de la régression, Christofferson ( 1974) et Lacourly (1974) ont estimé, quant à eux, les composantes et les axes principaux de l'analyse en composantes principales.
Dans le cadre de la méthode du maximum de vraisemblance, de nombreuses techniques se sont développées pour estimer les paramètres des densités. Anderson (1957) développe l'une d'elles lorsque l'échantillon est extrait d'une population normale et lorsqu'une seule variable comporte des données manquantes. Marini, Olsen et Rubin (1980) donnent une illustration numérique du maximum de vraisemblance dans le cas particulier où les cases correspondant aux données observées de chaque variable sont emboîtées. Citons également Beale et Little (1975) qui ont estimé les paramètres d'une loi normale multidimensionnelle par la méthode du point fixe et qui ont organisé un algorithme de reconstitution. Ces auteurs se sont basés essentiellement sur le principe de l'Information Manquante (IM) développé par Orchard et Woodbury (1972) et ont organisé une procédure cyclique jusqu'à la convergence à partir de valeurs initiales du vecteur de moyenne et de la matrice de covariance. Un algorithme, construit en deux étapes, Estimation et Maximisation et dit de type EM, accomplit la même tâche.
Le terme EM est introduit pour la première fois par Dempster, Laird et Rubin (1977) bien que ce processus ait déjà été utilisé auparavant par McKendrick (1926). Ce type d'algorithme consiste à : (1) remplacer la valeur manquante par son estimation, (2) estimer les paramètres, (3) réestimer les valeurs manquantes en connaissant les nouveaux paramètres, (4) réestimer les paramètres jusqu'à la convergence. L'étude du comportement théorique de cet algorithme est difficile. Les résultats les plus significatifs sont ceux de Wu (1983) pour l'aspect algorithme d'optimisation, et ceux de Redner et Walker (1984) pour l'aspect "maximum de vraisemblance". Un premier exemple intéressant d'application de cet algorithme est présenté par Grundy en 1952. L'article fondamental de Blight (1970) traite des familles exponentielles en général, et reconnaît explicitement l'interprétation à deux pas de chaque itération de EM. Efron et Morris (1967) proposent l'algorithme EM pour les données uniquement censurées et Turnbull (1974, 1976) étend l'approche d'Efron aux données arbitrairement censurées ou tronquées. Pour ce type de données manquantes, nous pouvons également citer les travaux de (Hartley 1958, Irwin 1959 et 1963). Après estimation des paramètres des modèles log-linéaires en présence de données incomplètes par l'algorithme EM, Fuchs

(1982) propose de rendre cet algorithme plus efficace en considérant des tests d'ajustement.

Cependant, l'algorithme EM ne sert pas uniquement lorsque les données de l'échantillon sont incomplètes. En effet, il est aussi utilisé dans le cadre de la reconnaissance de mélanges à laquelle nous nous intéressons également dans ce travail. Les problèmes de mélanges finis peuvent être considérés comme des problèmes d'information manquante car le problème de l'appartenance à un composant est une information manquante. Nous pouvons citer les travaux sur l'algorithme EM simple ou auquel est ajouté une étape Stochastique (algorithme SEM) de Celeux et Diebolt (1985, 1986, 1988), et de Redner et Walker (1984). En utilisant le principe IM, Orchard et Woodbury (1972) estiment également les paramètres du modèle de mélanges finis.

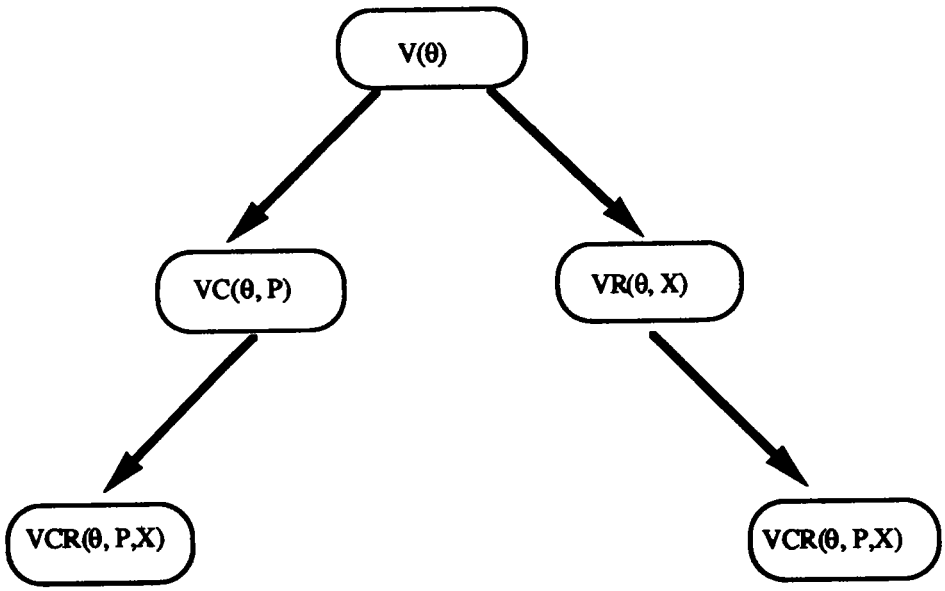
Comme nous nous intéressons plus particulièrement au problème de la classification automatique en présence des données manquantes, nous sommes confrontés à deux types de données manquantes qui sont dues d'une part à la structure du mélange ( appartenance aux classes ) et d'autre part aux valeurs non observées de l'échantillon. La prise en compte des données manquantes en classification est peu envisagée. Citons Fèvre (1980) qui a développé des méthodes de classification dans le cas continu. Quelques méthodes de classification adaptées aux données manquantes sont également proposées par Ok-Sakun (1975), Diday (1975) et Hartigan (1975). Dans le logiciel SICLA (Système Interactif de Classification Automatique), il n'existe pas de commandes permettant de classifier en présence de données manquantes. Par contre, Celeux (1988) propose dans les cas quantitatifs et qualitatifs, plusieurs stratégies d'attributions de valeurs aux données manquantes en tenant compte des différents types de ces données. Puis, il étend ses résultats en utilisant l'algorithme SEM.

Pour aborder le problème de la classification, nous nous basons essentiellement sur la notion de modèle probabiliste qui est souvent située derrière une méthode de classification. Ceci permet d'interpréter certaines méthodes existantes et d'en développer d'autres. Nous nous limitons dans ce travail aux cas des données binaires et qualitatives nominales.

Avant de citer brièvement les contenus des chapitres de notre travail, nous tenons à donner une vue générale sur la démarche entreprise qui est basée essentiellement sur le processus de l'algorithme EM. Pour cela, nous notons **P** une partition de l'ensemble à classifier, **X** l'ensemble des données manquantes et $\theta$ les paramètres qui indexent la

vraisemblance V. La maximisation de cette vraisemblance V(θ) est remplacée par la maximisation de la "moyenne" des vraisemblances classifiantes VC(θ, P) pour tous les P possibles. Rappelons qu'en classification, l'objectif est la maximisation de VC(θ, P). Selon le même mode de raisonnement que précédemment, cette maximisation peut être remplacée par la maximisation de la "moyenne" des vraisemblances classifiantes "reconstituantes" VCR(θ, P, X) pour tous les X possibles. Cette maximisation s'appuie sur la maximisation d'une espérance conditionnelle.

Nous pouvons aborder cette démarche d'une manière légèrement différente en cherchant à maximiser la vraisemblance "reconstituante" VR(θ, X) et en utilisant également la maximisation de la vraisemblance VCR(θ, P, X) pour tous les X possibles. Ainsi, bien que l'objectif de notre travail soit la classification automatique, nous nous intéressons également au problème de la reconstitution des données manquantes. De cette façon, nous nous situons dans les deux grandes approches décrites précédemment. Dans le graphe ci-dessous, nous résumons les différentes étapes envisagées dans cette démarche.




Le premier chapitre comporte des rappels sur les deux approches estimation et classification utilisées dans le cadre de l'estimation des paramètres du mélange. Dans le deuxième chapitre, nous élaborons une méthode de classification en utilisant les trois variantes du modèle associé aux données binaires (Govaert 1989) et en faisant des hypothèses sur les données manquantes. Dans le troisième chapitre, nous

une méthode de classification avec reconstitution des valeurs non observées. Dans le quatrième chapitre, nous élargissons l'utilisation de l'algorithme EM à notre situation dans le cadre de l'estimation des paramètres du modèle de mélanges puis nous utilisons les estimations des paramètres en vue de rechercher une partition. Dans le cinquième chapitre, nous montrons comment adapter la méthode des nuées dynamiques dans le cas quantitatif (MNDQAN) (Celeux *et al.* 1989) à notre problème. Les trois derniers chapitres sont consacrés aux données qualitatives. Dans le sixième et le septième chapitres, nous faisons une extension des différentes méthodes développées dans le cas binaire et dans le dernier chapitre nous proposons une variante de la méthode MNDDIK (**M**éthode des **N**uées **D**ynamiques sur un tableau **DI**sjonctif complet et utilisant la distance du **K**hi2) (Marchetti 1989), en utilisant une variante de la distance du Khi2. Les différentes méthodes proposées dans cette étude sont présentées ci dessous.

**MNDM**

méthode des nuées dynamiques en présence de données manquantes qui sont supposées ne pas suivre le modèle (Chapitre 2).

**MNDMIN**

une variante de la méthode MNDM lorsque les données manquantes suivent le modèle (Chapitre 3).

**MNDMRE**

méthode des nuées dynamiques avec reconstitution des données manquantes (Chapitre 3).

**EMDM**

une variante de l'algorithme EM lorsque les données manquantes sont dues d'une part à la structure de mélange et d'autre part aux valeurs non observées de l'échantillon (Chapitre 4).

**MNDKIDM**

une variante de la méthode MNDDIK utilisant une variante de la distance du Khi2, adaptée aux données manquantes (Chapitre 8).

Les programmes qui correspondent à ces méthodes ont été écrits et intégrés au logiciel d'analyse des données SICLA développé par l'équipe de "classification automatique et reconnaissance des formes" de l'INRIA.

# DONNEES BINAIRES

# CHAPITRE I

# CLASSIFICATION ET MODELES PROBABILISTES

## INTRODUCTION

Les modèles probabilistes étant notre principal outil pour étudier et proposer des solutions au problème de la classification en présence de données manquantes, nous commençons donc, dans ce premier chapitre, par rappeler comment la classification peut être vue comme une solution à un problème d'estimation de paramètres d'un modèle de mélanges.

De nombreuses méthodes de classification reposent essentiellement sur la définition d'une distance ou plus généralement sur celle d'une mesure de dissimilarité et d'un critère associé, sans faire référence explicitement à des modèles probabilistes. En réalité, comme Celeux le propose (1988), et en particulier dans le cas du modèle gaussien, il est souvent possible de montrer qu'il existe un modèle sous-jacent. Celui-ci permet alors de donner une interprétation du critère et de justifier de son choix. Ainsi, le critère d'inertie interclasse est associé aux mélanges gaussiens. Dans le cas binaire, Govaert (1988) a montré que, lorsque les données sont binaires, la classification utilisant la distance $L_1$ correspond à l'hypothèse d'une population issue d'un mélange de distributions de Bernoulli avec le même paramètre pour toutes les classes et pour toutes les variables.

Nous rappelons ici cette approche, en particulier dans le cas des données binaires. Au delà des liens existant entre les critères à optimiser dans le cas du mélange (vraisemblance) et de la classification (vraisemblance classifiante), il est possible d'associer à l'algorithme EM ( Estimation, Maximisation ) un algorithme CEM (Classification, Estimation, Maximisation) pour estimer les paramètres du mélange (Celeux et Govaert 1991).

Dans le premier paragraphe, nous précisons rapidement les modèles de mélanges, puis nous abordons le problème du cas binaire pour lequel plusieurs variantes peuvent être proposées (Govaert 1988). Dans le second paragraphe, nous rappelons comment

l'algorithme **EM** peut résoudre le problème de l'estimation des paramètres du modèle de mélanges, et nous étudions en particulier ce que devient cet algorithme pour les différents modèles de Bernoulli associés aux données binaires. Enfin, dans le troisième paragraphe, nous rappelons comment la classification peut être vue comme une solution au problème d'estimation des paramètres d'un mélange, puis comment nous pouvons définir un algorithme de classification CEM à partir de l'algorithme EM utilisé pour la résolution du problème initial. Ainsi, l'algorithme CEM correspondant au modèle de mélanges gaussiens avec des variances égales et des proportions constantes est simplement celui des centres mobiles. De même, sous la contrainte de proportions égales, l'algorithme CEM correspondant au modèle de mélanges de Bernoulli dont nous avons parlé, est la Méthode des Nuées Dynamiques dans le cas des données **BIN**aires ( **MNDBIN** ) (Marchetti 1989).

## 1. MODELE DE MELANGES

Soit $\Omega$ un ensemble de n individus $\{x_1, ..., x_n\}$ mesurés par p variables binaires $\{x^1, ..., x^p\}$, $I = \{1, ..., n\}$ et $J = \{1, ..., p\}$ correspondent respectivement aux indices des n individus et p variables. Nous notons **X** la matrice de données (nxp).

### 1.1 MODELE GENERAL

Le tableau de données initial de dimension (nxp) et noté **X**, est considéré comme un échantillon $\Omega$ de taille n d'une variable aléatoire à valeurs dans $\{0, 1\}^p$, et dont la loi de probabilité admet la distribution de probabilité f.

Pour tout $x_i \in \Omega$, nous avons :

$$f(x_i) = \sum_{k=1}^{K} p_k \, f(x_i ; a_k)$$

$$\text{avec} \quad \forall k = 1, K \quad p_k \in ]0, 1[ \quad \text{et} \quad \sum_{k=1}^{K} p_k = 1$$

où $f(. ; a_k)$ est une distribution de probabilité sur $\{0, 1\}^p$ appartenant à une famille de distributions de probabilités et $p_k$ est la probabilité qu'un point de l'échantillon suive la loi $f(. ; a_k)$, c'est-à-dire le poids spécifique du composant dans la population générale. Nous appelons ces $p_k$ les proportions du mélange, K étant le nombre de composants de ce mélange.

**Remarque :**

*Le paramètre $a_k$ peut définir aussi bien le centre de groupement des observations correspondantes ( auquel cas il s'interprète comme un paramètre de localisation) que le degré de leur dispersion aléatoire (il est alors interprété comme un paramètre d'échelle). Cette définition permet souvent de simplifier le modèle de mélanges en imposant des contraintes sur le paramètre de dispersion (constant, forme précise, etc...). Ainsi, dans l'algorithme des centres mobiles, sous la contrainte des proportions égales, nous supposons de plus que les lois associées à chaque classe ont une matrice de variance égale à l'identité.*

## 1.2 LOI DE BERNOULLI

pour chaque composant du mélange, nous supposons que les p variables sont indépendantes et que chacune d'elles suit une des deux lois de Bernoulli suivantes (Govaert 1988) :

$$\begin{cases} 1 \text{ avec la probabilité } 1-\varepsilon \text{ et } 0 \text{ avec la probabilité } \varepsilon \\ 1 \text{ avec la probabilité } \varepsilon \text{ et } 0 \text{ avec la probabilité } 1-\varepsilon \end{cases}$$

où $\varepsilon \in ]0, \frac{1}{2}[$ ; c'est-à-dire les lois de Bernoulli respectivement de paramètre $(1-\varepsilon)$ et de paramètre $\varepsilon$.

Nous pouvons alors écrire : $$f(x_i ; a_k) = \prod_{j=1}^{p} \varepsilon^{|x_i^j - a_k^j|} (1-\varepsilon)^{1-|x_i^j - a_k^j|} \qquad (1.2.1)$$

où $a_k = (a_k^1, ..., a_k^p)$ et où les $a_k^j$ indiquent la distribution retenue :

$$\begin{cases} a_k^j = 1 \text{ pour la première distribution} \\ a_k^j = 0 \text{ pour la seconde distribution} \end{cases}$$

Nous pouvons généraliser en remplaçant la valeur réelle de $\varepsilon$ par un vecteur de p valeurs $\varepsilon^j$ dépendant de chaque variable ou des valeurs $\varepsilon_k^j$ dépendant à la fois des classes et des variables. L'expression (1.2.1) s'écrit alors respectivement :

$$f(x_i ; a_k) = \prod_{j=1}^{p} \varepsilon^{j\,|x_i^j - a_k^j|} (1-\varepsilon^j)^{1-|x_i^j - a_k^j|} \qquad (1.2.2)$$

et

$$f(x_i ; a_k) = \prod_{j=1}^{p} \varepsilon_k^{j\,|x_i^j - a_k^j|} \, (1-\varepsilon_k^j)^{1-|x_i^j - a_k^j|} \qquad (1.2.3)$$

Govaert (1988) et, Celeux et Govaert (1989) ont montré qu'il existe un lien étroit entre le modèle associé aux données binaires dans le cas le plus général (les paramètres du modèle dépendant à la fois des classes et des variables) et le modèle des classes latentes que nous rappelons ci-dessous.

L'hypothèse du modèle des classes latentes (Goodman 1974, Everitt 1981) est la suivante : il existe une variable qualitative "cachée" à K modalités telle que conditionnellement à la connaissance de l'une de ces modalités, les p variables soient mutuellement indépendantes.

Les paramètres de ce modèle sont les fréquences relatives ($p_k$, k=1, K) des K modalités de la variable cachée ou latente et les probabilités $\alpha_k^j$ (probabilité que l'individu $x_i$ présente 1 pour la variable $x^j$ c'est-à-dire que $x_i^j = 1$ sachant que cet individu présente la modalité k de la variable latente).
Ce modèle revient à supposer (Everitt 1981) que les n vecteurs binaires à p coordonnées décrivant les individus sont un échantillon du mélange de densités.

$$\text{Pour tout } x_i \quad f(x_i) = \sum_{k=1}^{K} p_k f(x_i ; \alpha_k) \quad \text{avec} \quad f(x_i, \alpha_k) = \prod_{j \in J} (\alpha_{k,j})^{x_i^j} (1-\alpha_{k,j})^{1-x_i^j}$$

où $\alpha_k = (\alpha_{k,j} , j \in J)$

et $f(x_i ; \alpha_k)$ est la densité d'une loi binomiale multivariée de paramètre $\alpha_k$.

Pour mettre en évidence le lien dont nous avons parlé, nous reprenons l'expression de (1.2.3) et celle de $f(x_i, \alpha_k)$ et nous constatons que pour retrouver le modèle des classes latentes, il suffit de poser :

-si $\alpha_{k,j} \in [0, 1/2[$ $\quad \varepsilon_k^j = \alpha_{k,j}$ et $a_k^j = 0$

-si $\alpha_{k,j} \in ]1/2, 1]$ $\quad \varepsilon_k^j = 1 - \alpha_{k,j}$ et $a_k^j = 1$.

Le problème posé est l'estimation des paramètres inconnus $\{p_k, a_k / k = 1, K\}$ au vu de l'échantillon. Nous étudions dans le paragraphe suivant l'utilisation de l'algorithme EM pour résoudre ce problème.

## 2. APPROCHE ESTIMATION

Cette approche, bien que très ancienne, présente l'intérêt d'être directe. Les principales techniques d'estimation utilisées pour estimer les paramètres d'un modèle de mélanges sont celles des moments (Pearson 1894) et celles du maximum de vraisemblance (Day 1969, Wolfe 1970). Ces dernières consistent à résoudre itérativement les équations de vraisemblance et, à des variantes près, les algorithmes les plus efficaces sont de type **EM.**

### 2.1 ALGORITHME EM

Soit Y un échantillon de taille n d'une variable réelle à valeurs dans $\{0, 1\}^p$. Supposons que nous ayons un modèle pour des données complètes Y avec la densité $f(Y/\theta)$ où $\theta$ est un paramètre inconnu. Nous notons $Y = (Y_o, Y_m)$ où $Y_o$ représente les données observées et $Y_m$ représente les données manquantes. Nous supposons que les données manquantes sont de type DMH (Données Manquantes au Hasard) et notre objectif est d'estimer la vraisemblance :

$$f(Y_o ; \theta) = \int f(Y_o, Y_m ; \theta)\, dY_m \qquad (2.1.1)$$

La distribution des données complètes Y peut s'écrire :

$$f(Y; \theta) = f(Y_o, Y_m ; \theta) = f(Y_o ; \theta) f(Y_m / Y_o, \theta)$$

où

$\begin{cases} f(Y_o ; \theta) \text{ est la densité des données observées} \\ f(Y_m / Y_o, \theta) \text{ est la densité de la loi conditionnelle de } Y_m \text{ sachant } Y_o \text{ et } \theta \end{cases}$

La décomposition du Log-vraisemblance correspond à :

$$\text{Log } f(Y ; \theta) = \text{Log } f(Y_o, Y_m ; \theta) = \text{Log } f(Y_o ; \theta) + \text{Log } f(Y_m / Y_o, \theta)$$

L'estimation de $\hat{\theta}$ qui maximise $\text{Log } f(Y_o ; \theta)$ pour $Y_o$ fixe est difficile à obtenir par des calculs directs. En revanche, si les données manquantes sont reconstituées, il est bien plus facile de trouver $\hat{\theta}$ qui maximise $\text{Log } f(Y_o, Y_m / \theta)$.

Nous avons : $\text{Log } f(Y_o ; \theta) = \text{Log } f(Y ; \theta) - \text{Log } f(Y_m / Y_o, \theta)$

Fixons maintenant une valeur de $\theta$ (Soit $\theta^{(t)}$ cette valeur) ce qui nous permettra de définir complètement la densité Log $f(Y_m/Y_o, \theta)$. L'espérance conditionnelle s'écrit :

$$\text{Log } f(Y_o ; \theta) = Q(\theta, \theta^{(t)}) - H(\theta, \theta^{(t)})$$

où $$Q(\theta, \theta^{(t)}) = E[\text{Log } f(Y_o , Y_m ; \theta) / Y_o, \theta^{(t)})]$$

et $$H(\theta, \theta^{(t)}) = E[\text{Log } f(Y_m / Y_o , \theta) / Y_o, \theta^{(t)})]$$

*Notons que $H(\theta, \theta^{(t)}) \leq H(\theta^{(t)}, \theta^{(t)})$ par l'inégalité de Jensen (voir Rao 1972, p 47).*

L'algorithme EM remplace la maximisation de la vraisemblance des données complètes, inobservées, par celle de son espérance conditionnelle aux observations. Nous décrivons les deux étapes qui constituent cet algorithme :

***Etape d'estimation :***

Recherche de l'espérance conditionnelle $Q(\theta, \theta^{(t)})$ sachant les données observées et l'estimation courante du paramètre $\theta$.

***Etape maximisation :***

Recherche des estimations $\theta^{(t+1)}$ maximisant $Q(\theta, \theta^{(t)})$.
*Notons que pour tout $\theta$, nous avons : $Q(\theta^{(t+1)}, \theta^{(t)}) \geq Q(\theta, \theta^{(t)})$.*

Nous allons étudier maintenant l'application de cet algorithme dans le cadre de l'estimation des paramètres d'un modèle de mélanges. Dans ce cas, les données considérées manquantes sont les composants d'où sont issus les individus $x_i$.

## 2.2 MODELE DE MELANGES

L'algorithme EM est utilisé pour estimer le paramètre $\theta$ qui, dans ce cas, est $(p_k, a_k)$, en considérant $Y = (Y_o, Y_m)$ où $Y_o$ représente les données observées c'est-à-dire toutes les données de l'échantillon et où $Y_m = (z_1, ..., z_n)$ avec $z_i$ appartenant à $\{1, ..., K\}$ et indiquant le composant du mélange duquel $x_i$ est issu. En notant $p(z_i) = p_k$ si $z_i = k$, la vraisemblance s'écrit :

$$f(Y ; \theta) = \prod_{i \in I} p(z_i) f(x_i ; a(z_i)) \qquad \text{où} \qquad a(z_i) = a_k \text{ si } z_i = k.$$

avec $a(z_i) = ( a_j(z_i) ; j \in J)$ et $a_k = (a_{k,j} , j \in J)$

Dans tous les chapitres et pour tout x appartenant à [0, 1], les notations $x_{k,j}$ et $x_k^j$ sont équivalentes. Par commodité, nous opterons pour l'une ou l'autre de ces notations suivant la situation rencontrée.

La vraisemblance des données complètes s'écrit :

$$f(Y ; \theta) = \prod_{i \in I} p(z_i) f(x_i, a(z_i))$$

Notons qu'à partir de (2.1.1) nous avons :

$$f(Y_o ; \theta) = \int_Z f(Y_o, Z ; \theta) \, d Z$$

Nous pouvons alors écrire :

$$f(Y_o ; \theta) = \prod_{i \in I} \sum_{k=1}^{K} p_k f(x_i, a_k)$$

**Remarque :**

*L'algorithme EM présente les caractéristiques suivantes :*
*-Il fonctionne pour un grand nombre de composants et dans le cas multidimensionnel.*
*-Il fournit en général de bons résultats si le nombre de composants est connu.*
*-Malheureusement, il converge extrêmement lentement. Cette lenteur peut rendre son utilisation redhibitoire. C'est en particulier le cas lorsque la solution initiale est éloignée de la solution limite accessible*

Nous développons dans le paragraphe suivant l'algorithme EM pour les différents modèles de mélanges associés aux données binaires.

## 2.3 MODELES DE BERNOULLI

Il nous reste à caractériser les deux étapes de cet algorithme. Dans l'étape d'estimation, nous donnons la forme générale de l'espérance conditionnelle. Par contre, dans l'étape de maximisation, nous distinguons les cas des trois variantes du modèle de Bernoulli.

***Etape estimation :***

Dans cette étape, nous calculons l'espérance conditionnelle $Q(\theta, \theta^{(t)})$.

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= E[\text{Log } f(Y_o, Y_m ; \theta) / Y_o, \theta^{(t)}) \\ &= E[\text{Log } \prod_{i \in I} p(z_i) f(x_i ; a(z_i)) / Y_o, p_k^{(t)}, a_k^{(t)}] \\ &= \sum_{i \in I} E[\text{Log } p(z_i) f(x_i ; a(z_i)) / Y_o, p_k^{(t)}, a_k^{(t)}] \\ &= \sum_{i \in I} E[\text{Log } p(z_i) / Y_o, p_k^{(t)}, a_k^{(t)}] + \sum_{i \in I} E[\text{Log } f(x_i ; a(z_i)) / Y_o, p_k^{(t)}, a_k^{(t)}] \end{aligned}$$

En notant $s_k^{(t)}(x_i)$ la probabilité conditionnelle qu'un individu $x_i$ soit issu du composant k, nous en déduisons l'écriture de $Q(\theta, \theta^{(t)})$ :

$$Q(\theta, \theta^{(t)}) = \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log } p_k + \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log } f(x_i ; a_k) \qquad (2.3.1)$$

$$\text{avec } s_k^{(t)}(x_i) = \frac{p_k^{(t)} f(x_i ; a_k^{(t)})}{\sum_{k=1}^{K} p_k^{(t)} f(x_i ; a_k^{(t)})}$$

***Etape de maximisation :***

Cette étape consiste à trouver $p_k^{(t+1)}$, $a_k^{(t+1)}$ maximisant $Q(\theta^{(t+1)}, \theta^{(t)})$.

***Calcul des $p_k^{(t+1)}$***

Le lagrangien (Lag) de ce problème s'écrit :

$$\text{Lag} = Q(\theta^{(t+1)}, \theta^{(t)}) - \lambda(\sum_{k=1}^{K} p_k^{(t+1)} - 1)$$

$$\frac{\partial \text{ Lag}}{\partial p_k^{(t+1)}} = \frac{\sum_{i \in I} s_k^{(t)}(x_i)}{p_k^{(t+1)}} - \lambda$$

$$\frac{\partial\, Lag}{\partial p_k^{(t+1)}} = 0 \quad \Leftrightarrow \quad \lambda = \frac{\sum_{i\in I} s_k^{(t)}(x_i)}{p_k^{(t+1)}}$$

$$\Leftrightarrow \quad p_k^{(t+1)} = \frac{\sum_{i\in I} s_k^{(t)}(x_i)}{\lambda}$$

De $\sum_{k=1}^{K} p_k^{(t+1)} = 1$ et $\sum_{k=1}^{K} \sum_{i\in I} s_k^{(t)}(x_i) = n$ nous en tirons $\lambda = n$

$$\text{d'où} \quad p_k^{(t+1)} = \frac{\sum_{i\in I} s_k^{(t)}(x_i)}{n}$$

***Calcul des*** $a_k^{(t+1)}$

Les vecteurs $a_k^{(t+1)}$ qui maximisent $Q(\theta^{(t+1)}, \theta^{(t)})$ sont ceux qui maximisent :

$$\sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log } f(x_i ; a_k^{(t+1)}) \qquad (2.3.2)$$

Nous allons considérer en détail les trois variantes du modèle de mélanges de Bernoulli afin de connaître l'estimation des paramètres $a_k$ de chaque modèle.

### 2.3.1 ε paramètre fixe

A partir de (1.2.1), nous pouvons écrire :

$$\text{Log } f(x_i ; a_k^{(t+1)}) = \text{Log} \left( \prod_{j\in J} \varepsilon^{(t+1)^{|x_i^j - a_{k,j}^{(t+1)}|}} . (1-\varepsilon^{(t+1)})^{1-|x_i^j - a_{k,j}^{(t+1)}|} \right)$$

$$= \text{Log} \frac{\varepsilon^{(t+1)}}{1-\varepsilon^{(t+1)}} \left( \sum_{j\in J} |x_i^j - a_{k,j}^{(t+1)}| \right) + p \text{ Log}(1-\varepsilon^{(t+1)})$$

Le terme (2.3.2) à maximiser s'écrit :

$$\text{Log} \frac{\varepsilon^{(t+1)}}{1-\varepsilon^{(t+1)}} \sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \sum_{j\in J} |x_i^j - a_{k,j}^{(t+1)}| + p \sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log}(1-\varepsilon^{(t+1)})$$

(2.3.1.1)

Comme $\sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) = n$, l'expression (2.3.1.1) peut également s'écrire :

$$\text{Log} \frac{\varepsilon^{(t+1)}}{1-\varepsilon^{(t+1)}} \sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \sum_{j\in J} |x_i^j - a_{k,j}^{(t+1)}| + np \text{ Log}(1-\varepsilon^{(t+1)}) \qquad (2.3.1.2)$$

Puisque $\text{Log}(\varepsilon^{(t+1)}/1-\varepsilon^{(t+1)})$ est négatif ($\varepsilon^{(t+1)} \in ]0, 1/2[$ ), maximiser (2.3.1.2) revient à minimiser :

$$\sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \sum_{j\in J} | x_i^j - a_{k,j}^{(t+1)}|$$

ou encore à minimiser chacun des termes suivants :

$$L_{k,j} = \sum_{i\in I} s_k^{(t)}(x_i) | x_i^j - a_{k,j}^{(t+1)}|$$

Les solutions de ce problème sont définies par :

$\forall\, k = 1,\ K, \forall\, j \in J$

$$a_{k,j}^{(t+1)} = \text{médiane binaire de l'ensemble } \{(x_i^j\ ,\ s_k^{(t)}(x_i)),\ i \in I\}.$$

Pour une variable j et un k fixés, nous pouvons considérer les $\{s_k^{(t)}(x_i)), i \in I\}$ comme des pondérations de chaque valeur 0 ou 1. La valeur $a_{k,j}^{(t+1)}$ est alors simplement la valeur majoritaire de ces valeurs 0 et 1 pondérées.

Les composantes $a_{k,j}^{(t+1)}$ étant définies, si nous notons

$$e^{(t+1)} = \sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \sum_{j\in J} | x_i^j - a_{k,j}^{(t+1)}|\ ,$$

nous pouvons alors écrire à partir de (2.3.1.2) :

$$e^{(t+1)}\ \text{Log}\frac{\varepsilon^{(t+1)}}{1-\varepsilon^{(t+1)}} + np\ \text{Log}(1-\varepsilon^{(t+1)}) = D(\varepsilon^{(t+1)}) \qquad (2.3.1.3)$$

La valeur $\varepsilon^{(t+1)}$ qui maximise (2.3.1.3) est alors définie par :

$$\frac{\partial D}{\partial \varepsilon^{(t+1)}} = 0 \quad \Rightarrow \quad \varepsilon^{(t+1)} = \frac{e^{(t+1)}}{np}$$

### 2.3.2 ε paramètre dépendant de chaque variable

A partir de (1.2.2), nous pouvons écrire :

$$\text{Log } f(x_i ; a_k^{(t+1)}) = \text{Log} \left( \prod_{j \in J} \varepsilon_j^{(t+1)\,|x_i^j - a_{k,j}^{(t+1)}|} \cdot (1 - \varepsilon_j^{(t+1)})^{1-|x_i^j - a_{k,j}^{(t+1)}|} \right)$$

$$= \sum_{j \in J} \text{Log} \frac{\varepsilon_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}} \, |x_i^j - a_{k,j}^{(t+1)}| + \sum_{j \in J} \text{Log}(1-\varepsilon_j^{(t+1)})$$

Le terme (2.3.2) s'écrit alors :

$$\sum_{i \in I} \sum_{k=1}^{K} \sum_{j \in J} s_k^{(t)}(x_i) \, \text{Log} \frac{\varepsilon_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}} \, |x_i^j - a_{k,j}^{(t+1)}| + \sum_{i \in I} \sum_{k=1}^{K} \sum_{j \in J} s_k^{(t)}(x_i) \, \text{Log}(1-\varepsilon_j^{(t+1)})$$

(2.3.2.1)

Nous pouvons en déduire que les $a_k^{(t+1)}$ qui maximisent $Q(\theta^{(t+1)}, \theta^{(t)})$ sont les vecteurs dont les composantes minimisent chacun des termes correspondants :

$$L_{k,j} = \sum_{i \in I} s_k^{(t)}(x_i) \, |x_i^j - a_{k,j}^{(t+1)}|$$

Nous avons comme précédemment :

$\forall\, k = 1,\ K, \forall\, j \in J \quad a_{k,j}^{(t+1)}$ = médiane binaire de l'ensemble $\{(x_i^j , s_k^{(t)}(x_i)),\ i \in I\}$.

Les composantes $a_{k,j}^{(t+1)}$ étant définies, si nous notons $e_j^{(t+1)} = \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \, |x_i^j - a_{k,j}^{(t+1)}|$ nous pouvons alors écrire à partir de (2.3.2.1) :

$$e_j^{(t+1)} \sum_{j \in J} \text{Log} \frac{\varepsilon_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}} + n \sum_{j \in J} \text{Log}(1-\varepsilon_j^{(t+1)}) = \sum_{j \in J} F(\varepsilon_j^{(t+1)}) \qquad (2.3.2.2)$$

Les valeurs $(\varepsilon_j^{(t+1)}, j \in J)$ qui maximisent (2.3.2.2) sont alors définies par :

$$\frac{\partial F}{\partial \varepsilon_j^{(t+1)}} = 0 \quad \Rightarrow \quad \frac{e_j^{(t+1)}}{\varepsilon_j^{(t+1)}} - \frac{n - e_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}} = 0$$

$$\Rightarrow \qquad \varepsilon_j^{(t+1)} = \frac{e_j^{(t+1)}}{n}$$

Il est facile de vérifier que les $\varepsilon_j^{(t+1)}$ appartiennent à ]0, 1/2[ sauf dans les cas très particuliers et très rares où $e_j^{(t+1)} = 0$ ou $e_j^{(t+1)} = n/2$.

Pour pallier à cet inconvénient, lorsque la valeur $\varepsilon_j^{(t+1)}$ est égale à 0 ou 1/2, il suffit de prendre la valeur du paramètre $\varepsilon_j^{(t)}$ calculée dans l'étape précédente et dans ce cas, nous ne maximisons pas le critère mais nous l'améliorons.

### 2.3.3 ε paramètre dépendant de chaque classe et de chaque variable

A partir de (1.2.3) nous pouvons écrire :

$$\text{Log } f(x_i ; a_k^{(t+1)}) = \text{Log} \Big( \prod_{j \in J} \varepsilon_{k,j}^{(t+1)\,|x_i^j - a_{k,j}^{(t+1)}|} \cdot (1-\varepsilon_{k,j}^{(t+1)})^{1-|x_i^j - a_{k,j}^{(t+1)}|} \Big)$$

$$= \sum_{j \in J} \text{Log} \frac{\varepsilon_{k,j}^{(t+1)}}{1-\varepsilon_{k,j}^{(t+1)}} \, |x_i^j - a_{k,j}^{(t+1)}| + \sum_{j \in J} \text{Log}(1-\varepsilon_{k,j}^{(t+1)})$$

Le terme (2.3.2) s'écrit alors :

$$\sum_{i \in I} \sum_{k=1}^{K} \sum_{j \in J} s_k^{(t)}(x_i) \, \text{Log} \frac{\varepsilon_{k,j}^{(t+1)}}{1-\varepsilon_{k,j}^{(t+1)}} \, |x_i^j - a_{k,j}^{(t+1)}| + \sum_{i \in I} \sum_{k=1}^{K} \sum_{j \in J} s_k^{(t)}(x_i) \, \text{Log}(1-\varepsilon_{k,j}^{(t+1)})$$

(2.3.3.1)

Nous pouvons en déduire que les $a_k^{(t+1)}$ qui maximisent $Q(\theta^{(t+1)}, \theta^{(t)})$ sont les vecteurs dont les composantes minimisent chacun des termes correspondants :

$$L_{k,j} = \sum_{i \in I} s_k^{(t)}(x_i) \, |x_i^j - a_{k,j}^{(t+1)}|$$

$\forall\, k = 1, K, \forall\, j \in J \quad a_{k,j}^{(t+1)}$ = médiane binaire de l'ensemble $\{(x_i^j, s_k^{(t)}(x_i)), i \in I\}$.

Les composantes $a_{k,j}^{(t+1)}$ étant définies, si nous notons $e_{k,j}^{(t+1)} = \sum_{i \in I} s_k^{(t)}(x_i) \, |x_i^j - a_{k,j}^{(t+1)}|$, La quantité à maximiser s'écrit alors :

$$e_{k,j}^{(t+1)} \text{ Log} \frac{\varepsilon_{k,j}^{(t+1)}}{1-\varepsilon_{k,j}^{(t+1)}} + \sum_{i \in I} s_k^{(t)}(x_i) \text{ Log}(1-\varepsilon_{k,j}^{(t+1)}) = G(\varepsilon_{k,j}^{(t+1)}) \qquad (2.3.3.2)$$

Les valeurs $(\varepsilon_{k,j}^{(t+1)}, j \in J, k \in \{1, ..., K\})$ qui maximisent (2.3.3.2) sont définies par:

$$\frac{\partial G}{\partial \varepsilon_{k,j}^{(t+1)}} = 0 \quad \Rightarrow \quad \varepsilon_{k,j}^{(t+1)} = \frac{e_{k,j}^{(t+1)}}{\sum_{i \in I} s_k^{(t)}(x_i)}$$

Il est facile de vérifier que les $\varepsilon_{k,j}^{(t+1)}$ appartiennent à [0, 1/2[ sauf dans le cas très particulier et très rare où $e_{k,j}^{(t+1)} = 0$ ou $e_{k,j}^{(t+1)} = n/2$. La stratégie citée précédemment peut être adoptée pour traiter ces cas.

## 3. APPROCHE CLASSIFICATION

### 3.1 PROBLEME

Dans ce paragraphe, nous rappelons comment l'approche "classification" peut être utilisée pour identifier un mélange de lois de probabilités (Scott et Symons 1971, Schroeder 1976).

Dans cette approche, le problème initial d'estimation est remplacé par le problème suivant :

*Rechercher une partition $P = (P_1, ..., P_K)$, K étant supposé connu, telle que chaque classe $P_k$ soit assimilable à un sous-échantillon qui suit une loi $f(., a_k)$.*

L'espace de représentation **L** d'une classe étant l'espace de définition des paramètres a dont dépendent les densités f(., a), la méthode vise à maximiser le critère de **vraisemblance classifiante** suivant :

$$W(P, a) = \sum_{k=1}^{K} \text{Log } L(P_k, a_k) \qquad (3.1.1)$$

où a est le p-uplet $(a_1, ..., a_K)$ et $L(P_k, a_k)$ est la vraisemblance du sous-échantillon $P_k$ suivant la loi $f(./a_k)$ :

$$L(P_k, a_k) = \prod_{i \in P_k} f(x_i ; a_k).$$

Notons que l'approche classification ne s'attaque pas directement à l'estimation des paramètres du mélange mais qu'elle présente surtout l'intérêt d'être rapide.

### 3.2 ALGORITHME CEM

Pour maximiser le critère précédent, nous pouvons utiliser l'algorithme **CEM** (Celeux et Govaert 1991) construit à partir de **EM** en ajoutant une étape de "classification". Nous décrivons ci-dessous les différentes étapes en utilisant les notations précédentes.

**(1) Estimation** des probabilités conditionnelles $s_k$.

**(2) Classification** : les individus $x_i$ sont rangés dans la classe k qui maximise $s_k(x_i)$. Cette étape supplémentaire revient à remplacer le vecteur :

$$s(x_i) = (s_k(x_i) ; k = 1, ..., K)$$

par le vecteur binaire dont l'unique composante non nulle et égale à 1 correspond au plus grand des $s_k$.

*Ex : $s(x_i)$ = (1/3, 1/4, 5/12) alors le vecteur binaire associé est (0, 0, 1) et l'individu $x_i$ est affecté à la classe 3.*

**(3) Maximisation** : calcul des proportions $p_k$ et des paramètres $a_k$.

Si les proportions $p_k$ sont fixées (égales à 1/K), nous retrouvons un algorithme du type de l'algorithme des nuées dynamiques construit à partir de deux étapes.

- une étape de représentation caractérisée par :
une fonction g définie par $g(P) = g(P_1, ..., P_K) = (a_1, ..., a_K)$ où $a_k$ est l'estimation du maximum de vraisemblance du paramètre de la densité associée au sous-échantillon $P_k$.
- une étape d'affectation caractérisée par :
une fonction h définie par $h(a) = h(a_1, ..., a_K) = (P_1, ..., P_K)$ où $P_k = \{x_i \in \Omega / f(x_i/a_k) \geq f(x_i/a_m)$ avec $k<m$ en cas d'égalité$\}$.

Cet algorithme fait croître le critère à chaque itération et nous avons le résultat de convergence suivant :

**Théorème :**
*Sous l'hypothèse que la famille de densités f(x, a) soit bornée supérieurement pour tout* $x \in R^p$ *et pour tout* $a \in L$, *la suite* $v_n = (L^n, P^n)$ *converge dans* $L_k x P_k$ *en un nombre fini d'itérations et atteint sa limite : la suite réelle* $u_n = W(v_n)$ *converge en croissant vers un maximum local.*

Nous obtenons à la convergence une partition P et une estimation des paramètres $a_k$. Les proportions $p_k$ du mélange supposées constantes dans l'algorithme peuvent être alors estimées par les quantités (Card( $P_k$)/n, k =1, K).

### 3.3 DONNEES BINAIRES : MNDBIN

L'application de l'algorithme précédent à des modèles de mélanges est alors simple (Govaert 1988). Nous decrivons ci-dessous la méthode MNDBIN en distinguant les trois variantes du modèle de Bernoulli.

#### 3.3.1 ε paramètre fixe

Le critère de vraisemblance classifiante s'écrit :

$$W(P, a, \varepsilon)] = \text{Log}\frac{\varepsilon}{1-\varepsilon} \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} |x_i^j - a_k^j| + np\,\text{Log}(1-\varepsilon)$$

***Etape d'affectation** : (recherche des classes)*

Lors de cette étape nous affectons l'individu $x_i$ à la classe $P_k$ qui minimise :

$$\sum_{j \in J} |x_i^j - a_k^j|$$

***Etape de représentation** : ( recherche des* $a_k^j$ *)*

Les $a_k^j$ maximisant le critère sont les valeurs majoritaires de chaque classe pour chaque variable. Le paramètre ε n'intervient pas dans les deux étapes, et son estimation ne nous semble pas nécessaire.

**Remarque :**

*Nous avons noté que dans l'algorithme EM les probabilités conditionnelles $s_k$ dépendaient du paramètre $\varepsilon$ dont l'estimation à chaque itération était par conséquent nécessaire contrairement à l'algorithme MNDBIN. En parallèle, nous rappelons que dans le cas gaussien, lorsque les matrices de variance sont supposées identiques pour toutes les classes, elles ne jouent aucun rôle en classification contrairement à l'approche estimation.*

### 3.3.2 $\varepsilon$ paramètre dépendant de chaque variable

Le critère de vraisemblance classifiante s'écrit :

$$W(P, a, \varepsilon) = - \sum_{k=1}^{K} \sum_{i \in P_k} d_\varepsilon(x_i, a_k) + A$$

où $\quad d_\varepsilon(x_i, a_k) = \sum_{j \in J} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | x_i^j - a_k^j | \quad$ et $\quad A = n \sum_{j \in J} \text{Log}(1-\varepsilon^j).$

***Etape d'affectation** : (recherche des classes)*

Lors de cette étape, nous affectons $x_i$ à la classe $P_k$ qui minimise :

$$d_\varepsilon(x_i, a_k)$$

***Etape de représentation** : ( recherche des $a_k^j$ et $\varepsilon^j$ )*

Les $a_k^j$ qui maximisent le critère sont les valeurs majoritaires de chaque classe pour chaque variable. En posant $e_j = \sum_{k=1}^{K} \sum_{i \in P_k} | x_i^j - a_k^j |$ qui exprime le nombre de fois où la valeur majoritaire n'a pas été prise dans une classe, les valeurs $\varepsilon^j$ qui maximisent le critère sont définies par :

$$\varepsilon^j = \frac{e_j}{n}$$

Les $\varepsilon^j$ appartiennent à ]0, 1/2[ sauf dans le cas très particulier où les $e_j$ correspondant sont nulles. Cette approche est similaire à la méthode des distances adaptatives avec une distance unique (Govaert 1975, Diday et Govaert 1977).

### 3.3.3 ε paramètre dépendant de chaque classe et de chaque variable

Le critère de vraisemblance classifiante s'écrit :

$$W(P, a, \varepsilon) = \sum_{k=1}^{K} \sum_{i \in P_k} \{-d_{\varepsilon_K}(x_i, a_k) + A_k\}$$

où $d_{\varepsilon_K}(x_i, a_k) = \sum_{j \in J} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j |$ et $A_k = \sum_{j \in J} \text{Log}(1-\varepsilon_k^j)$.

***Etape d'affectation*** *: (recherche des classes)*

Lors de cette étape, le terme $A_k$ n'est pas constant. Nous affectons $x_i$ à la classe $P_k$ qui minimise $d_{\varepsilon_K}(x_i, a_k) - A_k$.

***Etape de représentation*** *: ( recherche des $a_k^j$ et $\varepsilon_k^j$ )*

Les $a_k^j$ maximisant le critère sont les valeurs majoritaires de chaque classe pour chaque variable. En posant $e_k^j = \sum_{i \in P_k} | x_i^j - a_k^j |$ qui exprime le nombre de fois où la valeur majoritaire n'a pas été prise dans la classe k et pour la variable $x^j$, les valeurs $\varepsilon_k^j$ qui maximisent le critère sont définies par :

$$\varepsilon_k^j = \frac{e_k^j}{n_k}$$

où $n_k$ est l'effectif de la classe $P_k$.

**Remarques :**

1-*Cette approche revient à imposer aux noyaux des classes la même structure que les données initiales : les données étant binaires, chaque noyau peut être considéré comme un élément de $\{0, 1\}^p$, où p représente le nombre total de variables. Ceci permet la description et l'interprétation des classes obtenues. En effet, chaque classe est ainsi résumée par un vecteur binaire.*

*2- Nous avons noté dans ce chapitre les liens étroits existant entre les algorithmes EM et CEM. De cette façon, l'algorithme des centres mobiles s'avère être un cas particulier de la version classification de l'algorithme EM.*

# CHAPITRE II

# DONNEES MANQUANTES UNE PREMIERE APPROCHE

## INTRODUCTION

Dans ce chapitre, nous nous plaçons dans le cas où le problème de la classification automatique se ramène à un problème d'optimisation d'un critère de classification. Dans notre situation où toutes les données binaires du tableau initial ne sont pas observées, le problème est impossible à résoudre. Néanmoins, nous pouvons noter que l'ensemble des valeurs manquantes appartient à un ensemble fini et par conséquent l'ensemble des valeurs du critère est fini. Ainsi, dans ce chapitre, après avoir fait une hypothèse sur la distribution des données manquantes, nous proposons de remplacer le critère de classification par son espérance. Nous étudions cette approche dans le cadre du critère de vraisemblance classifiante rappelé dans le premier chapitre.

Dans le premier paragraphe, nous présentons cette approche et les hypothèses émises sur les données manquantes. Nous proposons dans le deuxième paragraphe une méthode adaptée à notre situation qui utilise un algorithme itératif pour chercher la partition optimale. Dans le troisième paragraphe, nous analysons et comparons les différentes mesures de dissimilarité qui définissent les divers critères métriques obtenus. En particulier, nous montrons les liens qui existent entre notre travail et celui de Fèvre (1980). Enfin, le dernier paragraphe contient des résultats de classification obtenus à partir de données simulées et réelles ayant subi une destruction au hasard.

## 1. APPROCHE PROPOSEE

### 1.1 DONNEES

Nous reprenons les notations du chapitre précédent. Soit $\Omega$ un ensemble de n individus $\{x_1, ..., x_n\}$ mesurés par p variables binaires $\{x^1, ..., x^p\}$. Nous notons $I = \{1, ..., n\}$ et $J = \{1, ..., p\}$ les ensembles qui désignent les indices correspondant

respectivement aux n individus et p variables. Nous notons X la matrice de données (nxp). Cette matrice comportant des données non observées, nous utilisons les notations suivantes:

Soit $x_i$ un élément de $\Omega$. Nous écrivons $x_i = (x_i^o, x_i^m)$ où

$x_i^o$ représente les valeurs observées.

$x_i^m$ représente les valeurs manquantes.

Soit $O_i$ l'ensemble des indices j pour lesquels les valeurs du vecteur $x_i$ sont observées et $M_i$ l'ensemble des indices j pour lesquels les valeurs du vecteur $x_i$ sont manquantes.

Nous avons évidemment pour tout $x_i$ élément de $\Omega$ : $O_i \cup M_i = J$

Nous notons m le nombre total de données manquantes et $m^j$ le nombre de données manquantes pour la variable $x^j$.

## 1.2 PROBLEME

Nous cherchons une partition $P = (P_1, ..., P_K)$ de $\Omega$ en K classes "homogènes" sachant que certaines variables n'ont pas été relevées sur certains individus.

L'utilisation des critères de classification habituels est impossible. Nous aurions pu alors envisager de définir de nouveaux critères adaptés aux données manquantes. Cependant, nous avons préféré nous appuyer sur les critères existants pour les données complètes. Ainsi, nous pouvons remarquer que le tableau initial peut être reconstitué d'un nombre fini de façons. En effet, il y a $2^m$ façons de reconstituer les données manquantes ; il en découle alors tout un ensemble de tableaux de données possibles $X_q$, noté **H**.

$$\mathbf{H} = \{X_q \ / q \in \{1, ..., 2^m\}\}$$

Pour chacun des tableaux reconstitués, nous pouvons alors rechercher la partition optimale pour le critère de classification. Nous obtenons donc non pas une valeur du critère mais une distribution de valeurs. Il suffit alors de retenir comme critère de classification pour les données manquantes une caractéristique de cette distribution. Par exemple, le minimum, le maximum, la moyenne, etc...

Dans la suite de notre chapitre, nous retenons comme critère, l'espérance de cette distribution. Ceci nécessite une hypothèse sur la distribution des données manquantes.

### 1.3 HYPOTHESES SUR LES DONNEES MANQUANTES ET DESCRIPTION

Tout d'abord, nous supposons que les données manquantes sont de type DMH (Données Manquantes au Hasard) et indépendantes. Chaque valeur non observée ne pouvant appartenir qu'à {0, 1}, nous supposons alors que les données manquantes suivent la distribution de Bernoulli de paramètre $\alpha$ **connu** à priori. Nous avons alors :

$$f(x_i^j = 1) = \alpha \quad \text{et} \quad f(x_i^j = 0) = 1-\alpha \qquad \text{pour } j \in M_i$$

Nous verrons que le paramètre $\alpha$ peut être remplacé par un vecteur dont les composantes dépendent de chaque variable.

Le critère de vraisemblance classifiante **W(P, a)** (voir I.3.1.1) peut être considéré comme une variable aléatoire de **H** dans **R**, et son espérance existe. Ainsi, nous remplaçons le critère de vraisemblance classifiante par son espérance.

Chaque échantillon $X_q$ reconstitué est pondéré par une probabilité $\theta_q = \prod_{a=1}^{m} p_a$ où $p_a$ désigne la probabilité associé à chaque valeur de l'ensemble des données manquantes. Nous avons alors :

$$p_a = f(x_i^j = 1 \; ; \; j \in M_i) = \alpha \quad \text{ou} \quad p_a = f(x_i^j = 0 \; ; \; j \in M_i) = 1-\alpha.$$

Il est alors facile de montrer que $\sum_{q=1}^{2^m} \theta_q = 1$.

Nous allons étudier maintenant le modèle binaire le plus simple et décrire la Méthode des Nuées dynamiques dans le cas des Données Manquantes (MNDM).

## 2. METHODE MNDM

Dans ce paragraphe, nous décrivons la méthode de classification dans le cas binaire en présence de données manquantes. Cette méthode présente plusieurs variantes qui sont dues d'une part au choix du modèle et d'autre part au choix du paramètre $\alpha$. Nous considérons tout d'abord le cas le plus simple c'est-à-dire lorsque le paramètre du modèle de Bernoulli $\varepsilon$ est fixe et lorsque $\alpha$ est un réel.

## 2.1 EXPRESSION DU CRITERE

La proposition suivante donne l'expression du critère dans notre cas.

**Proposition 1 :**

$$E[W(P,X,a,\alpha,\varepsilon)] = Log\frac{\varepsilon}{1-\varepsilon}\sum_{k=1}^{K}\sum_{i\in P_k}\Big(\sum_{j\in O_i}|x_i^j-a_k^j| + \sum_{j\in M_i}|\alpha-a_k^j|\Big) + npLog(1-\varepsilon)$$

**Preuve :**

$$E[W(P, X, a, \alpha, \varepsilon)] = E\Big[\sum_{k=1}^{K} Log\ \{\prod_{i\in P_k} f(x_i ; a_k)\}\Big]$$

$$= E\Big[\sum_{k=1}^{K}\sum_{i\in P_k}(Log\ f(x_i ; a_k))\Big]$$

$$= \sum_{k=1}^{K}\sum_{i\in P_k} E[Log\ f(x_i ; a_k)]$$

$$= \sum_{k=1}^{K}\sum_{i\in P_k} E[Log(f(x_i^o, x_i^m ; a_k))]$$

$$= \sum_{k=1}^{K}\sum_{i\in P_k} E[Log(f(x_i^o ; a_k)\, f(x_i^m ; a_k))]$$

(hypothèse d'indépendance à l'intérieur des classes)

$$= \sum_{k=1}^{K}\sum_{i\in P_k}(Log\ f(x_i^o ; a_k) + E[Log\ f(x_i^m ; a_k)]) \qquad (2.1.1)$$

Explicitons le terme $E[Log\ f(x_i^m ; a_k)]$

$$E[Log\ f(x_i^m ; a_k)] = E[Log\prod_{j\in M_i} f(x_i^j ; a_k^j)]$$

$$= E[\sum_{j\in M_i} Log\ f(x_i^j ; a_k^j)]$$

$$= \sum_{j\in M_i} E[Log\ f(x_i^j ; a_k^j)]$$

Le terme $E[Log\ f(x_i^j ; a_k^j)]$ peut s'écrire :

$$E[Log\ f(x_i^j ; a_k^j)] = f(x_i^j = 1)\ Log\ f(x_i^j = 1 ; a_k^j) + f(x_i^j = 0)\ Log\ f(x_i^j = 0 ; a_k^j)$$

$$\text{Si } a_k^j=1 \quad E[\text{Log } f(x_i^j ; a_k^j)] = (1-\alpha)\,\text{Log}\varepsilon + \alpha\,\text{Log}(1-\varepsilon)$$

$$\text{Si } a_k^j=0 \quad E[\text{Log } f(x_i^j ; a_k^j)] = \alpha\,\text{Log}\varepsilon + (1-\alpha)\,\text{Log}(1-\varepsilon)$$

Nous pouvons écrire : $E[\text{Log } f(x_i^j ; a_k^j)] = |\alpha - a_k^j|\,\text{Log}\varepsilon + (1-|\alpha-a_k^j|)\,\text{Log}(1-\varepsilon)$

ou encore

$$E[\text{Log } f(x_i^j ; a_k^j)] = \text{Log}\frac{\varepsilon}{1-\varepsilon}\,|\alpha-a_k^j| + \text{Log}(1-\varepsilon)$$

Nous en déduisons la forme générale de $E[\text{Log } p(x_i^m ; a_k)]$ :

$$E[\text{Log } f(x_i^m ; a_k)] = \text{Log}\frac{\varepsilon}{1-\varepsilon}\sum_{j\in M_i}|\alpha-a_k^j| + m_i\,\text{Log}(1-\varepsilon)$$

où $m_i$ désigne le nombre de données manquantes pour l'individu $x_i$.

Explicitons maintenant le terme $\text{Log } f(x_i^o ; a_k)$.

A partir de (I.1.2.1), nous pouvons écrire : $f(x_i^o ; a_k) = \prod_{j\in O_i} \varepsilon^{|x_i^j - a_k^j|}\,(1-\varepsilon)^{1-|x_i^j - a_k^j|}$

Nous en déduisons le logarithme de $f(x_i^o ; a_k)$ :

$$\text{Log } f(x_i^o ; a_k) = \text{Log}\left(\prod_{j\in O_i} \varepsilon^{|x_i^j - a_k^j|}\,.\,(1-\varepsilon)^{1-|x_i^j - a_k^j|}\right)$$

$$= \text{Log}\frac{\varepsilon}{1-\varepsilon}\left(\sum_{j\in O_i}|x_i^j - a_k^j|\right) + (p-m_i)\,\text{Log}(1-\varepsilon)$$

D'où :

$$E[W(P,X,a,\alpha,\varepsilon)] = \text{Log}\frac{\varepsilon}{1-\varepsilon}\sum_{k=1}^{K}\sum_{i\in P_k}\left(\sum_{j\in O_i}|x_i^j - a_k^j| + \sum_{j\in M_i}|\alpha-a_k^j|\right) + np\,\text{Log}(1-\varepsilon)$$

#

**Remarques :**

1- *Lorsque toutes les valeurs sont observées,* $\sum_{j\in M_i}|\alpha-a_k^j| = 0$ *et nous retrouvons l'expression du critère de vraisemblance classifiante dans le cas où toutes les variables suivent une loi de Bernoulli de paramètre* $\varepsilon$ *(Govaert 1988).*

2- *Tout se passe comme si les données manquantes étaient reconstituées par la valeur* $\alpha$ *qui n'est pas binaire, contrairement aux valeurs observées du tableau de données. Nous exploiterons cette remarque dans le chapitre V.*

L'expression E[W(P, X, a, α, ε)] montre que les recherches de ε et des $a_k$ sont indépendantes. Pour un ε fixé appartenant à $]0, \frac{1}{2}[$, Log(ε/(1-ε)) est négatif, et maximiser l'espérance du critère revient donc à minimiser :

$$C(P, a) = \sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} |\alpha - a_k^j| \right)$$

## 2.2 MINIMISATION DE C(P, .)

Pour les besoins de cette minimisation, nous complétons les notations en appelant :

$m_k^j$ le nombre de données manquantes pour la variable $x^j$ dans la classe k,

$t_k^j$ le nombre de 1 pour la variable $x^j$ dans la classe k, et

$q_k^j$ le nombre de 0 pour la variable $x^j$ dans la classe k.

Nous avons donc $m_k^j + t_k^j + q_k^j = Card(P_k) = n_k$ où $n_k$ est l'effectif de la classe $P_k$.

**Proposition 2 :**

*Les composantes des $a_k$ éléments de $\{0, 1\}^p$ minimisant C(P, a) sont définies par :*

*pour k = 1, ..., K et j =1, ..., p,*

*lorsque* $\alpha \in [0, \frac{1}{2}[$

$$\begin{cases} a_k^j = 0 \text{ si } m_k^j \geq \zeta \\ a_k^j = 1 \text{ si } m_k^j \leq \zeta \end{cases}$$

*lorsque* $\alpha = \frac{1}{2}$

$$\begin{cases} a_k^j = 0 \text{ si } q_k^j > t_k^j \\ a_k^j = 1 \text{ sinon} \end{cases}$$

*lorsque* $\alpha \in ]\frac{1}{2}, 1]$

$$\begin{cases} a_k^j = 0 \text{ si } m_k^j \leq \zeta \\ a_k^j = 1 \text{ si } m_k^j \geq \zeta \end{cases}$$

*où* $\xi = \dfrac{t_k^j - q_k^j}{1-2\alpha}$.

**Preuve :**

Minimiser C(P, a) revient à chercher pour tout k = 1, ... , K et j =1, ..., p les $a_k^j$ minimisant :

$$\sum_{i \in P_k} |x_i^j - a_k^j| + m_k^j |\alpha - a_k^j| = t_k^j + (q_k^j - t_k^j) a_k^j + m_k^j |\alpha - a_k^j|$$

---

Nous noterons : $B_k^j = t_k^j + (q_k^j - t_k^j)a_k^j + m_k^j \mid \alpha - a_k^j \mid$

Comme $a_k^j \in \{0, 1\}$, il suffit de comparer les deux termes :

$(t_k^j + m_k^j \alpha)$ et $(q_k^j + m_k^j (1-\alpha))$ qui correspondent respectivement à $B_k^j$ lorsque $a_k^j = 0$ et $a_k^j = 1$. Pour cela, nous étudions le signe de $(t_k^j - q_k^j) + m_k^j (2\alpha - 1) = D_k^j$ .

Lorsque $\alpha \in [0, \frac{1}{2}[$, nous avons :

$$a_k^j = 0 \Leftrightarrow D_k^j \leq 0 \Leftrightarrow m_k^j \geq \xi,$$

$$a_k^j = 1 \Leftrightarrow D_k^j \geq 0 \Leftrightarrow m_k^j \leq \xi.$$

Lorsque $\alpha \in ]\frac{1}{2}, 1]$, nous pouvons en déduire :

$$a_k^j = 0 \Leftrightarrow D_k^j \leq 0 \Leftrightarrow m_k^j \leq \xi,$$

$$a_k^j = 1 \Leftrightarrow D_k^j \geq 0 \Leftrightarrow m_k^j \geq \xi.$$

Dans le cas où $\alpha = 1/2$, le signe dépend uniquement du terme $(t_k^j - q_k^j)$ et nous avons :

$$a_k^j = 0 \text{ si } q_k^j > t_k^j \text{ et } a_k^j = 1 \text{ sinon.}$$

Nous notons que le choix des $a_k^j$ est arbitraire dans les cas suivants :

$$\alpha \neq 1/2 \text{ avec } m_k^j = \xi, \text{ et } \alpha = 1/2 \text{ avec } q_k^j = t_k^j.$$

#

**Remarque :**

*Quand le paramètre $\alpha$ est différent de 1/2, il a une influence sur le calcul des noyaux lorsque le nombre des données manquantes est important. D'autre part, lorsque $\alpha$ est égal à 1/2, son rôle est passif et tout se passe comme si lors de la recherche des noyaux, les données manquantes de chaque classe pour chaque variable étaient reconstituées par le noyau correspondant.*

Nous définissons dans le paragraphe suivant l'algorithme utilisé.

## 2.3 ALGORITHME

Comme dans la méthode MNDBIN, nous utilisons un algorithme itératif caractérisé par deux étapes : affectation et représentation.

***Etape d'affectation** : (recherche des classes)*

Le terme npLog(1-ε) de l'expression de **E[W(P, X, a, α, ε)]** est constant lors de cette étape. Puisque nous cherchons à maximiser **E[W(P, X, a, α, ε)]**, nous affectons $x_i$ à la classe $P_k$ qui minimise

$$\sum_{j\in O_i} |x_i^j - a_k^j| + \sum_{j\in M_i} |\alpha - a_k^j|$$

***Etape de représentation** : (recherche des $a_k^j$)*

Nous associons à chaque classe $P_k$ le vecteur $a_k$ défini dans la proposition 2.

Notons que la valeur ε n'est pas intervenue dans cet algorithme qui minimise le critère **C(P, a)**. S'il est nécessaire d'estimer ε, il suffit de maximiser l'espérance du critère.

Si nous notons $$e = \sum_{k=1}^{K} \sum_{i\in P_k} \left( \sum_{j\in O_i} |x_i^j - a_k^j| + \sum_{j\in M_i} |\alpha - a_k^j| \right),$$

la valeur du critère obtenu à la convergence de l'algorithme est :

$$E[W(P, X, a, \alpha, \varepsilon)] = (np-e)\text{Log}(1-\varepsilon) + e\,\text{Log}(\varepsilon) \qquad (2.3.1)$$

Le paramètre ε qui maximise (2.3.1) est défini par :

$$\frac{\partial E[W(P, X, a, \alpha, \varepsilon)]}{\partial \varepsilon} = 0 \quad \Rightarrow \quad \varepsilon = \frac{e}{np}$$

La convergence de l'algorithme se démontre de la même façon que pour les algorithmes de type Nuées Dynamiques.

**Remarque :**

*Dans le cas où $\alpha = 1/2$, la mesure de dissimilarité utilisée apparaît comme une "pseudo distance" $L_1$. Les écarts sont calculés uniquement sur les composantes observées simultanément. Nous traiterons ce cas particulier dans le paragraphe 3.*

## 2.4 EXEMPLE SIMPLE D'APPLICATION

Pour illustrer cette méthode, nous la développons sur un exemple très simple à partir d'un tableau de données binaires croisant 10 individus identifiés par les lettres **a** à **j** et 10 variables identifiées par les nombres **1** à **10** ( tableau 1).

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

Tableau 1: tableau initial.

Nous appliquons la méthode MNDBIN rappelée dans le chapitre I en demandant trois classes. A la convergence, la meilleure partition obtenue est : **A** = { {**a, b, c, i**}, {**h**}, {**d, e, f, g, j**} }. Dans le tableau 2, nous présentons les noyaux associés respectivement aux trois classes.

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{a_1}$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| $\mathbf{a_2}$ | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $\mathbf{a_3}$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 |

Tableau 2 : tableau des noyaux.

A partir du tableau initial, pour obtenir un tableau comportant des données manquantes, nous détruisons au hasard 20% des données (tableau 3).

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

Tableau 3 : tableau obtenu avec 20% de données manquantes.

Nous appliquons notre méthode MNDM en demandant trois classes. Nous prenons comme valeur de $\alpha$ la fréquence des 1 calculée sur les valeurs observées dans le tableau des données. Dans notre cas, la valeur du paramètre $\alpha$ est égale à **0.53**. Dans les tableaux 4 et 5, nous représentons respectivement le tableau initial réordonné et le tableau des noyaux.

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

Tableau 4 : la meilleure partition obtenue.

La valeur du critère est **10.13** et la meilleure partition n'est autre que **A**.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| $a_2$ | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 |
| $a_3$ | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |

Tableau 5 : tableau des noyaux.

## 2.5 GENERALISATION

L'algorithme précédent peut être développé avec les deux autres modèles binaires. En outre, nous avons noté que nous pouvions étendre l'hypothèse sur la distribution des données manquantes en considérant $\alpha$ comme un vecteur formé de p valeurs $\alpha_j$ dépendant de chaque variable. Il en résulte ainsi six possibilités dont la première a été étudiée dans le paragraphe précédent. Nous décrivons rapidement deux autres situations avec $\alpha$, un vecteur dont les composantes dépendent de chaque variable.

### 2.5.1 $\varepsilon$ paramètre dépendant de chaque variable

Nous considérons que chaque variable $x^j$ suit une distribution de Bernoulli de paramètre $\varepsilon^j$. L'espérance de vraisemblance classifiante s'écrit dans ce cas :

$$E[W_1(P,X,a,\alpha,\varepsilon)] = \sum_{k=1}^{K} \sum_{i \in P_k} \{-( \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | x_i^j - a_k^j | + \sum_{j \in M_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | \alpha^j - a_k^j | )\}$$

$$+ n \sum_{j \in J} \text{Log}(1-\varepsilon^j)$$

$$= - \sum_{k=1}^{K} \sum_{i \in P_k} d_\varepsilon(x_i, a_k) + A \qquad (2.5.1.1)$$

où $d_\varepsilon(x_i, a_k) = \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | x_i^j - a_k^j | + \sum_{j \in M_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | \alpha^j - a_k^j |$ et $A = n \sum_{j \in J} \text{Log}(1-\varepsilon^j)$

***Etape d'affectation** (recherche des classes)*

Lors de cette étape, le terme A est constant. Nous affectons $x_i$ à la classe $P_k$ qui minimise $d_\varepsilon(x_i, a_k)$.

***Etape de représentation** (recherche des $a_k^j$ et $\varepsilon^j$)*

Quelles que soient les valeurs $\varepsilon^j$, les $a_k^j$ sont nécessairement celles calculées dans la proposition 2. Il ne reste plus qu'à déterminer les valeurs $\varepsilon^j$ maximisant (2.5.1.1).

$E[W_1(P, X, a, \alpha, \varepsilon)]$ s'écrit également :

$$E[W_1(P, X, a, \alpha, \varepsilon)] = \sum_{j \in O_i} (\text{Log} \frac{\varepsilon^j}{1-\varepsilon^j}) e_j + \sum_{j \in M_i} (\text{Log} \frac{\varepsilon^j}{1-\varepsilon^j}) \beta_j + n \sum_{j \in J} \text{Log}(1-\varepsilon^j)$$

$$= \sum_{j \in J} F(\varepsilon^j) \qquad (2.5.1.2)$$

où $e_j = \sum_{k=1}^{K} \sum_{i \in P_k} | x_i^j - a_k^j |$ pour $j \in O_i$

$\beta_j = \sum_{k=1}^{K} \sum_{i \in P_k} | \alpha^j - a_k^j |$ pour $j \in M_i$

Les composantes $(\varepsilon^j, j \in J)$ qui maximisent (2.5.1.2) sont définies par :

$$\frac{\partial F}{\partial \varepsilon^j} = 0 \quad \Rightarrow \quad \frac{e_j + \beta_j}{\varepsilon^j} - \frac{n - e_j - \beta_j}{1-\varepsilon^j} = 0$$

$$\Rightarrow \quad \varepsilon^j = \frac{e_j + \beta_j}{n}$$

### 2.5.2 ε paramètre dépendant de chaque classe et de chaque variable

Nous considérons que dans chaque classe k, chaque variable $x^j$ suit une distribution de Bernoulli de paramètre $\varepsilon_k^j$. Dans ce cas, l'espérance du critère de vraisemblance classifiante s'écrit :

$$E[W_2(P,X,a,\alpha,\varepsilon)] = \sum_{k=1}^{K} \sum_{i\in P_k} \{-( \sum_{j\in O_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j| + \sum_{j\in M_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | \alpha^j - a_k^j| )$$

$$+ \sum_{j\in J} \text{Log}(1-\varepsilon_k^j) \}$$

$$= \sum_{k=1}^{K} \sum_{i\in P_k} \{-d_{\varepsilon_k}(x_i, a_k) + A_k\} \qquad (2.5.2.1)$$

où $d_{\varepsilon_k}(x_i, a_k) = \sum_{j\in O_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j| + \sum_{j\in M_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | \alpha^j - a_k^j|$ et $A_k = \sum_{j\in J} \text{Log}(1-\varepsilon_k^j)$.

***Etape d'affectation** : (recherche des classes)*

Lors de cette étape, le terme $A_k$ n'est pas constant. Nous affectons $x_i$ à la classe $P_k$ qui minimise $d_{\varepsilon_k}(x_i, a_k) - A_k$.

***Etape de représentation** : (recherche des $a_k^j$ et $\varepsilon_k^j$)*

Quelles que soient les valeurs $\varepsilon_k^j$, les $a_k^j$ sont nécessairement celles calculées dans la proposition 2. Il ne reste plus qu'à déterminer les $\varepsilon_k^j$ maximisant, pour chaque classe $P_k$ et chaque variable $x^j$, la quantité :

$$C = \sum_{i\in P_k} \{ (\text{Log} \frac{\varepsilon_k^j}{1-\varepsilon_k^j}) | x_i^j - a_k^j| + (\text{Log} \frac{\varepsilon_k^j}{1-\varepsilon_k^j}) | \alpha^j - a_k^j| + \text{Log}(1-\varepsilon_k^j) \}$$

$$= (\text{Log} \frac{\varepsilon_k^j}{1-\varepsilon_k^j}) \epsilon_k^j + (\text{Log} \frac{\varepsilon_k^j}{1-\varepsilon_k^j}) \beta_k^j + n_k \text{Log}(1-\varepsilon_k^j)$$

$$= G(\varepsilon_k^j) \qquad (2.5.2.2)$$

$$\text{où} \quad e_k^j = \sum_{i \in P_k} |x_i^j - a_k^j| \qquad \text{pour } k = 1, ..., K \text{ et } j \in O_i$$

$$\beta_k^j = \sum_{i \in P_k} |\alpha^j - a_k^j| \qquad \text{pour } k = 1, ..., K \text{ et } j \in M_i$$

Les composantes $(\varepsilon_k^j, k \in \{1, ..., K\}, j \in J)$ qui maximisent (2.5.2.2) sont définies par :

$$\frac{\partial G}{\partial \varepsilon_k^j} = 0 \quad \Rightarrow \quad \varepsilon_k^j = \frac{e_k^j + \beta_k^j}{n_k}$$

**Remarque :**

*Les $\varepsilon^j$ ( respectivement $\varepsilon_k^j$ ) maximisant $E[W_1(P, X, a, \alpha, \varepsilon)]$ ( respectivement $E[W_2(P, X, a, \alpha, \varepsilon)]$) appartiennent à $]0, \frac{1}{2}[$ sauf dans les cas très particuliers $e_j + \beta_j = 0$ et $e_j + \beta_j = n/2$ (respectivement $e_k^j + \beta_k^j = 0$ et $e_k^j + \beta_k^j = n_k/2$ ). Comme nous l'avons noté dans le premier chapitre, pour remédier à cet inconvénient, il suffit de prendre dans l'algorithme les valeurs de ces paramètres calculées dans l'étape précédente.*

Nous présentons l'organigramme qui décrit l'algorithme.

**(0) Initialisation**

a/ tirage d'une partition initiale
b/ calcul des fréquences de 1 pour chaque variable ou pour le tableau entier.

**(1) Représentation**

a/ calcul des noyaux
b/ calcul des paramètres des lois de Bernoulli

**(2) Affectation**

**(3) Etude de la convergence**

Si non convergence retour à **(1)**.

## 3. INTERPRETATIONS DES MESURES DE DISSIMILARITE

Comme toutes les méthodes de type Nuées Dynamiques, notre méthode s'appuie sur une mesure de dissimilarité entre un vecteur $x_i$ et un noyau. En effet, nous avions eu à optimiser un critère qui s'écrivait sous la forme :

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

Dans notre situation où il y a des données manquantes, il peut être intéressant d'étudier la mesure de ressemblance D que notre démarche nous a conduite à utiliser.

Lors de l'étape d'affectation, nous avons vu précédemment, que dans le cas le plus simple, les individus sont affectés aux centres les plus proches au sens de la mesure de dissimilarité suivante :

$$d(x_i, a_k) = \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} |\alpha - a_k^j|$$

Signalons tout d'abord que cette approximation n'est pas une distance. Cette mesure se calcule entre un vecteur qui peut être incomplet et un vecteur complet. Rappelons qu'en classification et en particulier pour la méthode des Nuées Dynamiques, il est moins utile de connaître les valeurs exactes des distances que leurs grandeurs relatives.

**Exemple** :

Soient $x_i = (1, 1, ?, 1, 0, ?)$, $a_1 = (1, 0, 1, 0, 0, 1)$ et $a_2 = (0, 0, 0, 0, 0, 1)$. En prenant $\alpha = 0.6$, nous avons :

$$d(x_i, a_1) = (1-1) + (1-0) + (1-0.6) + (1-0) + (0-0) + (1-0.6) = 2.8$$

et

$$d(x_i, a_2) = (1-0) + (1-0) + (0.6-0) + (1-0) + (0-0) + (0.6-0) = 4.2$$

Ainsi le vecteur $x_i$ est plus proche du vecteur complet $a_1$ que du vecteur complet $a_2$.

Dans ce qui suit, nous nous intéressons au cas où $\alpha = 1/2$. Notre mesure de dissimilarité s'écrit alors, entre deux vecteurs : un vecteur $x_i$ qui peut être incomplet et un vecteur complet $a_k$ :

$$d(x_i, a_k) = \sum_{j \in O_i} | x_i^j - a_k^j |$$

Cette "pseudo-distance" $L_1$ est la somme en valeur absolue des écarts entre les composantes relevées simultanément sur les deux vecteurs. Ainsi, nous admettons que deux vecteurs qui sont proches dans l'espace des variables mesurées, le seront aussi pour des variables non mesurées. Il est raisonnable d'accepter cette approximation au vu de l'information présente.

**Exemple** :

Soient $x_i = (1, 1, ?, 1, 0, ?)$, $a_1 = (1, 0, 1, 0, 0, 1)$ et $a_2 = (0, 0, 0, 0, 0, 1)$.
La 3ème et 6ème composantes ne seront pas prises en compte :

$$d(x_i, a_1) = (1-1) + (1-0) + (1-0) + (0-0) = 2$$

et

$$d(x_i, a_2) = (1-0) + (1-0) + (1-0) + (0-0) = 3$$

Le vecteur incomplet $x_i$ est plus proche de $a_1$ que de $a_2$.

Dans le cas des données continues, Fèvre (1980) projette le vecteur $x_i$ incomplet, élément de $\mathbf{R}^p$, sur un sous-espace vectoriel de $\mathbf{R}^p$ isomorphe à $\mathbf{R}^q$ où q désigne le nombre de composantes observées du vecteur $x_i$. En associant un projecteur à chaque individu, la distance euclidienne est approximée par une "pseudo-distance" euclidienne définie uniquement sur les composantes observées simultanément. L'algorithme utilisant cette mesure a donné des résultats satisfaisants. Le tableau 6 donne un aperçu sur les points communs entre la méthode de Fèvre et la méthode MNDM dans le cas où la mesure est la "pseudo-distance" $L_1$. Notons que si nous imposons la contrainte que les noyaux soient binaires, les deux méthodes sont identiques.

Nous pouvons ainsi donner une interprétation de la méthode de Fèvre en tant que modèle de mélanges qui est la suivante : dans ce cas, le critère métrique défini par cette pseudo-distance euclidienne s'interprète comme une vraisemblance induite par un mélange fini de lois gaussiennes dont les variances sont supposées connues et fixes et $p_k = 1/K$. Quant aux variances des données manquantes, elles sont supposées infinies. Ainsi, dans les deux cas continu et binaire, cela revient en sorte que la partie "localisation" (voir chapitre I) du paramétrage n'intervient pas.

| Données quantitatives | Données binaires |
|---|---|
| **Pseudo-distance euclidienne**<br>$d_e^2(x_i, g_k) = \sum_{j \in O_i} (x_i^j - g_k^j)^2$ | **Pseudo-distance $L_1$**<br>$d(x_i, a_k) = \sum_{j \in O_i} \lvert x_i^j - a_k^j \rvert$ |
| Etape de représentation<br>Les noyaux sont des **pseudo-centres de gravité**. Les composantes sont calculées sur les valeurs observées de chaque classe pour chaque variable | Etape de représentation<br>Les noyaux sont des **pseudo-centres médians**. Les médianes sont calculées sur les valeurs observées de chaque classe pour chaque variable |
| Etape d'affectation<br>Les individus sont affectés aux centres les plus proches au sens de $\mathbf{d_e}$. | Etape d'affectation<br>Les individus sont affectés aux centres les plus proches au sens de **d**. |

Tableau 6 : comparaion de la méthode de Fèvre et de la méthode MNDM lorsque $\alpha$=1/2

Nous avons considéré la "pseudo-distance" $L_1$ sous sa forme la plus simple. Rappelons qu'elle découle du modèle suivant une distribution de Bernoulli de paramètre identique pour toutes les variables. Dans le cas où les paramètres dépendent de chaque variable, la mesure de dissimilarité apparaît comme une distance $L_1$ pondérée par les quantités ( Log $((1-\varepsilon^j)/\varepsilon^j)$ ; $j \in J$) et s'écrit :

$$d(x_i, a_k) = \sum_{j \in O_i} \text{Log}\left(\frac{1-\varepsilon^j}{\varepsilon^j}\right) \lvert x_i^j - a_k^j \rvert$$

Dans le cas le plus général, la mesure s'écrit :

$$d(x_i, a_k) = \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} \lvert x_i^j - a_k^j \rvert + \sum_{j \in J} \text{Log}(1-\varepsilon_k^j)$$

En imposant aux coefficients de la mesure la contrainte $\prod_{j \in J} (1-\varepsilon_k^j)$ = constante, la mesure obtenue est une pseudo-distance $L_1$ pondérée avec un terme additif constant qui ne jouera aucun rôle. Le critère métrique obtenu est alors équivalent à un critère métrique défini uniquement à l'aide d'une pseudo distance $L_1$ pondérée (sans terme additif).

# 4. EXPERIENCES NUMERIQUES

## 4.1 APPLICATIONS A UN TABLEAU CONSTRUIT SUIVANT UN MODELE

G. Govaert (1988) a montré comment l'identification d'un mélange de distributions de Bernoulli avec le même paramètre pour toutes les classes et pour toutes les variables correspond au critère de la méthode MNDBIN (méthode des nuées dynamiques sur tableau binaire). Le programme HASBIN a été construit pour générer de tels tableaux. Celui-ci 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. De cette manière, nous créons un tableau de données complet 100x10 que nous nommerons XER. Ce tableau ainsi que les données correspondant aux mesures de 10 variables sur 3 classes sont présentés ci-dessous.

**Classe 1**
{1, 2, 3, 9, 16, 17, 22, 23, 27, 31, 33, 34, 35, 36, 40, 41, 44, 50, 51, 52, 53, 55, 57, 59, 63, 65, 67, 83, 84, 95, 96, 97, 99}

**Classe 2**
{8, 15, 18, 21, 25, 26, 28, 30, 43, 46, 49, 54, 56, 58, 62, 68, 70, 71, 74, 77, 78, 79, 80, 82, 85, 86, 87, 88, 90, 92, 93, 94, 98, 100}

**Classe 3**
{4, 5, 6, 7, 10, 11, 12, 13, 14, 19, 20, 24, 29, 32, 37, 38, 39, 42, 45, 47, 48, 60, 61, 64, 66, 69, 72, 73, 75, 76, 81, 89, 91}

| | | | |
|---|---|---|---|
| 1 0000111100 | 26 1100001101 | 51 0000111100 | 76 1110000011 |
| 2 0000111100 | 27 0000111110 | 52 0001111100 | 77 1110001101 |
| 3 0000111100 | 28 1100000100 | 53 0000111100 | 78 1100000101 |
| 4 1111000011 | 29 1111100011 | 54 1101001100 | 79 1100001001 |
| 5 1111010111 | 30 1100101101 | 55 0000101100 | 80 1100101101 |
| 6 1111000011 | 31 0001111100 | 56 1100001101 | 81 1111100011 |
| 7 1111010011 | 32 1111010011 | 57 0001111100 | 82 1100001101 |
| 8 1100101101 | 33 0000111100 | 58 1100001101 | 83 0000111100 |
| 9 0000111100 | 34 0000111100 | 59 0000111100 | 84 0000111100 |
| 10 1111000010 | 35 0000111100 | 60 1111000011 | 85 1100001101 |
| 11 1110000011 | 36 0010111100 | 61 1111000011 | 86 1100001101 |
| 12 1111000011 | 37 1111000011 | 62 1110001101 | 87 1100001101 |
| 13 1110000111 | 38 1111000010 | 63 1000111100 | 88 1100001101 |
| 14 1111000011 | 39 0111000011 | 64 1111000011 | 89 1111000011 |
| 15 1100001001 | 40 0000101100 | 65 0000111100 | 90 1100001101 |
| 16 0000111100 | 41 0000111101 | 66 1111000011 | 91 1111100011 |
| 17 0000111100 | 42 0111001011 | 67 0000111100 | 92 1000101111 |
| 18 1100000111 | 43 1100001101 | 68 1100001101 | 93 1100001101 |
| 19 1111000011 | 44 0010111100 | 69 1111000001 | 94 1100001101 |
| 20 1111100011 | 45 1011000011 | 70 1100001111 | 95 1000111100 |
| 21 1101001101 | 46 1100001101 | 71 1100001100 | 96 0000111100 |
| 22 0010111100 | 47 1111000011 | 72 1111010111 | 97 0000111100 |
| 23 0000111100 | 48 1111000011 | 73 1111000001 | 98 1100101101 |
| 24 1111000011 | 49 1100001001 | 74 1100001101 | 99 0000111100 |
| 25 1100001101 | 50 0000111100 | 75 1110000011 | 100 1100001101 |

A partir de ce tableau, nous avons simulé des données manquantes de façon aléatoire et progressive pour tous les essais. Nous appliquons ensuite notre méthode MNDM avec différentes options. Ces dernières dépendent des paramètres $\alpha$ et du modèle choisi. Il nous est donc facile de juger de la qualité des classifications que nous obtiendrons par rapport à la classification idéale. En effet, il suffit de calculer le nombre d'objets mal classés et d'en déduire les pourcentages de ces objets. Nous verrons ainsi l'évolution de ces pourcentages dans les différents cas.

### 4.1.1 Lorsque $\alpha$ est un réel estimé sur toutes les données observées du tableau

Nous avons représenté les résultats dans le tableau 7. Le paramètre $\alpha$ est estimé par la fréquence **f** des 1 calculée sur les valeurs observées du tableau :

**f** = nombre des 1 dans X / (nxp -m).

| % de destruction | pourcentage d'objets mal classés | | |
|---|---|---|---|
| | $\varepsilon$ | $\varepsilon^j$ | $\varepsilon_k^j$ |
| 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 |
| 16 | 1 | 1 | 1 |
| 20 | 2 | 2 | 2 |
| 24 | 2 | 2 | 2 |
| 28 | 3 | 3 | 3 |
| 32 | 6 | 6 | 6 |
| 36 | 9 | 9 | 6 |
| 40 | 22 | 37 | 4 |
| 44 | 38 | 41 | 8 |
| 48 | * | * | 9 |
| 50 | * | * | 9 |
| 52 | * | * | 21 |
| 56 | * | * | 23 |

Tableau 7 : évolution des pourcentages d'objets mal classés.

* Il nous est impossible d'avoir 3 classes.

Nous constatons que lorsque le pourcentage des données manquantes est en dessous de 36 %, la méthode MNDM avec les trois variantes donne des résultats presque identiques. D'autre part, lorsque le taux de destruction atteint 40 %, l'algorithme utilisant la mesure de dissimilarité adaptative pour chaque classe et pour chaque

variable donne des résultats satisfaisants. Notons toutefois que la convergence dans ce cas est lente. Par contre avec les deux premières variantes, la convergence est rapide mais le pourcentage d'objets mal classés est important. En effet, deux classes tendent à se confondre d'où une augmentation brusque du taux d'objets mal classés.

### 4.1.2 Lorsque $\alpha$ est un vecteur dont les composantes dépendent de chaque variable

Les composantes de $\alpha$ correspondent aux fréquences $f_j$ des 1 pour chaque variable :

$$f_j = \text{nombre des 1 pour la variable } x^j / (n - m^j).$$

Les résultats sont présentés comme précédemment dans le tableau 8.

| % de destruction | pourcentage d'objets mal classés | | |
|---|---|---|---|
| | $\varepsilon$ | $\varepsilon^j$ | $\varepsilon^j_k$ |
| 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 |
| 16 | 0 | 2 | 0 |
| 20 | 1 | 3 | 1 |
| 24 | 2 | 3 | 1 |
| 28 | 1 | 3 | 2 |
| 32 | 3 | 3 | 4 |
| 36 | 4 | 4 | 4 |
| 40 | 6 | 7 | 5 |
| 44 | 7 | 9 | 5 |
| 48 | 8 | 8 | 5 |
| 50 | 6 | 10 | 6 |
| 52 | 12 | 12 | 6 |
| 56 | 17 | 20 | 5 |
| 64 | 28 | 31 | 14 |

Tableau 8 : évolution des pourcentages d'objets mal classés.

Les résultats sont nettement meilleurs que dans le cas où $\alpha$ est le même pour toutes les variables. Comme précédemment, la troisième option donne de meilleurs résultats. Ceci est illustré dans la figure 1 en notant MNDM i (i = 1, 2, 3) la méthode MNDM correspondant aux trois options associées respectivement aux différents choix des paramètres de la distribution de Bernoulli ($\varepsilon(1)$, $\varepsilon^j(2)$ et $\varepsilon^j_k(3)$).

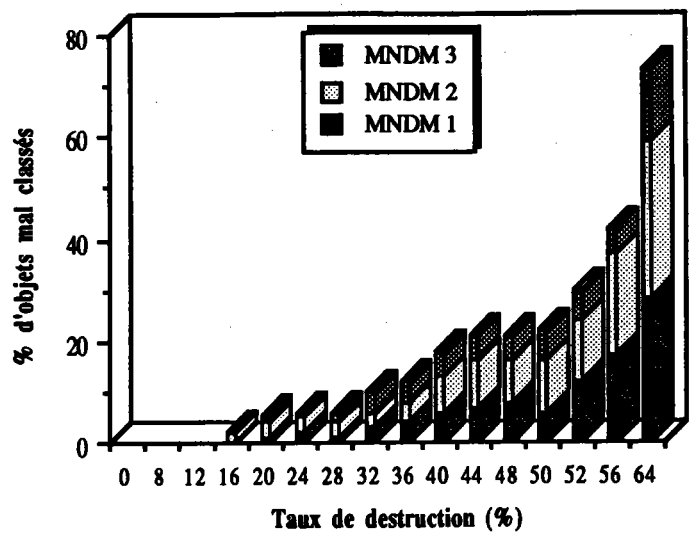



Figure 1 : comparaison des trois variantes de MNDM.

Nous présentons ci-dessous le tableau de données comportant 50 % de données manquantes XER50. Les données manquantes sont codées "9". Les pages qui suivent contiennent les résultats de la méthode MNDM avec les trois variantes.

| | | | |
|---|---|---|---|
| 1 0009111190 | 26 1100909109 | 51 0909191199 | 76 9199090991 |
| 2 9090111990 | 27 0909919110 | 52 0999911990 | 77 9110999999 |
| 3 9099919909 | 28 1199000909 | 53 0009919190 | 78 9199900991 |
| 4 1199000919 | 29 1199900991 | 54 1901999199 | 79 9999091909 |
| 5 9999090999 | 30 1109191191 | 55 0900991190 | 80 9100191101 |
| 6 1111999011 | 31 9091991190 | 56 1909901909 | 81 9191190999 |
| 7 1111019091 | 32 9199910911 | 57 9091919900 | 82 9990991101 |
| 8 9900991191 | 33 9000111900 | 58 9900009199 | 83 9990191190 |
| 9 9009111900 | 34 9000111100 | 59 0990999990 | 84 0909111999 |
| 10 1919990019 | 35 9990111199 | 60 9999909919 | 85 1190091991 |
| 11 9999990019 | 36 9919191109 | 61 9119909919 | 86 1999901109 |
| 12 1919000091 | 37 1991099911 | 62 9119901999 | 87 1100099991 |
| 13 9190990991 | 38 9919909999 | 63 9999999999 | 88 9109999901 |
| 14 9199000091 | 39 0119900991 | 64 9911009099 | 89 9911990999 |
| 15 9990991009 | 40 0009909909 | 65 9909991190 | 90 1190091199 |
| 16 9990999990 | 41 9090119999 | 66 1119999019 | 91 9991109091 |
| 17 9099199909 | 42 9919909911 | 67 0900191909 | 92 1900101991 |
| 18 1990090191 | 43 9199909999 | 68 1100901999 | 93 9909001191 |
| 19 1191000919 | 44 0010191900 | 69 9191000991 | 94 1990099191 |
| 20 9919999099 | 45 1099099919 | 70 9900091119 | 95 1909111909 |
| 21 1901991999 | 46 1199009101 | 71 9190999109 | 96 9909919909 |
| 22 0010199109 | 47 9911999999 | 72 9199019199 | 97 9000191100 |
| 23 0999111909 | 48 9919990911 | 73 9199900009 | 98 9909909909 |
| 24 1119009099 | 49 1100001001 | 74 1990901999 | 99 0999191190 |
| 25 9990999109 | 50 0999191190 | 75 9199090911 | 100 1990901901 |

COMMANDE : MNDM <>  utilisant une mesure sans ponderation.

valeur du critère obtenu  257.1028000

Partition

classe 1 : 38 elements

8 15 18 21 25 26 28 30 36 38 43 46 47 49 54 56 58 62 63 68 70
71 72 74 77 79 80 82 85 86 87 88 90 92 93 94 98 100

classe 2 : 31 elements

4 5 6 7 10 11 12 13 14 19 20 24 29 32 37 39 42 45 48 60 61
64 66 69 73 75 76 78 81 89 91

classe 3 : 31 elements

1 2 3 9 16 17 22 23 27 31 33 34 35 40 41 44 50 51 52 53 55
57 59 65 67 83 84 95 96 97 99

Objets mal classes = (36, 38, 47, 63, 72, 78)

TABLEAU DES VALEURS IDEALES

```
  VVVVVVVVVV
  AAAAAAAAAA
           1
  1234567890
1 11    11 1
2 111     11
3    1111
```

TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

| | VA 1 | VA 2 | VA 3 | VA 4 | VA 5 | VA 6 | VA 7 | VA 8 | VA 9 | VA 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 54.0 | 50.0 | 49.0 | 52.0 | 49.0 | 49.0 | 39.0 | 54.0 | 53.0 | 51.0 |

TABLEAU DES ALPHA

| | VA 1 | VA 2 | VA 3 | VA 4 | VA 5 | VA 6 | VA 7 | VA 8 | VA 9 | VA 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.65 | 0.68 | 0.39 | 0.29 | 0.49 | 0.37 | 0.66 | 0.70 | 0.34 | 0.61 |

TABLEAU DES ECARTS

| | VA 1 | VA 2 | VA 3 | VA 4 | VA 5 | VA 6 | VA 7 | VA 8 | VA 9 | VA 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 6.96 | 7.04 | 11.67 | 7.37 | 14.29 | 8.82 | 7.51 | 7.48 | 8.15 | 9.31 |
| 2 | 7.61 | 4.84 | 10.94 | 15.13 | 9.84 | 7.22 | 7.87 | 13.22 | 11.55 | 5.82 |
| 3 | 10.78 | 10.88 | 7.49 | 6.67 | 6.12 | 9.78 | 3.79 | 5.17 | 6.45 | 7.35 |

TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

| | VA 1 | VA 2 | VA 3 | VA 4 | VA 5 | VA 6 | VA 7 | VA 8 | VA 9 | VA 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |
| 2 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |
| 3 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 | 0.27 |

TABLEAU DES COEFFICIENTS DE PONDERATION

| | VA 1 | VA 2 | VA 3 | VA 4 | VA 5 | VA 6 | VA 7 | VA 8 | VA 9 | VA 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 2 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| 3 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |

```
COMMANDE : MNDM <> utilisant une mesure ponderee (ponderation differente par variable)

valeur du critere obtenu    566.6147000

          Partition

  classe  1 :    42 elements
  ----------
8    15   18   21   25   26   28   30   36   37   38   43   45   46   47   49   54   56   58   60   62
63   68   70   71   72   74   77   79   80   82   85   86   87   88   90   92   93   94   96   98  100

  classe  2 :    30 elements
  ----------
1    2    3    9   16   17   22   23   27   31   33   34   35   40   41   44   50   51   52   53   55
57   59   65   67   83   84   95   97   99

  classe  3 :    28 elements
  ----------
4    5    6    7   10   11   12   13   14   19   20   24   29   32   39   42   48   61   64   66   69
73   75   76   78   81   89   91

objets mal classes = (36, 37, 38, 45, 47, 60, 63, 72, 78, 96)

     TABLEAU DES VALEURS IDEALES

      VVVVVVVVVV
      AAAAAAAAAA
                1
      1234567890
   1 11    11 1
   2     1111
   3 111     11

          TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

           V     V     V     V     V     V     V     V     V     V
           A     A     A     A     A     A     A     A     A     A
                                                                  1
           1     2     3     4     5     6     7     8     9     0
   1   54.0  50.0  49.0  52.0  49.0  49.0  39.0  54.0  53.0  51.0

          TABLEAU DES ALPHA

           V     V     V     V     V     V     V     V     V     V
           A     A     A     A     A     A     A     A     A     A
                                                                  1
           1     2     3     4     5     6     7     8     9     0
   1    0.65  0.68  0.39  0.29  0.49  0.37  0.66  0.70  0.34  0.61

          TABLEAU DES ECARTS

           V     V     V     V     V     V     V     V     V     V
           A     A     A     A     A     A     A     A     A     A
                                                                  1
           1     2     3     4     5     6     7     8     9     0
   1    7.65  9.00 12.84  9.25 15.27 10.57  8.89  8.70 11.15 10.47
   2   10.13 10.20  7.49  6.38  5.61  9.78  3.44  4.87  6.45  6.73
   3    7.26  3.20  9.12 13.54  9.35  6.47  5.90 11.13 11.55  5.04

          TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

           V     V     V     V     V     V     V     V     V     V
           A     A     A     A     A     A     A     A     A     A
                                                                  1
           1     2     3     4     5     6     7     8     9     0
   1    0.25  0.22  0.29  0.29  0.30  0.27  0.18  0.25  0.29  0.22
   2    0.25  0.22  0.29  0.29  0.30  0.27  0.18  0.25  0.29  0.22
   3    0.25  0.22  0.29  0.29  0.30  0.27  0.18  0.25  0.29  0.22

          TABLEAU DES COEFFICIENTS DE PONDERATION

           V     V     V     V     V     V     V     V     V     V
           A     A     A     A     A     A     A     A     A     A
                                                                  1
           1     2     3     4     5     6     7     8     9     0
   1    1.1   1.2   0.9   0.9   0.8   1.0   1.5   1.1   0.9   1.3
   2    1.1   1.2   0.9   0.9   0.8   1.0   1.5   1.1   0.9   1.3
   3    1.1   1.2   0.9   0.9   0.8   1.0   1.5   1.1   0.9   1.3
```

COMMANDE : MNDM <> utilisant une mesure ponderee (ponderation differente par variable et par classe)

valeur du critere obtenu 549.0522000

Partition

classe 1 : 36 elements

4 5 6 7 10 11 12 13 14 18 19 20 24 28 29 32 37 38 39 42 45
47 48 60 61 64 66 69 73 75 76 77 78 81 89 91

classe 2 : 32 elements

8 15 21 25 26 30 43 46 49 54 56 58 62 63 68 70 71 72 74 79 80
82 85 86 87 88 90 92 93 94 98 100

classe 3 : 32 elements

1 2 3 9 16 17 22 23 27 31 33 34 35 36 40 41 44 50 51 52 53
55 57 59 65 67 83 84 95 96 97 99

objets mal classes =(18, 28, 63, 72, 77, 78)

TABLEAU DES VALEURS IDEALES

| | VVVVVVVVVV<br>AAAAAAAAAA<br>1<br>1234567890 |
|---|---|
| 1 | 111 11 |
| 2 | 11 11 1 |
| 3 | 1111 |

TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 54.0 | 50.0 | 49.0 | 52.0 | 49.0 | 49.0 | 39.0 | 54.0 | 53.0 | 51.0 |

TABLEAU DES ALPHA

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.65 | 0.68 | 0.39 | 0.29 | 0.49 | 0.37 | 0.66 | 0.70 | 0.34 | 0.61 |

TABLEAU DES ECARTS

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.65 | 5.80 | 12.16 | 16.71 | 11.31 | 8.33 | 9.84 | 17.00 | 15.19 | 7.37 |
| 2 | 5.57 | 5.76 | 6.88 | 5.50 | 11.82 | 7.33 | 4.48 | 6.26 | 6.79 | 7.37 |
| 3 | 11.43 | 11.56 | 8.49 | 6.96 | 6.12 | 10.41 | 3.79 | 5.17 | 6.45 | 7.96 |

TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.24 | 0.16 | 0.34 | 0.46 | 0.31 | 0.23 | 0.27 | 0.47 | 0.42 | 0.20 |
| 2 | 0.17 | 0.18 | 0.22 | 0.17 | 0.37 | 0.23 | 0.14 | 0.20 | 0.21 | 0.23 |
| 3 | 0.36 | 0.36 | 0.27 | 0.22 | 0.19 | 0.33 | 0.12 | 0.16 | 0.20 | 0.25 |

TABLEAU DES COEFFICIENTS DE PONDERATION

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 0 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.2 | 1.6 | 0.7 | 0.1 | 0.8 | 1.2 | 1.0 | 0.1 | 0.3 | 1.4 |
| 2 | 1.6 | 1.5 | 1.3 | 1.6 | 0.5 | 1.2 | 1.8 | 1.4 | 1.3 | 1.2 |
| 3 | 0.6 | 0.6 | 1.0 | 1.3 | 1.4 | 0.7 | 2.0 | 1.6 | 1.4 | 1.1 |

### 4.1.3 Lorsque $\alpha$ est égal à 1/2

Comme nous l'avons noté, lorsque $\alpha$ = 1/2, la mesure de dissimilarité utilise uniquement les composantes observées simultanément. Dans ce cas, la méthode MNDM se présente sous la forme la plus simple. Le paramètre $\alpha$ n'a aucune influence sur les étapes d'affectation et de représentation qui caractérisent la méthode MNDM. Nous procédons de la même manière en demandant trois classes. L'évolution du pourcentage d'objets mal classés est présentée pour les trois options dans le tableau 9.

| | pourcentage d'objets mal classés | | |
|---|---|---|---|
| % de destruction | $\varepsilon$ | $\varepsilon^j$ | $\varepsilon_k^j$ |
| 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 |
| 16 | 0 | 0 | 0 |
| 20 | 1 | 2 | 3 |
| 24 | 2 | 2 | 2 |
| 28 | 2 | 3 | 2 |
| 32 | 3 | 3 | 4 |
| 36 | 2 | 4 | 3 |
| 40 | 2 | 3 | 2 |
| 44 | 4 | 4 | 3 |
| 48 | 4 | 5 | 4 |
| 50 | 8 | 5 | 4 |
| 52 | 7 | 9 | 6 |
| 56 | 8 | 8 | 6 |
| 64 | 13 | 11 | 13 |

Tableau 9 : évolution des pourcentages d'objets mal classés.

Les résultats dans ce cas sont meilleurs que dans les cas précédents et la convergence est très rapide. C'est cette option que nous retiendrons. La mesure de dissimilarité utilisée semble être la plus raisonnable. Nous réutiliserons cette mesure sous sa forme la plus généralisée dans le chapitre IV. Comme précédemment, dans la figure 2, nous comparons les trois variantes de la méthode MNDM lorsque $\alpha$ = 1/2.

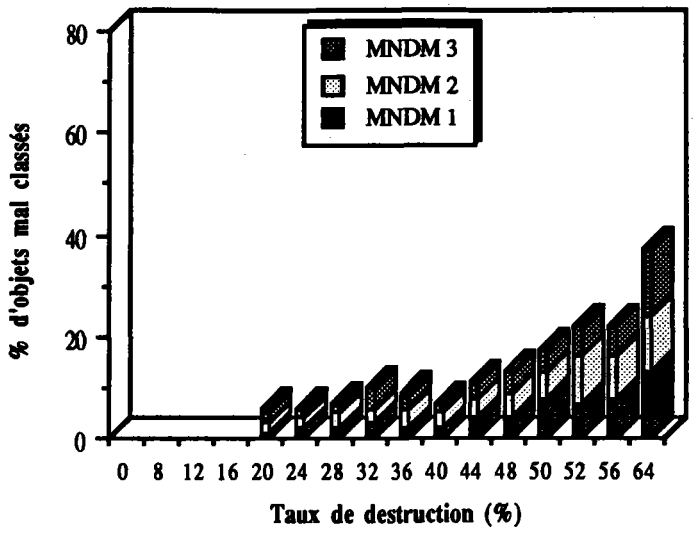



Figure 2 : comparaison des trois variantes de MNDM.

Dans la figure 3 nous comparons la méthode MNDM3 lorsque chacune des composantes de $\alpha$ dépend des variables (A) et MNDM 3 lorsque $\alpha = 1/2$ (B).

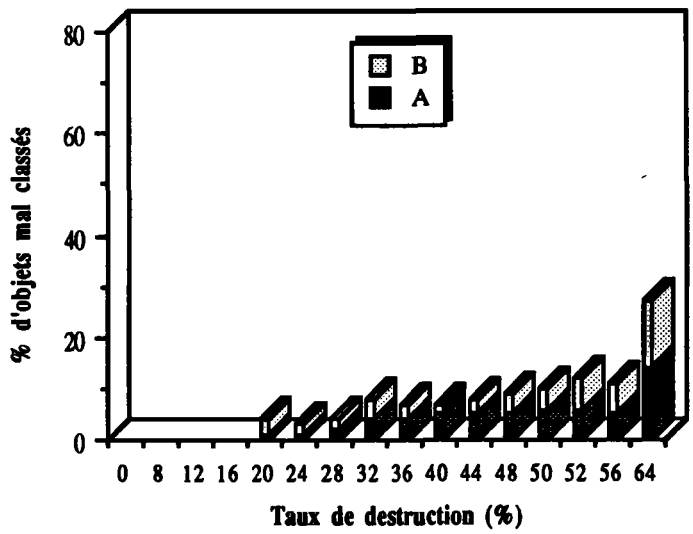



Figure 3 : comparaison de MNDM 3 lorsque $\alpha$ dépend des variables et lorsque $\alpha = 1/2$.

Les trois pages qui suivent contiennent les résultats de la méthode MNDM (avec les trois variantes) appliquée sur XER50.

```
COMMANDE : MNDM <> utilisant une mesure sans ponderation

valeur du critere obtenu      29.0000000

         Partition

  classe  1 :    35 elements
  -----------
1    2    3    9   15   16   17   22   23   25   27   31   33   34   35   36   40   41   44   50   51
52   53   55   57   59   63   65   67   83   84   95   96   97   99

  classe  2 :    37 elements
  -----------
4    5    6    7   10   11   12   13   14   19   20   24   28   29   32   37   38   39   42   43   45
47   48   60   61   62   64   66   69   73   75   76   77   78   81   89   91

  classe  3 :    28 elements
  -----------
8   18   21   26   30   46   49   54   56   58   68   70   71   72   74   79   80   82   85   86   87
88   90   92   93   94   98  100

objets mal classes = {15, 25, 28, 43, 62, 72, 77, 78}

     TABLEAU DES VALEURS IDEALES

      VVVVVVVVVV
      AAAAAAAAAA
               1
      1234567890
   1     1111
   2 1111    11
   3 11    11 1


          TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

           V    V    V    V    V    V    V    V    V    V
           A    A    A    A    A    A    A    A    A    A
                                                        1
           1    2    3    4    5    6    7    8    9    0
   1   54.0 50.0 49.0 52.0 49.0 49.0 39.0 54.0 53.0 51.0


          TABLEAU DES ALPHA

           V    V    V    V    V    V    V    V    V    V
           A    A    A    A    A    A    A    A    A    A
                                                        1
           1    2    3    4    5    6    7    8    9    0
   1    0.50 0.50 0.50 0.50 0.50 0.50 0.50 0.50 0.50 0.50


          TABLEAU DES ECARTS

           V    V    V    V    V    V    V    V    V    V
           A    A    A    A    A    A    A    A    A    A
                                                        1
           1    2    3    4    5    6    7    8    9    0
   1    1.00 0.00 3.00 2.00 0.00 1.00 0.00 1.00 1.00 0.00
   2    1.00 1.00 0.00 2.00 2.00 2.00 1.00 0.00 2.00 0.00
   3    0.00 0.00 0.00 2.00 3.00 1.00 1.00 1.00 1.00 0.00


          TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

           V    V    V    V    V    V    V    V    V    V
           A    A    A    A    A    A    A    A    A    A
                                                        1
           1    2    3    4    5    6    7    8    9    0
   1    0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27
   2    0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27
   3    0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27 0.27


          TABLEAU DES COEFFICIENTS DE PONDERATION

           V    V    V    V    V    V    V    V    V    V
           A    A    A    A    A    A    A    A    A    A
                                                        1
           1    2    3    4    5    6    7    8    9    0
   1    1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
   2    1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
   3    1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0  1.0
```

COMMANDE <> MNDM utilisant une mesure ponderee (ponderation differente par variable)

valeur du critere obtenu 188.1551000

Partition

classe 1 : 33 elements

8 15 18 21 25 26 30 43 46 49 54 56 58 62 63 68 70 71 72 74 79
80 82 85 86 87 88 90 92 93 94 98 100

classe 2 : 35 elements

4 5 6 7 10 11 12 13 14 19 20 24 28 29 32 37 38 39 42 45 47
48 60 61 64 66 69 73 75 76 77 78 81 89 91

classe 3 : 32 elements

1 2 3 9 16 17 22 23 27 31 33 34 35 36 40 41 44 50 51 52 53
55 57 59 65 67 83 84 95 96 97 99

Objets mal classes = {28, 63, 72, 77, 78}

TABLEAU DES VALEURS IDEALES

```
  VVVVVVVVVV
  AAAAAAAAAA
            1
  1234567890
1 11    11 1
2 1111    11
3     1111
```

TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 54.0 | 50.0 | 49.0 | 52.0 | 49.0 | 49.0 | 39.0 | 54.0 | 53.0 | 51.0 |

TABLEAU DES ALPHA

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

TABLEAU DES ECARTS

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.00 | 0.00 | 1.00 | 2.00 | 3.00 | 1.00 | 1.00 | 2.00 | 1.00 | 0.00 |
| 2 | 1.00 | 1.00 | 0.00 | 2.00 | 2.00 | 2.00 | 0.00 | 0.00 | 2.00 | 0.00 |
| 3 | 1.00 | 0.00 | 3.00 | 2.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |

TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.02 | 0.01 | 0.04 | 0.06 | 0.05 | 0.04 | 0.01 | 0.02 | 0.04 | 0.25 |
| 2 | 0.02 | 0.01 | 0.04 | 0.06 | 0.05 | 0.04 | 0.01 | 0.02 | 0.04 | 0.25 |
| 3 | 0.02 | 0.01 | 0.04 | 0.06 | 0.05 | 0.04 | 0.01 | 0.02 | 0.04 | 0.25 |

TABLEAU DES COEFFICIENTS DE PONDERATION

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.9 | 4.6 | 3.2 | 2.8 | 2.9 | 3.2 | 4.6 | 3.9 | 3.2 | 1.1 |
| 2 | 3.9 | 4.6 | 3.2 | 2.8 | 2.9 | 3.2 | 4.6 | 3.9 | 3.2 | 1.1 |
| 3 | 3.9 | 4.6 | 3.2 | 2.8 | 2.9 | 3.2 | 4.6 | 3.9 | 3.2 | 1.1 |

COMMANDE : MNDM <> utilisant une mesure ponderee ( ponderation differente par classe et par variable)

valeur du critere obtenu 108.9583000

Partition

classe 1 : 34 elements

1 2 3 9 16 17 22 23 25 27 31 33 34 35 36 40 41 44 50 51 52
53 55 57 59 63 65 67 83 84 95 96 97 99

classe 2 : 34 elements

8 13 15 18 21 26 28 30 43 46 49 54 56 58 62 68 70 71 72 74 77
79 80 82 85 86 87 88 90 92 93 94 98 100

classe 3 : 32 elements

4 5 6 7 10 11 12 14 19 20 24 29 32 37 38 39 42 45 47 48 60
61 64 66 69 73 75 76 78 81 89 91

Objets mal classes = (13, 25, 72, 78)

TABLEAU DES VALEURS IDEALES

```
  VVVVVVVVVV
  AAAAAAAAAA
            1
  1234567890
1     1111
2 11    11 1
3 1111    11
```

TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 54.0 | 50.0 | 49.0 | 52.0 | 49.0 | 49.0 | 39.0 | 54.0 | 53.0 | 51.0 |

TABLEAU DES ALPHA

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |

TABLEAU DES ECARTS

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.00 | 0.00 | 3.00 | 2.00 | 0.00 | 1.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| 2 | 0.00 | 0.00 | 2.00 | 2.00 | 3.00 | 1.00 | 3.00 | 2.00 | 1.00 | 0.00 |
| 3 | 1.00 | 1.00 | 0.00 | 0.00 | 2.00 | 2.00 | 0.00 | 0.00 | 1.00 | 0.00 |

TABLEAU DES PARAMETRES DE LA LOI DE BERNOUILLI

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.03 | 0.09 | 0.09 | 0.06 | 0.13 | 0.03 | 0.02 | 0.02 | 0.03 | 0.16 |
| 2 | 0.14 | 0.14 | 0.06 | 0.06 | 0.09 | 0.03 | 0.09 | 0.06 | 0.03 | 0.19 |
| 3 | 0.03 | 0.03 | 0.03 | 0.03 | 0.06 | 0.06 | 0.03 | 0.11 | 0.03 | 0.18 |

TABLEAU DES COEFFICIENTS DE PONDERATION

| | V A 1 | V A 2 | V A 3 | V A 4 | V A 5 | V A 6 | V A 7 | V A 8 | V A 9 | V A 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.5 | 2.3 | 2.3 | 2.8 | 1.9 | 3.5 | 3.7 | 3.7 | 3.5 | 1.6 |
| 2 | 1.8 | 1.8 | 2.8 | 2.8 | 2.3 | 3.5 | 2.3 | 2.8 | 3.5 | 1.4 |
| 3 | 3.4 | 3.4 | 3.6 | 3.5 | 2.7 | 2.7 | 3.6 | 2.1 | 3.4 | 1.5 |

## 4.2 APPLICATIONS AUX DONNEES REELLES

L'ensemble de données est constitué de 59 plaques-boucles sur lesquelles ont été relevées la réponse ou l'absence d'une sélection de 26 critères techniques de fabrication, de forme et de décor. En demandant deux classes, l'algorithme MNDBIN (Marchetti 1989) nous permet de constater une nette différence entre deux grands groupes de plaques-boucles. En effet, la valeur du critère égale à **224** traduit le fait que **84.5%** des données initiales sont égales aux valeurs idéales fournies par les noyaux. Les deux classes sont les suivantes :

**Classe 1** = {01, 02, 06, 08, 10, 12, 13, 16, 18, 19, 29, 37, ..., 45}

**Classe 2** = {03, 04, 05, 07, 09, 11, 14, 15, 17, 20, ..., 28, 30, ..., 36, 46, ..., 59}

Nous énumérons les caractères les plus explicites des classes 1 et 2 :

**Classe 1** : damasquinure par place prédominante (identifiée par C25),
damasquinure bichrome (C23),
fixation par clous de fer (C16),
fond plaqué argent (C37).

**Classe 2** : damasquinure par incrustation dominante (C24),
damasquinure monochrome argent (C22),
fixation par bossette de bronze (C15).

Nous avons détruit 50% de ces données, en les tirant au hasard parmi 59x26 = 1534 observations possibles. La méthode MNDM avec les trois variantes, est ensuite appliquée à ces données incomplètes avec les différents choix du paramètre $\alpha$. Les classes obtenues sont à un élément près identiques (seul l'élément 25 quitte sa classe, et cela représente 1,6% d'objets mal classés). Dans la page qui suit nous présentons les classes obtenues en appliquant la méthode MNDM utilisant la première variante lorsque $\alpha = 1/2$.

```
COMMANDE <> MNDM utilisant une mesure ponderee (ponderation differente par variable)

valeur du critere obtenu    110.0000000


          Partition

  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 INITIAL REORDONNE


     CCCCCCCCCCCCCCCCCCCCCCCCCC
     01111122222223333333333444
     14567923456890123456789012

   03   1 9999199999  9 99  9
   04   919 9999 9199 9   999 99
   05   1 991919 19 9999 9991
   07  99 199 991 99999999 1 911
   09   9 999 19991 999999 1191
   11   9 1 191999 99 9 9  9 9 9
   14   99  9919999 9 99   1999
   15 191   991 9999 99999 9 99
   17   1  91 9 199199  9999 9 9
   20  9999 1 999999 99  9 919 9
   21  9 1 91 99 999 9199  199
   22    1199 19 999 91  9911 99
   23 1 99  9 1  1999 9999 9   9
   24  99 1 1999  199999 9911  9
   25  9991 99 1 999 9 1 99119 9
   26  11 99191  99 999999 11 99
   27   1 9 1 9 19 199  19919999
   28  91  91 19  999 199  11999
   30   99  1 9  9199 1 99 1199
   31  9999919199 999 1 9  11 91
   32   1 9 191 9 19 999 9 99
   33  99 1 1 1999199 999991199
   34   19199 19  99 9999991999
   35   1 999 9 9 9   99  991 9
   36  999  199 9 199 9 9  11 99
   46   19991 19 99 9 19 9 11 99
   47 1 19999 9 91 99   999 9 99
   48  91   1 9 9999 99 9   99 1
   49   1 1 9919999 991  9 9199
   50   1 99191 199 19919999  19
   51   19 99 199999    1  9   9
   52   1 9 9999999999 9  911999
   53   19111 999999999 99911  9
   54 9999 1199 99   9  99  9 99
   55 1 99911999 1999  9 99 9999
   56 1 19911 9  1999 9 9 9 9999
   57 191  9991999  9 9999999 99
   58 1919911 1991 999 999   9 9
   59 9 1  91 9 1991999 999 1999
------------------------------
   01  9 199919199 99999 99 99 9
   02  9 19991 99 9 99 9911  999
   06  9 199 19199991999 9199999
   08 99 99999919  99999991 9 99
   10  1 199 999  99 199 19 99
   12  999 9 1 999  9999 91  911
   13    91 91 199  19999999 119
   16  1 19  1 9 99 9999911  191
   18 9  199 991 99999  991 9119
   19    1 9 1 1999 99 1 99 1 19
   29  9 119 1 1999 1999 999991
   37  9999 91 11999 9 99199
   38  1 99999 9 99999   91   99
   39  191 999919   919 99199  1
   40  999 999 9199  99  11 9  1
   41  9999999 11 99 19  11  9 1
   42    1 9999119999199 19   91
   43 9  99999999  9919  199   1
   44  9 1   199999 99 9  99   1
   45   999  1 1  99999 991 99 1
```

# CHAPITRE III

# DONNEES MANQUANTES LIEN AVEC LE MODELE

## INTRODUCTION

Dans l'algorithme EM, la maximisation de la vraisemblance est remplacée par la maximisation d'une espérance de vraisemblance, plus facile à réaliser. Elle est obtenue en considérant certaines données, telles que l'appartenance des individus aux composants, comme manquantes.

Notre problème consiste à maximiser un critère de vraisemblance classifiante. Dans le chapitre précédent, en présence des données manquantes, nous avons choisi comme critère d'optimisation l'espérance de vraisemblance classifiante. Cette approche bien que présentant des similitudes avec l'algorithme EM, en diffère cependant par l'utilisation dans le calcul de l'espérance de vraisemblance classifiante, d'hypothèses supplémentaires sur les données manquantes. La distribution de ces données ne dépendait pas du modèle et pouvait paraître injustifiée. Dans ce chapitre, nous allons appliquer l'algorithme EM à notre situation, ce qui revient à considérer que les données manquantes suivent le modèle.

Comme précédemment, nous cherchons à classifier des données binaires en présence de données manquantes. Soit le couple $(P, \theta)$ où P désigne la partition $(P_1, ..., P_K)$ avec K supposé connu et $\theta = (a, \varepsilon)$ correspond aux paramètres du modèle associé aux données binaires. Notre objectif, difficile à réaliser par des calculs directs, est de maximiser le critère de vraisemblance classifiante (VC). Ainsi, nous proposons de maximiser la "moyenne" des vraisemblances classifiantes "reconstituantes" $VCR(P, \theta, X)$ pour tous les X possibles. Pour cela, après avoir fixé les paramètres $\theta$ et P, nous proposons de maximiser l'espérance de la vraisemblance classifiante conditionnelle. Rappelons que c'est exactement la position de l'algorithme EM. En respectant les notations précédentes, nous pouvons écrire :

$$Q[(P, \theta), (P^{(t)}, \theta^{(t)})] = E[\text{Log}VCR(P, \theta, X)) / P^{(t)}, \theta^{(t)}] = E[W(P, \theta, X) / P^{(t)}, \theta^{(t)}]$$

Notons toutefois que lors de l'étape de maximisation nous allons chercher uniquement à améliorer le critère et non à le maximiser. Nous reconnaîtrons dans cette étape les deux fonctions, d'affectation et de représentation.

Dans les trois premiers paragraphes, nous proposons une variante de la méthode **MNDM** pour les trois types de modèle de mélanges associés aux données binaires. Dans le quatrième paragraphe, nous décrivons cette méthode à partir du modèle des classes latentes. En utilisant les liens existant entre ce modèle et le modèle associé aux données binaires dans le cas le plus général, nous proposons dans le dernier paragraphe un algorithme de classification avec reconstitution des données manquantes.

Nous conservons les notations utilisées dans les chapitres précédents, sauf indication de notre part.

## 1. ε PARAMETRE FIXE

Pour tout $x_i$ élément de $\Omega$, le modèle s'écrit :

$$f(x_i) = \sum_{k=1}^{K} p_k \, f(x_i \, ; \, a_k)$$

$$\text{avec} \quad \forall\, k = 1, K \quad p_k \in \, ]0, 1[ \quad \text{et} \quad \sum_{k=1}^{K} p_k = 1$$

$$\text{et} \qquad f(x_i \, ; \, a_k) = \prod_{j=1}^{p} \varepsilon^{|x_i^j - a_{k,j}|} \, (1-\varepsilon)^{1-|x_i^j - a_{k,j}|} \qquad (1.1)$$

Nous allons expliciter l'expression de $E[W(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}]$. Les paramètres $\varepsilon^{(t)}$, $a^{(t)}$ désigneront par la suite les valeurs respectives de $\varepsilon$ et a à l'itération (t).

### 1.1 EXPRESSION DE L'ESPERANCE

Sachant $P^{(t)}$, $\varepsilon^{(t)}$ et $a^{(t)}$, nous pouvons énoncer la proposition suivante :

**Proposition 1:**

$$E[W(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] =$$

$$\text{Log}\frac{\varepsilon}{1-\varepsilon} \sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} |x_i^j - a_{k,j}| + \sum_{j \in M_i} |\,|\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}\,| \Big) + np \, \text{Log}(1-\varepsilon)$$

**Preuve :**

D'après (II 2.1.1) nous pouvons écrire :

$$E[W(P, \varepsilon, a, X) / \theta^{(t)}] = \sum_{k=1}^{K} \sum_{i \in P_k} ( \text{Log } f(x_i^o ; a_k) + E[\text{Log } f(x_i^m ; a_k) / \varepsilon^{(t)}, a^{(t)}] )$$

Explicitons le terme $E[\text{Log } f(x_i^m ; a_k) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}]$

$$E[\text{Log } f(x_i^m ; a_k) / P^{(t)}, \varepsilon^{(t)}, a_k^{(t)}] = \sum_{j \in M_i} E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}]$$

Le terme $E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}]$ peut s'écrire :

$$E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, a_{k,j}^{(t)}, \varepsilon^{(t)}] = f(x_i^j = 1 / P^{(t)}, a_{k,j}^{(t)}, \varepsilon^{(t)}) \text{ Log } f(x_i^j = 1 ; a_{k,j}) + f(x_i^j = 0 / P^{(t)}, a_{k,j}^{(t)}, \varepsilon^{(t)}) \text{ Log } f(x_i^j = 0 ; a_{k,j})$$

or $\quad \text{Log } f(x_i^j = 1 ; a_{k,j}) = (1 - a_{k,j}) \text{Log } (\varepsilon) + a_{k,j} \text{Log } (1-\varepsilon)$

et $\quad \text{Log } f(x_i^j = 0 ; a_{k,j}) = a_{k,j} \text{Log } (\varepsilon) + (1-a_{k,j}) \text{Log } (1-\varepsilon).$

En distinguant les deux cas sur $a_{k,j}^{(t)}$ nous avons :

$$\text{Si } a_{k,j}^{(t)} = 1 \quad E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}] = (1-\varepsilon^{(t)}) \text{Log } f(x_i^j = 1 ; a_{k,j}) + \varepsilon^{(t)} \text{ Log } f(x_i^j = 0 ; a_{k,j})$$

$$\text{Si } a_{k,j}^{(t)} = 0 \quad E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}] = \varepsilon^{(t)} \text{Log } f(x_i^j = 1 ; a_{k,j}) + (1-\varepsilon^{(t)}) \text{ Log } f(x_i^j = 0 ; a_{k,j})$$

Pour donner une forme simple de l'expression $E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}]$, nous envisageons les quatre possibilités dépendant des choix de $a_{k,j}$ et de $a_{k,j}^{(t)}$. Ces possibilités sont représentées dans le tableau 1.

| | $a_{k,j}^{(t)} = 1$ | $a_{k,j}^{(t)} = 0$ |
|---|---|---|
| $a_{k,j} = 1$ | $(1- \varepsilon^{(t)})\text{Log } (1-\varepsilon) + \varepsilon^{(t)}\text{Log } (\varepsilon)$ | $\varepsilon^{(t)}\text{Log } (1-\varepsilon) + (1-\varepsilon^{(t)})\text{Log } (\varepsilon)$ |
| $a_{k,j} = 0$ | $(1- \varepsilon^{(t)})\text{Log } (\varepsilon) + \varepsilon^{(t)}\text{Log } (1-\varepsilon)$ | $\varepsilon^{(t)}\text{Log } (\varepsilon) + (1-\varepsilon^{(t)}) \text{Log } (1-\varepsilon)$ |

Tableau 1 : différentes expressions de cette espérance conditionnelle.

Nous pouvons alors écrire :

$$E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}] =$$

$$|\varepsilon^{(t)} - |a_{k,j} - a_{k,j}^{(t)}|| \text{Log}(\varepsilon) + (1 - |\varepsilon^{(t)} - |a_{k,j} - a_{k,j}^{(t)}|) \text{Log}(1-\varepsilon)$$

ou encore

$$E[\text{Log } f(x_i^j ; a_{k,j}) / P^{(t)}, \varepsilon^{(t)}, a_{k,j}^{(t)}] = \text{Log} \frac{\varepsilon}{1-\varepsilon} |\varepsilon^{(t)} - |a_{k,j} - a_{k,j}^{(t)}|| + \text{Log}(1-\varepsilon)$$

Notons que $|\varepsilon^{(t)} - |a_{k,j} - a_{k,j}^{(t)}|| = ||\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}|$

Nous en déduisons la forme générale de $E[\text{Log } f(x_i^m ; a_k) / P^{(t)}, \varepsilon^{(t)}, a_k^{(t)}]$ :

$$E[\text{Log } f(x_i^m ; a_k) / P^{(t)}, \varepsilon^{(t)}, a_k^{(t)}] = \text{Log} \frac{\varepsilon}{1-\varepsilon} \sum_{j \in M_i} ||\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}| + m_i \text{Log}(1-\varepsilon)$$

où $m_i$ désigne le nombre de données manquantes pour l'individu $x_i$.

Nous avons vu dans le deuxième chapitre que le terme $\text{Log } f(x_i^o ; a_k)$ s'écrit :

$$\text{Log } f(x_i^o ; a_k) = \text{Log} \left( \prod_{j \in O_i} \varepsilon^{|x_i^j - a_{k,j}|} . (1-\varepsilon)^{1-|x_i^j - a_{k,j}|} \right)$$

$$= \text{Log} \frac{\varepsilon}{1-\varepsilon} \left( \sum_{j \in O_i} |x_i^j - a_{k,j}| \right) + (p - m_i) \text{Log}(1-\varepsilon)$$

D'où :

$$E[W(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] =$$

$$\text{Log} \frac{\varepsilon}{1-\varepsilon} \sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} |x_i^j - a_{k,j}| + \sum_{j \in M_i} ||\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}| \right) + np \text{ Log}(1-\varepsilon)$$

#

**Remarque :**

*Nous constatons que cette expression s'écrit de telle manière que les données manquantes semblent être reconstituées par le terme $|\varepsilon^{(t)} - a_{k,j}^{(t)}|$ qui est égal à $\varepsilon^{(t)}$ ou à $1-\varepsilon^{(t)}$ suivant que $a_{k,j}^{(t)}$ est égal à 0 ou $a_{k,j}^{(t)}$ est égal à 1 respectivement . Ces probabilités jouent le même rôle que les probabilités conditionnelles $s_k$ déjà vues dans le premier chapitre.*

L'expression E[W(P, ε, a, X)] montre que les recherches de ε et des $a_k$ sont indépendantes. Pour un ε fixé appartenant à ]0, $\frac{1}{2}$[, Log(ε/(1-ε)) est négatif et maximiser l'espérance du critère revient donc à minimiser :

$$C(P, a) = \sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} |x_i^j - a_{k,j}| + \sum_{j \in M_i} ||\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}| \right)$$

Notons que cette expression dépend des valeurs $\varepsilon^{(t)}$ et $a_{k,j}^{(t)}$ qui sont connues. Comme dans le chapitre précédent où la recherche des noyaux dépendait du paramètre α, ici elle dépendra de $\varepsilon^{(t)}$ et $a_{k,j}^{(t)}$.

## 1.2 RECHERCHE DES NOYAUX

**Proposition 2 :**

*Les composantes des $a_k$ éléments de $\{0, 1\}^p$ minimisant C(P, a) sont définies par :*

*pour k = 1, ..., K et j =1, ..., p,*

| *lorsque $a_{k,j}^{(t)} = 1$* | *lorsque $a_{k,j}^{(t)} = 0$* |
|---|---|
| $\begin{cases} a_{kj}=0 \text{ si } m_k^j \leq \zeta \\ a_{k,j}=1 \text{ si } m_k^j \geq \zeta \end{cases}$ | $\begin{cases} a_{k,j}=1 \text{ si } m_k^j \leq -\zeta \\ a_{k,j}=0 \text{ si } m_k^j \geq -\zeta \end{cases}$ |

*où* $\xi = \dfrac{q_k^j - t_k^j}{1 - 2\varepsilon^{(t)}}$

**Preuve :**

Nous nous basons essentiellement sur la proposition 2 du chapitre II en nous limitant au cas où la paramètre α est inférieur à 1/2 ( car $\varepsilon^{(t)}$ est inférieur à 1/2). Dans notre situation $a_{k,j}^{(t)}$ joue un rôle important dans la détermination de $a_{k,j}$. En effet, il existe quatre possibilités qui sont résumées dans le tableau 2, et il s'agit de choisir les $a_{k,j}$ qui appartiennent à {0, 1} et qui minimisent C(P, a).

| | $a_{k,j}^{(t)} = 1$ | $a_{k,j}^{(t)} = 0$ |
|---|---|---|
| $a_{k,j} = 1$ | $q_k^j + m_k^j \varepsilon^{(t)}$ | $q_k^j + m_k^j (1-\varepsilon^{(t)})$ |
| $a_{k,j} = 0$ | $t_k^j + m_k^j (1- \varepsilon^{(t)})$ | $t_k^j + m_k^j \varepsilon^{(t)}$ |

Tableau 2 : différentes valeurs de C(P, a) dans la classe k et pour la variable $x^j$.

A partir des différentes expressions, nous pouvons déterminer de la même manière que dans le chapitre II les valeurs $a_{k,j}$ recherchées. En effet, si nous posons :

$$D_k^j = (t_k^j - q_k^j) + m_k^j (1 - 2\varepsilon^{(t)}) \qquad \text{et} \qquad E_k^j = (t_k^j - q_k^j) + m_k^j (2\varepsilon^{(t)} - 1),$$

les $a_{k,j}$ dépendent des signes de ces deux termes. Par commodité, nous présentons les résultats dans le tableau 3.

| | $a_{k,j}^{(t)} = 1$ | $a_{k,j}^{(t)} = 0$ |
|---|---|---|
| $a_{k,j} = 1$ | $D_k^j \geq 0 \Leftrightarrow m_k^j \geq \xi$ | $E_k^j \geq 0 \Leftrightarrow m_k^j \leq -\xi$ |
| $a_{k,j} = 0$ | $D_k^j \leq 0 \Leftrightarrow m_k^j \leq \xi$ | $E_k^j \leq 0 \Leftrightarrow m_k^j \geq -\xi$ |

Tableau 3 : solution du problème.

### 1.3 METHODE

La méthode **MNDMIN** que nous proposons est une variante de MNDM et est initialisée par $P^{(0)}$ et $(\varepsilon^{(0)}, a^{(0)})$. Les vecteurs $a_k^{(0)}$ sont estimés sur les données observées et l'étape de maximisation est décrite à l'itération (t+1).

**Etape d'estimation :**

Calcul de $Q[(P, \theta), (P^{(t)}, \theta^{(t)})]$ (voir proposition 1).

**Etape de maximisation :**

*Recherche des $a_{k,j}^{(t+1)}$: (recherche des noyaux)*

Les $a_{k,j}^{(t+1)}$ qui maximisent $Q[(P^{(t+1)}, \theta^{(t+1)}), (P^{(t)}, \theta^{(t)})]$ sont définies dans la proposition 2.

En posant $e^{(t+1)} = \sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} | x_i^j - a_{k,j}| + \sum_{j \in M_i} | |\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j} | \right)$, $\varepsilon^{(t+1)}$ qui maximise l'espérance conditionnelle est défini par : $\varepsilon^{(t+1)} = e^{(t+1)}/np$.

*Recherche de la partition $P^{(t+1)}$ : (recherche des classes)*

Nous affecterons $x_i$ à la classe $P_k^{(t+1)}$ qui minimise :

$$\sum_{j \in O_i} | x_i^j - a_{k,j}^{(t+1)}| + \sum_{j \in M_i} | |\varepsilon^{(t)} - a_{k,j}^{(t)}| - a_{k,j}^{(t+1)}|$$

## 2. ε PARAMETRE DEPENDANT DE CHAQUE VARIABLE

Pour tout $x_i$ élément de $\Omega$, le modèle s'écrit :

$$f(x_i) = \sum_{k=1}^{K} p_k \, f(x_i ; a_k)$$

$$\text{avec} \quad \forall k = 1, K \quad p_k \in ]0, 1[ \quad \text{et} \quad \sum_{k=1}^{K} p_k = 1$$

$$\text{et} \qquad f(x_i ; a_k) = \prod_{j=1}^{p} \varepsilon_j^{|x_i^j - a_{k,j}|} \, (1-\varepsilon_j)^{1-|x_i^j - a_{k,j}|} \qquad (2.1)$$

Dans ce qui suit, nous nous passerons des démonstrations qui se déduisent immédiatement à partir du premier paragraphe, nous donnerons uniquement les résultats. Nous notons $W_1$ la vraisemblance classifiante reconstituante.

### 2.1 EXPRESSION DE L'ESPERANCE

**Proposition 3 :**

$$E[W_1(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] =$$
$$\sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} \text{Log} \frac{\varepsilon_j}{1-\varepsilon_j} | x_i^j - a_{k,j}| + \sum_{j \in M_i} \text{Log} \frac{\varepsilon_j}{1-\varepsilon_j} | |\varepsilon_j^{(t)} - a_{k,j}^{(t)}| - a_{k,j} | \right) + n \sum_{j \in J} \text{Log}(1-\varepsilon_j)$$

**Preuve :**

La démonstration est identique à celle de la proposition 1.#

Cette expression peut également s'écrire :

$$E[W_1(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] = - \sum_{k=1}^{K} \sum_{i \in P_k} d_\varepsilon(x_i, a_k) + A \qquad (2.1.1)$$

où

$$d_\varepsilon(x_i, a_k) = \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | x_i^j - a_{k,j} | + \sum_{j \in M_i} \text{Log} \frac{1-\varepsilon^j}{\varepsilon^j} | | \varepsilon_j^{(t)} - a_{k,j}^{(t)} | - a_{k,j} |$$

et $A = n \sum_{j \in J} \text{Log}(1-\varepsilon^j)$.

## 2.2 METHODE

**Etape d'estimation :**
L'expression $Q[(P, \theta), (P^{(t)}, \theta^{(t)})]$ est définie dans la proposition 3.

**Etape de maximisation :**
*Recherche des $a_{k,j}^{(t+1)}$: (recherche des noyaux)*

Les $a_{k,j}^{(t+1)}$ sont définies dans la proposition 2. Il ne reste plus qu'à déterminer les $\varepsilon_j^{(t+1)}$ maximisant (2.1.1). La quantité $Q[(P^{(t+1)}, \theta^{(t+1)}), (P^{(t)}, \theta^{(t)})]$ peut s'écrire :

$$Q[(P^{(t+1)}, \theta^{(t+1)}), (P^{(t)}, \theta^{(t)})] = \sum_{j \in O_i} (\text{Log} \frac{\varepsilon_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}}) e_j + \sum_{j \in M_i} (\text{Log} \frac{\varepsilon_j^{(t+1)}}{1-\varepsilon_j^{(t+1)}}) \beta_j + n \sum_{j \in J} \text{Log}(1-\varepsilon_j^{(t+1)})$$

$$= \sum_{j \in J} F(\varepsilon_j^{(t+1)}) \qquad (2.2.1)$$

où

$$e_j = \sum_{k=1}^{K} \sum_{i \in P_k} | x_i^j - a_{k,j}^{(t+1)} | \qquad \text{pour } j \in O_i.$$

$$\beta_j = \sum_{k=1}^{K} \sum_{i \in P_k} | | \varepsilon_j^{(t)} - a_{k,j}^{(t)} | - a_{k,j}^{(t+1)} | \qquad \text{pour } j \in M_i.$$

Les valeurs $(\varepsilon_j^{(t+1)}, j \in J)$ qui maximisent (2.2.1) sont définies par :

$$\frac{\partial F}{\partial \varepsilon_j^{(t+1)}} = 0 \quad \Rightarrow \quad \frac{e_j + \beta_j}{\varepsilon_j^{(t+1)}} - \frac{n - e_j - \beta_j}{1-\varepsilon_j^{(t+1)}} = 0$$

$$\Rightarrow \qquad \varepsilon_j^{(t+1)} = \frac{e_j + \beta_j}{n}$$

En distinguant les différents cas rencontrés, nous pouvons montrer que les $\varepsilon_j^{(t+1)}$ appartiennent à ]0, 1/2[.

*Recherche de la partition $P^{(t+1)}$ : (recherche des classes)*

Puisque nous cherchons à maximiser $Q[(P^{(t+1)}, \theta^{(t+1)}), (P^{(t)}, \theta^{(t)})]$, nous affecterons $x_i$ à la classe $P_k^{(t+1)}$ qui minimise :

$$d_\varepsilon(x_i, a_k).$$

## 3. ε PARAMETRE DEPENDANT DE CHAQUE CLASSE ET DE CHAQUE VARIABLE

Pour tout $x_i$ élément de $\Omega$, le modèle s'écrit :

$$f(x_i) = \sum_{k=1}^{K} p_k \, f(x_i \, ; \, a_k)$$

$$\text{avec} \quad \forall \, k = 1, K \quad p_k \in \, ]0, 1[ \quad \text{et} \quad \sum_{k=1}^{K} p_k = 1$$

$$\text{et} \qquad f(x_i \, ; \, a_k) = \prod_{j=1}^{p} \varepsilon_{k,j}^{\,|x_i^j - a_{k,j}|} \, (1-\varepsilon_{k,j})^{1-|x_i^j - a_{k,j}|} \qquad (3.1)$$

Comme précédemment, nous nous baserons sur les résultats développés dans le premier paragraphe. Nous notons $W_2$ la vraisemblance classifiante reconstituante.

### 3.1 EXPRESSION DE L'ESPERANCE

**Proposition 4 :**

$$E[W_2(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] =$$

$$\sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} \text{Log} \frac{\varepsilon_{k,j}}{1-\varepsilon_{k,j}} |x_i^j - a_{k,j}| + \sum_{j \in M_i} \text{Log} \frac{\varepsilon_{k,j}}{1-\varepsilon_{k,j}} |\,|\varepsilon_{k,j}^{(t)} - a_{k,j}^{(t)}| - a_{k,j}| + \sum_{j \in J} \text{Log}(1-\varepsilon_{k,j}) \Big)$$

Cette expression peut également s'écrire :

$$E[W_2(P, \varepsilon, a, X) / P^{(t)}, \varepsilon^{(t)}, a^{(t)}] = \sum_{k=1}^{K} \sum_{i \in P_k} \{-d_{\varepsilon_K}(x_i, a_k) + A_k\} \qquad (3.1.1)$$

où

$$d_{\varepsilon_k}(x_i, a_k) = \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon_{k,j}}{\varepsilon_{k,j}} |x_i^j - a_{k,j}| + \sum_{j \in M_i} \text{Log} \frac{1-\varepsilon_{k,j}}{\varepsilon_{k,j}} ||\varepsilon_{k,j}^{(t)} - a_{k,j}^{(t)}| - a_{k,j}|$$

et $\quad A_k = \sum_{j \in J} \text{Log}(1-\varepsilon_{k,j}).$

## 3.2 METHODE

**Etape d'estimation :**
l'expression de $Q[(P, \theta), (P^{(t)}, \theta^{(t)})]$ est définie dans la proposition 4.

**Etape de maximisation :**
Les $a_{k,j}^{(t+1)}$ sont définies dans la proposition 2. Il ne reste plus qu'à déterminer les $\varepsilon_{k,j}^{(t+1)}$ qui maximisent la quantité C suivante :

$$C = \sum_{i \in P_k} \left\{ \text{Log} \frac{1-\varepsilon_{k,j}^{(t+1)}}{\varepsilon_{k,j}^{(t+1)}} |x_i^j - a_{k,j}^{(t+1)}| + \text{Log} \frac{1-\varepsilon_{k,j}^{(t+1)}}{\varepsilon_{k,j}^{(t+1)}} ||\varepsilon_{k,j}^{(t)} - a_{k,j}^{(t)}| - a_{k,j}^{(t+1)}| + \text{Log}(1-\varepsilon_{k,j}^{(t+1)}) \right\}$$

$$= (\text{Log} \frac{\varepsilon_{k,j}^{(t+1)}}{1-\varepsilon_{k,j}^{(t+1)}}) e_k^j + (\text{Log} \frac{\varepsilon_{k,j}^{(t+1)}}{1-\varepsilon_{k,j}^{(t+1)}}) \beta_k^j + n_k \text{Log}(1-\varepsilon_{k,j}^{(t+1)})$$

$$= G(\varepsilon_{k,j}^{(t+1)}) \qquad (3.2.1)$$

où

$$e_k^j = \sum_{i \in P_k} |x_i^j - a_{k,j}^{(t+1)}| \qquad \text{pour } k = 1, K \text{ et } j \in O_i$$

$$\beta_k^j = \sum_{i \in P_k} ||\varepsilon_{k,j}^{(t)} - a_{k,j}^{(t)}| - a_{k,j}^{(t+1)}| \qquad \text{pour } k = 1, K \text{ et } j \in M_i$$

Les composantes $(\varepsilon_{k,j}^{(t+1)}, k \in \{1, ..., K\}, j \in J)$ maximisant (3.2.1) sont définies par :

$$\frac{\partial G}{\partial \varepsilon_{k,j}^{(t+1)}} = 0 \quad \Rightarrow \quad \varepsilon_{k,j}^{(t+1)} = \frac{e_k^j + \beta_k^j}{n_k}$$

En distinguant les différents cas, nous pouvons montrer que les $\varepsilon_{k,j}^{(t+1)}$ appartiennent à ]0, 1/2[. Les classes sont définies en affectant chaque individu $x_i$ à la classe $P_k^{(t+1)}$ qui minimise :

$$d_{\varepsilon_k}(x_i, a_k).$$

**Remarque :**

*Nous avons rappelé dans le chapitre I les liens étroits existant entre ce modèle et celui des classes latentes. Notons toutefois que le modèle des classes latentes est plus aisé à utiliser car, dans les calculs, la manipulation d'un unique paramètre s'avère plus facile. Nous pouvons ainsi donner une analyse du "comportement" des cases vides correspondant aux données manquantes.*

## 4. MODELE DES CLASSES LATENTES

Dans le premier chapitre, nous avons rappelé le modèle des classes latentes. Dans notre cas, $\theta$ joue à la fois le rôle de $\varepsilon$ et celui de a, nous le notons $\alpha$. D'après ce qui précède, nous pouvons décomposer $Q[(P, \alpha), (P^{(t)}, \alpha^{(t)})]$ en :

$$\sum_{k=1}^{K} \sum_{i \in P_k} ( \text{Log } f(x_i^o ; \alpha_k) + E[\text{Log } f(x_i^m ; a_k / P^{(t)}, \alpha^{(t)}] )$$

Explicitons maintenant les deux termes de cette expression. Le premier s'écrit pour tout k appartenant à {1, ..., K} :

$$\text{Log } f(x_i^o ; \alpha_k) = \sum_{j \in O_i} \{ x_i^j \text{ Log } \alpha_{k,j} + (1-x_i^j) \text{ Log } (1- \alpha_{k,j}) \}$$

et le deuxième terme s'écrit pour tout k appartenant à {1, ..., K} :

$$\begin{aligned} &E[\text{Log } f(x_i^m ; \alpha_k) / P^{(t)}, \alpha^{(t)}] \\ &= \sum_{i \in P_k} \sum_{j \in M_i} E[ f(x_i^j ; \alpha_{k,j}) / P^{(t)}, \alpha_{k,j}^{(t)}] \\ &= \sum_{i \in P_k} \sum_{j \in M_i} \{ f(x_i^j = 1 / P^{(t)}, \alpha_{k,j}^{(t)}) \text{ Log } f(x_i^j = 1 ; \alpha_{k,j}) \\ &\qquad + f(x_i^j = 0 / P^{(t)}, \alpha_{k,j}^{(t)}) \text{ Log } f(x^j = 0 ; \alpha_{k,j}) \} \\ &= \sum_{i \in P_k} \sum_{j \in M_i} \{ \alpha_{k,j}^{(t)} \text{ Log } \alpha_{k,j} + (1-\alpha_{k,j}^{(t)}) \text{ Log } (1- \alpha_{k,j}) \} \end{aligned}$$

L'expression de l'espérance du critère de vraisemblance classifiante conditionnelle est la suivante pour $0 < \alpha_{k,j} < 1$ :

$$E[W(P, X, \alpha) / P^{(t)}, \alpha^{(t)}] =$$

$$\sum_{k=1}^{K} \{ \sum_{i \in P_k} \sum_{j \in O_i} x_i^j \text{Log} \frac{(\alpha_{k,j})}{(1-\alpha_{k,j})} + \sum_{j \in M_i} \alpha_{k,j}^{(t)} \text{ Log} \frac{(\alpha_{k,j})}{(1-\alpha_{k,j})} + \sum_{j \in J} \text{Log } (1-\alpha_{k,j}) \} \quad (4.1)$$

Nous allons chercher à calculer les ($\alpha_k$ ; k = 1, .., K) qui maximisent (4.1).

**proposition 5 :**

*Les composantes de $\alpha_k$ élément de $R^p$ maximisant $E[W(P, X, \alpha) / P^{(t)}, \alpha^{(t)}]$ sont définies par : pour tout k = 1, ..., K et j =1, ..., p*

$$\alpha_{k,j} = \frac{t_k^j + \alpha_{k,j}^{(t)} m_k^j}{n_k}$$

**Preuve :**

Maximiser $E[W(P, X, \alpha) / \alpha^{(t)}]$ revient à minimiser chacun des termes $L_{k,j}$ :

$$L_{k,j} = (t_k^j + \alpha_{k,j}^{(t)} m_k^j) \operatorname{Log} \frac{(\alpha_{k,j})}{(1-\alpha_{k,j})} + n_k \operatorname{Log} (1- \alpha_{k,j}).$$

Pour tout k = 1, ..., K et j =1, ..., p nous remarquons que $L_{k,j} \leq 0$ car $0 < \alpha_{k,j} < 1$.

Nous avons :

$$\frac{\partial L_{k,j}}{\partial \alpha_{k,j}} = \frac{t_k^j + \alpha_{k,j}^{(t)} m_k^j}{\alpha_{k,j}} + \frac{t_k^j + \alpha_{k,j}^{(t)} m_k^j - n_k}{1- \alpha_{k,j}} = \frac{t_k^j + \alpha_{k,j}^{(t)} m_k^j - \alpha_{k,j} n_k}{\alpha_{k,j}(1- \alpha_{k,j})}$$

Les $\alpha_{k,j}$ qui maximisent $L_{k,j}$ sont alors définies par :

$$\frac{\partial L_{k,j}}{\partial \alpha_{k,j}} = 0 \Rightarrow \alpha_{k,j} = \frac{t_k^j + \alpha_{k,j}^{(t)} m_k^j}{n_k}$$

Lorsque $(t_k^j + \alpha_{k,j}^{(t)} m_k^j)$ tend vers 0 ou $t_k^j$ tend vers $n_k$, le terme $L_{k,j}$ tend vers 0. La solution du problème est donc toujours celle que nous venons de donner. #

**Remarques :**

1. *Lors de l'étape de maximisation, tout se passe comme si les données manquantes de chaque variable étaient reconstituées par le paramètre correspondant et estimé*

*auparavant. En effet, pour k =1, ..., K et j =1, ..., p, les $\alpha_{k,j}$ désignent les centres de gravité calculés sur les valeurs observées et les données manquantes reconstituées par le paramètre $\alpha_{k,j}^{(t)}$ de chaque classe pour chaque variable.*

2. *Lorsque $n_k = m_k^j$, nous avons $t_k^j = 0$ et $\alpha_{k,j} = \alpha_{k,j}^{(t)}$.*

### 4.1 METHODE

Elle est initialisée par une partition initiale $P^{(0)}$ et les centres de gravité $\{(\alpha_{k,1}^{(0)}, ..., \alpha_{k,p}^{(0)}) ; k = 1, ..., K\}$ calculés sur les valeurs observées. L'étape de maximisation est décrite à l'itération (t+1).

**Etape d'estimation :**

Lors de cette étape, nous calculons $Q[(P, \alpha), (P^{(t)}, \alpha^{(t)})]$.

**Etape de maximisation :**

Nous associons à chaque classe $P_k^{(t)}$ le vecteur $\alpha_k^{(t+1)}$ défini dans la proposition 1.

Lors de cette étape, l'individu $x_i$ est affecté à la classe $P_k^{(t+1)}$ dont le noyau est le plus proche. Les classes sont alors définies de la façon suivante :

$\forall\ k = 1, ..., K$

$$P_k^{(t+1)} = \{ x_i / \sum_{j \in O_i} x_i^j \operatorname{Log} \frac{(\alpha_{k,j}^{(t+1)})}{(1-\alpha_{k,j}^{(t+1)})} + \sum_{j \in M_i} \alpha_{k,j}^{(t)} \operatorname{Log} \frac{(\alpha_{k,j}^{(t+1)})}{(1-\alpha_{k,j}^{(t+1)})} + \sum_{j \in J} \operatorname{Log} (1- \alpha_{k,j}) \geq$$

$$\sum_{j \in O_i} x_i^j \operatorname{Log} \frac{(\alpha_{m,j}^{(t+1)})}{(1-\alpha_{m,j}^{(t+1)})} + \sum_{j \in M_i} \alpha_{m,j}^{(t)} \operatorname{Log} \frac{(\alpha_{m,j}^{(t+1)})}{(1-\alpha_{m,j}^{(t+1)})} + \sum_{j \in J} \operatorname{Log} (1- \alpha_{m,j}^{(t+1)}) \quad \forall\ m \neq k \}$$

**Remarque :**

*Les liens existant entre ce modèle et le modèle utilisé dans le troisième paragraphe sont exactement ceux déjà vus dans le cas où toutes les données sont observées. Néanmoins, ils dépendent de deux itérations successives.*

## 4.2 EXEMPLE SIMPLE D'APPLICATION

Nous disposons d'un tableau de données binaires croisant 10 individus identifiés par les lettres **a** à **j** et 10 variables binaires identifiées par les nombres **1** à **10** (tableau 4).

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

Tableau 4 : tableau de données.

Nous appliquons notre méthode MNDMIN en demandant deux classes. Dans les tableaux 5 et 6, nous représentons respectivement le tableau initial réordonné et le tableau des noyaux.

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

Tableau 5 : meilleure partition obtenue.

La valeur du critère obtenu est **-9.46**. La partition obtenue est la même que celle obtenue sur le tableau initial avant la génération des données manquantes.

| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | 0.60 | 0.20 | 0.20 | 0.40 | 1.00 | 0.90 | 0.80 | 0.60 | 0.20 | 0.40 |
| $a_2$ | 0.87 | 0.80 | 0.80 | 0.80 | 0.40 | 0.80 | 0.40 | 0.40 | 0.60 | 0.55 |

Tableau 6 : tableau des noyaux.

## 5. EXPERIENCES NUMERIQUES

Nous appliquons notre méthode MNDMIN sur les données simulées utilisées dans le premier chapitre en demandant 3 classes. Les résultats sont présentés dans le tableau 7.

| % de destruction | % de mal classés | objets mal classés |
|---|---|---|
| 0 | 0 | Ø |
| 8 | 0 | Ø |
| 12 | 0 | Ø |
| 16 | 1 | {77} |
| 20 | 3 | {13, 77, 78} |
| 24 | 3 | {13, 77, 78} |
| 28 | 4 | {13, 73, 77, 78} |
| 32 | 4 | {13, 73, 77, 78} |
| 36 | 4 | {18, 38, 77, 78} |
| 40 | 4 | {72, 73, 77, 78} |
| 44 | 4 | {72, 73, 77, 78} |
| 48 | 4 | {13, 28, 72, 78} |
| 50 | 5 | {13, 25, 43, 73, 78} |
| 52 | 6 | {13, 43, 73, 78, 82, 95} |
| 56 | 5 | {43, 72, 78, 82, 95} |
| 64 | 12 | {18, 21, 25, 43, 71, 73, 78, 79, 81, 82, 91, 95} |

Tableau 7 : évolution des pourcentages d'objets mal classés.

La méthode MNDMIN donne de bons résultats. Notons toutefois la lenteur de la convergence de l'algorithme. Nous présentons ci-dessous le tableau de données XER56 comportant 56 % de données manquantes sur lequel nous appliquons notre méthode.

```
 1 0009111190    26 1100909109    51 0900191199    76 9199090991
 2 9090111990    27 0909919110    52 0909911990    77 9110999999
 3 9099919109    28 1199000909    53 0009919190    78 9199900991
 4 1199000919    29 1111100991    54 1901999199    79 9999091909
 5 9999090999    30 1109191191    55 0900991190    80 9100191101
 6 1111999011    31 9091991190    56 1909901909    81 9191190999
 7 1111019091    32 9199910911    57 9091919900    82 9990991101
 8 9900191191    33 9000111900    58 9900009109    83 9990191190
 9 9009111900    34 9000111100    59 0900999990    84 0909111999
10 1919990019    35 9990111199    60 9919909919    85 1190091991
11 9999990019    36 9919191109    61 9119909919    86 1999901109
12 1919000011    37 1991099911    62 9119901999    87 1100099991
13 9190990991    38 9919909999    63 9999999199    88 9109999901
14 9199000091    39 0119900991    64 9911009099    89 9911990999
15 9990991009    40 0009909909    65 9909991190    90 1190091199
16 9990991990    41 9000119999    66 1119999019    91 9991109091
17 9099199900    42 9119909911    67 0900191909    92 1900101991
18 1990090191    43 9199909909    68 1100901999    93 9909001101
19 1191000919    44 0010191900    69 9191000991    94 1990099191
20 9919999099    45 1099099919    70 9900091119    95 1909111909
21 1901991999    46 1199009101    71 9190999109    96 9909919909
22 0010199109    47 9911999099    72 9199019199    97 9000191100
23 0999111909    48 9919990911    73 9199900009    98 9909909909
24 1119009099    49 1100001001    74 1990901999    99 0999191190
25 9190999109    50 0999191190    75 9199090911   100 1990901901
```

Nous appliquons la méthode MNDMIN en demandant 3 classes :

**Commande <> MNDMIN methode de classification ( modele des classes latentes)**

```
                          ANALYSE DU MEILLEUR TIRAGE

valeur du critere obtenu   -183.8517000


          Partition

  classe  1 :    31 elements
  -----------
8   15   18   26   28   30   46   49   54   56   58   62   68   70   71   72   74
77   79   80   85   86   87   88   90   92   93   94   95   98  100

  classe  2 :    34 elements
  -----------
4    5    6    7   10   11   12   13   14   19   20   24   29   32   37   38   39
42   43   45   47   48   60   61   64   66   69   73   75   76   78   81   89   91

  classe  3 :    35 elements
  -----------
1    2    3    9   16   17   21   22   23   25   27   31   33   34   35   36   40
41   44   50   51   52   53   55   57   59   63   65   67   82   83   84   96   97
99


Objets mal classes = (43, 72, 78, 82, 95)

          TABLEAU DES NOYAUX A L'ETAPE "CONVERGENCE -1"

             V      V      V      V      V      V      V      V      V      V
             A      A      A      A      A      A      A      A      A      A
                                                                        1
             1      2      3      4      5      6      7      8      9      0
   1      0.07   0.00   0.13   0.25   1.00   0.93   1.00   1.00   0.06   0.00
   2      1.00   1.00   0.13   0.00   0.21   0.07   0.94   0.86   0.08   1.00
   3      0.91   0.94   1.00   0.91   0.14   0.13   0.00   0.00   0.91   1.00


          TABLEAU DES VALEURS IDEALES A LA CONVERGENCE

             V      V      V      V      V      V      V      V      V      V
             A      A      A      A      A      A      A      A      A      A
                                                                        1
             1      2      3      4      5      6      7      8      9      0
   1      0.61   0.42   0.12   0.15   0.65   0.51   0.97   0.94   0.07   0.32
   2      0.97   0.97   0.43   0.29   0.18   0.10   0.44   0.61   0.35   1.00
   3      0.57   0.59   0.66   0.61   0.53   0.45   0.57   0.40   0.52   0.46


          TABLEAU DES EFFECTIFS DES DONNEES MANQUANTES

             V      V      V      V      V      V      V      V      V      V
             A      A      A      A      A      A      A      A      A      A
                                                                        1
             1      2      3      4      5      6      7      8      9      0
   1      58.0   56.0   57.0   57.0   56.0   56.0   44.0   61.0   60.0   55.0
```

**Nous appliquons notre méthode sur le tableau de données "méro" qui a subi une destruction de 50%. Seul l'individu 25 a quitté sa classe. Les résultats figurent dans la page suivante.**

**Commande <> MNDMIN methode de classification ( modele des classes latentes)**

ANALYSE DU MEILLEUR TIRAGE

valeur du critère obtenu -464.0053000

Partition

classe 1 : 21 elements

01 02 06 08 10 12 13 16 18 19 25 29 37 38 39 40 41 42 43 44 45

classe 2 : 38 elements

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

TABLEAU INITIAL REORDONNE

```
     CCCCCCCCCCCCCCCCCCCCCCCCCC
     01111122222223333333333444
     14567923456890123456789012

  01  9 199919199 99999 99 99 9
  02  9 19991 99 9 99 9911  999
  06  9 199 19199991999 9199999
  08 99 99999919  99999991 9 99
  10  1 199 999  99 199 19 99
  12  999 9 1 999  9999 91  911
  13    91 91 199  19999999 119
  16  1 19  1 9 99 9999911  191
  18 9  199 991 99999  991 9119
  19    1 9 1 1999 99 1 99 1 19
  25  9991 99 1 999 9 1 99119 9
  29  9 119 1 1999 1999 999991
  37  9999 91 11999 9 99199
  38  1 99999 9 99999   91   99
  39  191 999919   919 99199  1
  40  999 999 9199  99  11 9  1
  41  9999999 11 99 19  11  9 1
  42    1 9999119999199 19   91
  43 9  99999999  9919  199   1
  44  9 1   199999 99 9  99   1
  45   999  1 1  99999 991 99 1
 ------------------------------
  03   1 9999199999  9 99  9
  04   919 9999 9199 9   999 99
  05   1 991919 19 9999 9991
  07  99 199 991 99999999 1 911
  09   9 999 19991 999999 1191
  11   9 1 191999 99 9 9  9 9 9
  14   99  9919999 9 99   1999
  15 191   991 9999 99999 9 99
  17   1  91 9 199199  9999 9 9
  20  9999 1 999999 99  9 919 9
  21  9 1 91 99 999 9199  199
  22    1199 19 999 91  9911 99
  23 1 99  9 1  1999 9999 9   9
  24  99 1 1999  199999 9911  9
  26  11 99191  99 999999 11 99
  27   1 9 1 9 19 199  19919999
  28  91  91 19  999 199  11999
  30   99  1 9  9199 1 99 1199
  31  9999919199 999 1 9  11 91
  32   1 9 191 9 19 999 9 99
  33  99 1 1 1999199 999991199
  34   19199 19  99 9999991999
  35   1 999 9 9 9   99  991 9
  36  999  199 9 199 9 9  11 99
  46   19991 19 99 9 19 9 11 99
  47 1 19999 9 91 99   999 9 99
  48  91   1 9 9999 99 9   99 1
  49   1 1 9919999 991  9 9199
  50   1 99191 199 19919999  19
  51   19 99 199999    1 9   9
  52   1 9 9999999999 9  911999
  53   19111 999999999 99911  9
  54 9999 1199 99   9  99  9 99
  55 1 99911999 1999  9 99 9999
  56 1 19911 9  1999 9 9 9 9999
  57 191  9991999  9 9999999 99
  58 1919911 1991 999 999   9 9
  59 9 1  91 9 1991999 999 1999
```

# 6. METHODE DE CLASSIFICATION AVEC RECONSTITUTION

## 6.1 INTRODUCTION

Dans le paragraphe précédent, nous avons noté que le processus est tel que les données manquantes sont reconstituées à chaque itération. Par analogie et en utilisant les liens étroits qui existent entre le modèle associé aux données binaires dans le cas le plus général (les paramètres du modèle dépendent à la fois des classes et des variables) et le modèle des classes latentes (voir chapitre I), nous proposons la méthode **MNDMRE** (MNDM + REconstitution). A partir d'une reconstitution initiale, nous disposons d'un tableau complété. Ainsi, le critère à maximiser est celui de la vraisemblance classifiante dans le cas où toutes les valeurs de l'échantillon sont observées. Ce critère s'écrit :

$$W(P, a, \varepsilon) = \sum_{k=1}^{K} \sum_{i \in P_k} \{ - \sum_{j \in J} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j | + \sum_{j \in J} \text{Log}(1-\varepsilon_k^j) \}$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \{ -d_{\varepsilon_k}(x_i, a_k) + A_k \} \qquad (6.1.1)$$

$$\text{où} \quad d_{\varepsilon_k}(x_i, a_k) = \sum_{j \in J} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j | \quad \text{et} \quad A_k = \sum_{j \in J} \text{Log}(1-\varepsilon_k^j).$$

La méthode MNDMRE est de type Nuées Dynamiques. Elle contient une étape de reconstitution qui est intercalée entre les étapes d'affectation et de représentation.

## 6.2 METHODE

Les données manquantes sont initialement reconstituées par les valeurs majoritaires estimées sur les données observées de chaque classe pour chaque variable.

**Etape de représentation** :*(recherche des $a_k^j$ et $\varepsilon_k^j$)*

Quelles que soient les valeurs $\varepsilon_k^j$, les $a_k^j$ sont nécessairement les valeurs majoritaires de chaque classe pour chaque variable. Par ailleurs, les $(\varepsilon_k^j, k \in \{1, \ldots, K\}, j \in J)$ maximisant (6.1.1) sont les valeurs $\frac{e_k^j}{n_k}$ où $e_k^j = \sum_{i \in P_k} | x_i^j - a_k^j |$.

**Etape d'affectation** :*(recherche des classes)*

Lors de cette étape, le terme $A_k$ n'est pas constant. Nous affectons $x_i$ à la classe $P_k$ qui minimise $d_{\varepsilon_K}(x_i, a_k) - A_k$.

**Etape de reconstitution** :*(reconstitution des données manquantes)*

Le but de cette étape est de reconstituer les données manquantes (par 0 ou 1) maximisant le critère. Il est clair qu'une donnée non observée pour une variable $x^j$ de l'individu $x_i$ et se trouvant dans la classe k sera reconstituée par $a_k^j$. En effet, le critère W(P, a, ε) peut s'écrire :

$$W(P, a, \varepsilon) = \sum_{k=1}^{K} \sum_{i \in P_k} \{ - \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j | - \sum_{j \in M_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j | + \sum_{j \in J} \text{Log}(1-\varepsilon_k^j) \},$$

et les données manquantes qui maximisent W(P, a, ε) sont celles qui annulent :

$$\sum_{j \in M_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} | x_i^j - a_k^j | .$$

Ainsi la décroissance du critère augmente. Cette convergence se démontre de la façon suivante :

Notons **DM** l'ensemble des données manquantes. A l'itération (t), nous avons :

$$W(P^t, L^t, DM^t) \leq W(P^t, L^t, DM^{t-1}).$$

et, par suite, à l'itération (t+1) nous avons :

$$W(P^{t+1}, L^{t+1}, DM^t) \leq W(P^t, L^t, DM^t)$$

d'où la convergence de l'algorithme.

## 7. APPLICATIONS DE LA METHODE MNDMRE

Dans ce paragraphe, nous étudions les deux tâches de cette méthode : classification et reconstitution.

## 7.1 CLASSIFICATION

Nous appliquons notre méthode MNDMRE sur les données simulées utilisées dans les chapitres précédents en demandant trois classes. Le tableau 8 présente le Taux de destruction "T" dans la première ligne et le Pourcentage d'objets mal classés "P" dans la seconde.

| T en % | 0 | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 | 44 | 48 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P en % | 0 | 0 | 1 | 3 | 6 | 6 | 7 | 8 | 8 | 11 | 16 | 27 |

Tableau 8 : pourcentages d'objets mal classés.

Nous présentons dans la figure 1 l'évolution de ces pourcentages.

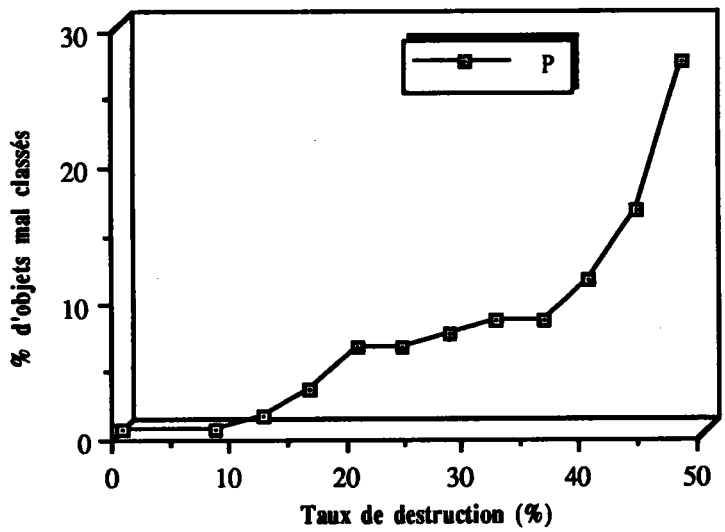



Figure 1 : évolution des pourcentages d'objets mal classés.

Nous constatons que cette méthode donne des résultats assez satisfaisants bien qu'elle accomplisse deux rôles : celui de la classification et celui de la reconstitution.

Sur le tableau XER36 présenté dans la page suivante, (Tableau binaire comportant 36% de données manquantes codées '?') nous appliquons la méthode MNDMRE en demandant trois classes.

```
  1 0000111100      26 1100?0?10?      51 0?001?1100      76 ?1?00?00?1
  2 ?0?0111??0      27 0?0?11?110      52 0?0??11??0      77 ?110??????
  3 000?11?10?      28 11??00010?      53 000??1?1?0      78 ?1????00??1
  4 11?1000?1?      29 1111100??1      54 11010??1??      79 ??0?001?0?
  5 ??1?0?0???      30 110?1?11?1      55 0?001?11?0      80 11001?1101
  6 1111???011      31 ?0?1??11?0      56 110?001?0?      81 ?1?11?0???
  7 111101?0?1      32 ?11?010?11      57 ?0?1?11?00      82 ??00??1101
  8 1?001?1101      33 ?000111?00      58 ??0000?10?      83 ???01?11?0
  9 ?00?111?00      34 ?000111100      59 0000111?0       84 0?00111???
 10 1111??001?      35 0??01111??      60 1?1??00?11      85 11?00?1?01
 11 1??0??001?      36 0?1?1?1100      61 ?11??0??1?      86 11???01107
 12 1?1?000011      37 1??100?011      62 ?11?0011??      87 11000??1?1
 13 ?1?0??0??1      38 ??1??0????0     63 ????????1??     88 ?100????01
 14 ?1??0000?1      39 011??00??1      64 ??1100?0?1      89 ??11??0???
 15 ????00?1001     40 0000?0?10?      65 0?0?1?11?0      90 11?00?11??
 16 00?0?11??0      41 0000111107      66 111?0??01?      91 1?1110?0?1
 17 ?0??11?100      42 ?11??01011      67 00001?1?00      92 1000101111
 18 1??00?0111      43 ?1????0??01     68 1100001???      93 1?0?001101
 19 1111000?1?      44 00101?1100      69 11?1000?1       94 11?00??1?1
 20 1?11???01?      45 10??0???1?      70 ?1000?111?      95 1?0?111?0?
 21 1?01?011??      46 110?00?101      71 ?1?0?0?100      96 ?00??1??0?
 22 001011?10?      47 ??11?000?1      72 111101?1?1      97 00001?1100
 23 0?00111?00      48 ??1????0?11     73 11???0000?      98 1?0??0?101
 24 111?00?0??      49 1100001001      74 1??0?01?01      9? 00001?11?0
 25 ?1?0???10?      50 0???1?11?0      75 ?11?0?0?11      100 1??0?01101
```

**Les classes obtenues sont :**

```
  4 1111    11
  5  11      1
  6 1111    11
  7 1111 1   1
 10 1111    11
 11 111     11
 12 111     11
 13  11      1
 14  11      1
 18 111    111
 19 1111    11
 20 1111    11
 24 111      1
 29 11111    1
 32  11  1  11
 37 1111    11
 38  11
 39  11      1
 42  11   1 11
 43  11      1
 45 1 1     11
 47  111     1
 48  11     11
 60 111     11
 61  11     11
 64  111     1
 66 111     11
 69 1111     1
 72 1111 1 1 1
 73 111      1
 75  11     11
 76  11      1
 77  11      1
 78  11      1
 81  1111    1
 88  1       1
 89  111     1
 91 11111    1
```

```
  1     1111
  2     1111
  3     1111
  9     1111
 16     1111
 17     1111
 22   1 1111
 23     1111
 27     11111
 33     1111
 34     1111
 35     1111
 36   1 1111
 40     1 11
 41     1111
 44   1 1111
 50     1111
 51     1111
 52     1111
 53     1111
 55     1111
 57    11111
 59     1111
 65     1111
 67     1111
 84     1111
 95 1   1111
 96     1111
 97     1111
 99     1111
```

```
  8 11  1 11 1
 15 11    1  1
 21 11 1  11
 25 11    11
 26 11    11
 28 11     1
 30 11  1 11 1
 31 1  1  11
 46 11    11 1
 49 11    1  1
 54 11 1  11
 56 11    11
 58 11    11
 62 111   11
 63 11    11
 68 11    11
 70 11    111
 71 11    11
 74 11    11 1
 79 11    11
 80 11  1 11 1
 82 11    11 1
 83 11  1 11
 85 11    11 1
 86 11    11
 87 11    11 1
 90 11    11
 92 1   1 1111
 93 11    11 1
 94 11    11 1
 98 11    11 1
100 11    11 1
```

## 7.2 RECONSTITUTION

Considérons maintenant le problème de l'estimation de la qualité de la reconstitution des données manquantes par la méthode MNDMRE. Il est certes très difficile d'introduire un critère formalisé de cette qualité. Nous proposons comme critère de qualité de cette méthode, la somme des écarts en valeurs absolues entre les valeurs véritables et les valeurs reconstituées.

Comme nous disposons d'un tableau initial complet, le processus de destruction nous permet de comparer les valeurs reconstituées par la méthode MNDMRE avec les véritables valeurs qui sont celles du tableau initial.

Soit D = {1, ..., m} où m est le nombre total des données manquantes. Notons $V_m$ le vecteur dont les composantes correspondent aux m valeurs véritables **v** . Soit $R_m$ le vecteur dont les composantes correspondent aux m valeurs reconstituées **r**. Notons **E** l'écart en valeur absolue entre les deux types de valeurs. Nous écrivons

$$\mathbf{E} = \sum_{j \in D} | v^j - r^j |$$

Notons **QR** la valeur indiquant la qualité de la reconstitution. Elle s'exprime par :

$$\mathbf{QR} = \frac{m - E}{m}$$

Cette quantité correspond au pourcentage de ressemblance entre les valeurs véritables et les valeurs reconstituées.

Le tableau 9 nous donne un aperçu de cette qualité suivant le terme **T** qui indique le taux de destruction.

| T en % | 0 | 8 | 12 | 16 | 20 | 24 | 28 | 32 | 36 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|
| QR en % | 100 | 97 | 95 | 96 | 90 | 92 | 90 | 90 | 76 | 68 |

Tableau 9 : qualité de la reconstitution.

L'évolution de cette qualité est présentée dans la figure 2.

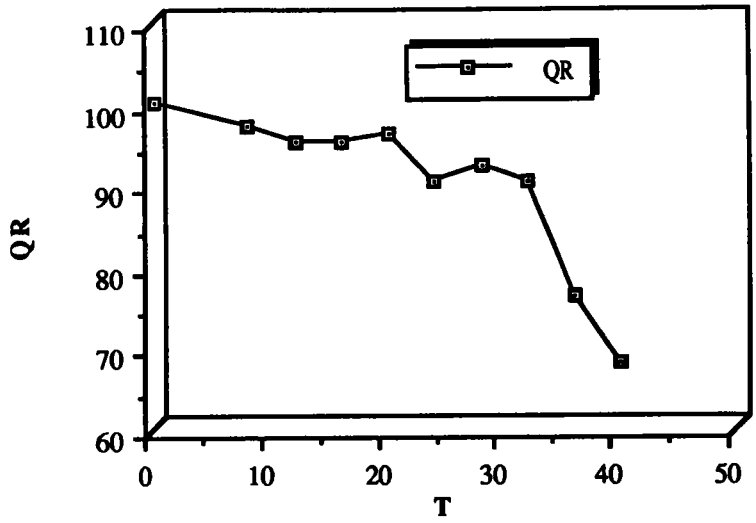



Figure 2 : évolution de la qualité de reconstitution.

Nous constatons que pour un pourcentage " raisonnable" de données manquantes cette méthode reconstitue convenablement les valeurs non observées. Les valeurs mal reconstituées sont très souvent les valeurs correspondant aux objets mal classés.

**Remarques :**

*1-Dans ce dernier paragraphe, nous pouvons nous demander quel est le critère que nous cherchons à maximiser. Pour développer la méthode MNDMIN nous avons avons cherché à maximiser la "moyenne" des vraisemblances classifiantes "reconstituantes" VCR($\theta$, P, X) pour tous les X possibles en nous appuyant sur la maximisation d'une espérance conditionnelle. Dans notre cas, le fait d'ajouter une étape de reconstitution revient simplement à chercher à maximiser une vraisemblance "reconstituante" (VR) qui dépend de (P, X ) où P désigne la partition ($P_1$, ..., $P_K$) avec K supposé connu et X désigne l'ensemble des données manquantes. Pour cela, nous nous basons également sur la maximisation de l'espérance conditionnelle E[VCR (P, $\theta$, X) / P', $\theta$')]. En effet, cela consiste à : (1) remplacer la donnée manquante par son estimation (binaire) c'est-à-dire initialiser $\theta$, (2) chercher une partition P maximisant VCR avec $\theta$ et X connu, (3) Chercher $\theta$ maximisant VCR, (4) chercher X qui maximise VCR, (5) réestimer les paramètres jusqu'à la convergence.*

*2-Nous avons également utilisé dans cette méthode la distances $L_1$ sans pondération et la distance $L_1$ pondérée et identique pour toutes les classes et nous avons remarqué que les résultats sont moins bons que ceux présentés dans le tableau 9.*

# CHAPITRE IV

# ALGORITHME EM EN PRESENCE DE DONNEES MANQUANTES

## INTRODUCTION

Dans notre situation, la méconnaissance de l'appartenance d'un individu à un composant du mélange peut être considérée comme une information manquante qui vient alors s'ajouter aux données non observées de l'échantillon. Dans ce chapitre, en utilisant le modèle des classes latentes, nous élargissons l'utilisation de l'algorithme EM à cette situation. Pour les autres variantes du modèle de mélanges de Bernoulli, c'est-à-dire lorsque le paramètre de la distribution est fixe et lorsque celui-ci est un vecteur dont les composantes dépendent de chaque variable, les étapes d'estimation et de maximisation de l'algorithme EM se déduisent de la méthode MNDMIN développée précédemment.

Dans le premier paragraphe, nous présentons notre travail qui consiste en la description de l'algorithme **EM** sur un échantillon en présence de **D**onnées **M**anquantes (**EMDM**). Les liens qui existent entre les deux approches "estimation" et "classification" sont étudiés dans le troisième paragraphe. Enfin, le dernier paragraphe est consacré à l'application de notre algorithme à des données simulées et à des données réelles.

## 1. EM EN PRESENCE DES DONNES MANQUANTES

Nous reprenons les notations précédentes $I = \{1, ..., n\}$ et $J = \{1, ..., p\}$. Le vecteur $x_i$ se décompose en deux vecteurs : $x_i^o$ qui représente les données observées et $x_i^q$ qui représente les données manquantes. Notons $O_i$ l'ensemble des indices j pour lesquels $x_i$ est observé et $M_i$ l'ensemble des indices pour lesquels $x_i$ est manquant. Dans ce contexte, nous avons : $-Y_o = (x_i^j ; j \in O_i ; i \in I)$

$-Y_m = DxZ$ où $D = (x_i^j ; j \in M_i ; i \in I)$ et $Z = (z_i ; i \in I)$.

Il est convenable d'écrire $Y = (Y_o, D, Z)$ avec $D \in \mathbf{D} = \mathbf{R}^m$ ( m est le nombre total des données manquantes) et $Z \in \mathbf{Z} = \{1, ..., K\}^n$.

Dans le chapitre précédent, nous avons vu que l'estimation de $\theta$ qui maximise la vraisemblance Log $f(Y_o ; \theta)$ était difficile à réaliser. Par ailleurs, il est plus facile de maximiser $\mathbf{E}[\text{Log}(f(Y ; \theta) / \theta^{(t)})]$.

## 1.1 HYPOTHESES SUR LES DONNEES MANQUANTES

Nous supposons que les données manquantes sont de type DMH et indépendantes. Nous supposons comme précédemment que les données manquantes suivent le modèle. Ce qui est équivalent à écrire :

$$f(\ x_i^j = 1 ;\ x_i \text{ est issu du composant } k\ ;\ j \in M_i\ ) = \alpha_{k,j}$$

$$\text{et} \qquad f(\ x_i^j = 0\ ;\ x_i \text{ est issu du composant } k\ ;\ j \in M_i\ ) = 1 - \alpha_{k,j}$$

La vraisemblance des données complètes s'écrit :

$$f(Y ; \theta) = \prod_{i \in I} p(z_i) f(x_i^o, x_i^q, \alpha(z_i))$$

Notons qu'à partir de (I.1.1) nous avons :

$$f(Y_o ; \theta) = \int_{D \times Z} f(Y_o, D, Z ; \theta)\, dD\, d\, Z$$

$$= \int_Z \left( \int_D f(Y_o, D, Z ; \theta)\, dD \right) d\, Z$$

Nous pouvons alors écrire :

$$f(Y_o ; \theta) = \prod_{i \in I} \sum_{k=1}^{K} p_k\, f(x_i,\ \alpha_k)$$

Nous décrivons maintenant les étapes de l'algorithme EM. A partir d'une solution initiale $(p_k^{(0)}, \alpha_k^{(0)} ; k = 1, K)$, l'algorithme est le suivant :

## 1.2 ETAPE ESTIMATION

Dans cette étape, nous calculons l'espérance conditionnelle $Q(\theta, \theta^{(t)})$.

$$Q(\theta, \theta^{(t)}) = E[\text{Log } f(Y_o, Y_m ; \theta) / Y_o, \theta^{(t)}]$$

$$= E[\text{Log} \prod_{i \in I} p(z_i) f(x_i^o, x_i^q ; \alpha(z_i)) / Y_o, p_k^{(t)}, \alpha_k^{(t)}]$$

$$= \sum_{i \in I} E[\text{Log } p(z_i) f(x_i^o, x_i^q ; \alpha(z_i)) / Y_o, p_k^{(t)}, \alpha_k^{(t)}]$$

$$= \sum_{i \in I} E[\text{Log } p(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}] + \sum_{i \in I} E[\text{Log } f(x_i^o ; \alpha(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}]$$
$$+ \sum_{i \in I} E[\text{Log } f(x_i^q ; \alpha(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}]$$

D'après (I.2.3.1), nous pouvons écrire :

$$Q(\theta, \theta^{(t)}) = \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log } p_k + \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) \text{ Log } f(x_i^o ; \alpha_k)$$
$$+ \sum_{i \in I} \sum_{k=1}^{K} s_k^{(t)}(x_i) E[\text{Log } f(x_i^q ; \alpha_k)]$$

où cette fois-ci, les probabilités conditionnelles indiquant que l'individu $x_i$ est issu du composant k sont définies par :

$$s_k^{(t)}(x_i) = \frac{p_k^{(t)} E[f(x_i, \alpha_k^{(t)})]}{\sum_{k=1}^{K} p_k^{(t)} E[f(x_i, \alpha_k^{(t)})]}$$

En effet, le troisième terme $\sum_{i \in I} E[\text{Log } f(x_i^q ; \alpha(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}]$ qui vient s'ajouter dans l'expression de $Q(\theta, \theta^{(t)})$ à cause des données manquantes de l'échantillon, peut s'écrire sachant que $\alpha(z_i)$ est $(\alpha_j(z_i) ; j \in J)$ :

$$\sum_{i \in I} E[\text{Log} f(x_i^q ; \alpha(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}] = \sum_{i \in I} E[\text{Log} \prod_{j \in M_i} f(x_i^j ; \alpha_j(z_i)) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}]$$
$$= \sum_{i \in I} \sum_{j \in M_i} E[\text{Log } f(x_i^j ; \alpha_j(z_i)) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}]$$

En considérant $f(x_i^j ; \alpha_j(z_i))$ comme une variable aléatoire définie de l'ensemble $\{f(x_i^j ; \alpha_j(z_i))\}$ dans l'ensemble $\{f(x_i^j ; \alpha_{k,j}) ; x_i^j \in \{0, 1\} ; \alpha_{k,j} \in \{\alpha_{1,j}, ..., \alpha_{K,j})\}\}$, nous pouvons alors écrire :

$$\sum_{i\in I} E[\text{Log } f(x_i^j ; \alpha_j(z_i)) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}]$$

$$= \sum_{i\in I} \Big[ \sum_{k=1}^{K} \{Pr [ f(x_i^j ; \alpha_j(z_i)) = f(1 ; \alpha_{k,j})) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(1; \alpha_{k,j})\}$$

$$+ \sum_{k=1}^{K} \{Pr [ f(x_i^j ; \alpha_j(z_i)) = f(0, \alpha_{k,j}) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(0 ; \alpha_{k,j})\} \Big]$$

$$= \sum_{i\in I} \Big[ \sum_{k=1}^{K} \{Pr [ x_i^j = 1 ; \alpha_j(z_i) = \alpha_{k,j} / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(1; \alpha_{k,j})\}$$

$$+ \sum_{k=1}^{K} \{Pr [ x_i^j = 0 ; \alpha_j(z_i) = \alpha_{k,j} / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(0 ; \alpha_{k,j}) )\} \Big]$$

En décomposant ces deux termes, nous obtenons :

$$\sum_{i\in I} E[\text{Log } f(x_i^j ; \alpha_j(z_i)) / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}]$$

$$= \sum_{i\in I} \Big[ \sum_{k=1}^{K} \{Pr [ x_i^j = 1 / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(1, \alpha_{k,j}) )$$

$$+ Pr [ x_i^j = 0 / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \text{ Log } f(0, \alpha_{k,j}) )\}$$

$$\cdot Pr [\alpha_j(z_i) = \alpha_{k,j} / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] \Big]$$

$$= \sum_{i\in I} \sum_{k=1}^{K} Pr [\alpha_j(z_i) = \alpha_{k,j} / Y_o, p_k^{(t)}, \alpha_{k,j}^{(t)}] E[\text{Log } f(x_i^j ; \alpha_{k,j})]$$

$$= \sum_{k=1}^{K} s_k^{(t)}(x_i) \; E[\text{Log } f(x_i^j ; \alpha_{k,j})]$$

d'où

$$\sum_{i\in I} E[\text{Log } f(x_i^q ; \alpha(z_i) / Y_o, p_k^{(t)}, \alpha_k^{(t)}] = \sum_{i\in I} \sum_{k=1}^{K} s_k^{(t)} (x_i) E[\text{Log } f(x_i , \alpha_k))] \quad \#$$

## 1.2 ETAPE MAXIMISATION

Cette étape consiste à trouver $p_k^{(t+1)}$, $\alpha_k^{(t+1)}$ maximisant $Q( \theta^{(t+1)} , \theta^{(t)} )$.

**Calcul des $p_k^{(t+1)}$ :**

Nous avons montré dans le premier chapitre que pour tout k =1, K, nous avons :

$$p_k^{(t+1)} = \frac{\sum_{i\in I} s_k^{(t)}(x_i)}{n}$$

**Calcul des** $\alpha_k^{(t+1)}$

Elles sont solutions de l'équation : $\dfrac{\partial\, E[\text{Log } f(x_i, \alpha_k^{(t+1)})]}{\partial\, \alpha_k} = 0$

ou $\sum_{i \in I} s_k^{(t)}(x_i) \dfrac{\partial\, Q(\theta^{(t+1)}, \theta^{(t)})}{\partial\, \alpha_k} = 0$ pour k = 1, K.

En utilisant la décomposition de $x_i$ en $x_i^o$ et $x_i^q$ cette équation s'écrit :

$$\sum_{i \in I} s_k^{(t)}(x_i) \frac{\partial\, E[\text{Log } f(x_i^o, \alpha_k^{(t+1)})]}{\partial\, \alpha_k} + \sum_{i \in I} s_k^{(t)}(x_i) \frac{\partial\, E[\text{Log } f(x_i^q, \alpha_k^{(t+1)})]}{\partial\, \alpha_k} = 0$$

Tout d'abord explicitons les deux termes $E[\text{Log } f(x_i^o ; \alpha_k^{(t+1)})]$ et $E[\text{Log } f(x_i^q ; \alpha_k^{(t+1)})]$.

$$\begin{aligned} E[\text{Log } f(x_i^o ; \alpha_k^{(t+1)})] &= \text{Log } f(x_i^o ; \alpha_k^{(t+1)}) \\ &= \sum_{j \in O_i} x_i^j \text{ Log } \alpha_{k,j}^{(t+1)} + (1-x_i^j) \text{ Log } (1- \alpha_{k,j}^{(t+1)}) \end{aligned}$$

et

$$\begin{aligned} E[\text{Log } f(x_i^q ; \alpha_k^{(t+1)})] &= \sum_{j \in M_i} \text{Log } f(x_i^j ; \alpha_{k,j}^{(t+1)}) \\ &= \sum_{j \in M_i} \{ f(x_i^j = 1) \text{ Log } f(x_i^j = 1; \alpha_{k,j}^{(t+1)}) \\ &\qquad + \; f(x_i^j = 0) \text{ Log } f(x_i^j = 0 ; \alpha_{k,j}^{(t+1)}) \} \\ &= \sum_{j \in M_i} \{ \alpha_{k,j}^{(t)} \text{ Log } \alpha_{k,j}^{(t+1)} + (1-\alpha_{k,j}^{(t)}) \text{ Log } (1- \alpha_{k,j}^{(t+1)}) \} \end{aligned}$$

Nous obtenons alors :

$$\sum_{i \in I} s_k^{(t)}(x_i) \{ \sum_{j \in O_i} \frac{x_i^j - \alpha_{k,j}^{(t+1)}}{\alpha_{k,j}^{(t+1)} ( 1- \alpha_{k,j}^{(t+1)})} + \sum_{j \in M_i} \frac{\alpha_{k,j}^{(t)} - \alpha_{k,j}^{(t+1)}}{\alpha_{k,j}^{(t+1)} ( 1- \alpha_{k,j}^{(t+1)})} \} = 0$$

Rappelons que pour chaque composant les p variables sont mutuellement indépendantes. De ce fait, chacune des composantes du vecteur $\alpha_k^{(t+1)}$ est solution de l'équation $L_{k,j} = 0$ où

$$L_{k,j} = \sum_{i \in I} s_k^{(t)}(x_i) \left\{ \sum_{i \in R_j} ( x_i^j - \alpha_{k,j}^{(t+1)} ) + \sum_{i \in S_j} ( \alpha_{k,j}^{(t)} - \alpha_{k,j}^{(t+1)} ) \right\}$$

d'où

$$\alpha_{k,j}^{(t+1)} = \frac{\sum_{i \in R_j} s_k^{(t)}(x_i)\, x_i^j + \sum_{i \in S_j} s_k^{(t)}(x_i)\, \alpha_{k,j}^{(t)}}{\sum_{i \in I} s_k^{(t)}(x_i)}$$

où $R_j = \{ i \in I ; x_i^j$ est observée$\}$ et $S_j = \{ i \in I ; x_i^j$ n'est pas observée$\}$.

Nous avons évidemment pour tout $j \in J = \{1, ..., p\}$ : $R_j \cup S_j = I$.

Dans le tableau 1, nous proposons un récapitulatif des estimations des paramètres du mélange dans le cas où les données de l'échantillon sont toutes observées et dans le cas où certaines valeurs sont manquantes.

| Toutes les données de l'échantillon sont observées | Les données de l'échantillon comportent des données manquantes |
|---|---|
| Etape d'estimation | |
| $s_k^{(t)}(x_i) = \frac{p_k^{(t)} f(x_i, \alpha_k^{(t)})}{\sum_{k=1}^{K} p_k^{(t)} f(x_i, \alpha_k^{(t)})}$ | $s_k^{(t)}(x_i) = \frac{p_k^{(t)} E[f(x_i, \alpha_k^{(t)})]}{\sum_{k=1}^{K} p_k^{(t)} E[f(x_i, \alpha_k^{(t)})]}$ |
| Etape de maximisation | |
| $p_k^{(t+1)} = \frac{\sum_{i \in I} s_k^{(t)}(x_i)}{n}$ | $p_k^{(t+1)} = \frac{\sum_{i \in I} s_k^{(t)}(x_i)}{n}$ |
| $\alpha_{k,j}^{(t+1)} = \frac{\sum_{i \in I} s_k^{(t)}(x_i)\, x_i^j}{\sum_{i \in I} s_k^{(t)}(x_i)}$ | $\alpha_{k,j}^{(t+1)} = \frac{\sum_{i \in R_j} s_k^{(t)}(x_i) x_i^j + \sum_{i \in S_j} s_k^{(t)}(x_i)\, \alpha_{k,j}^{(t)}}{\sum_{i \in I} s_k^{(t)}(x_i)}$ |

**Tableau 1 : comparaison des expressions des estimations des paramètres du modèle de mélanges obtenues par l'algorithme EM lorsque les données sont observées et par l'algorithme EMDM lorsque les données de l'échantillon ne sont pas toutes observées.**

**Remarque :**

*Lors des étapes d'estimation et de maximisation, le processus est tel que les données manquantes sont reconstituées à chaque étape. En effet, pour une variable $x_j$ les données non observées des individus issus d'un composant k sont reconstituées par le paramètre $a_{k,j}$ qui est estimé dans l'étape précédente. Dans le chapitre précédent nous avons relevé la même remarque.*

## 1.3 EXEMPLE D'APPLICATION

Soit un tableau de données binaires comportant des données manquantes codées "?", croisant 10 individus et 10 variables (tableau 2). Nous supposons toujours que les données manquantes sont de type DMH.

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

Tableau 2 : tableau de données.

Nous appliquons la méthode EMDM (l'algorithme EM avec Données Manquantes) en demandant deux composants du mélange. Les résultats figurent dans le tableau 3.

| | k = 1 | k = 2 |
|---|---|---|
| $p_k$ | 0.5 | 0.5 |
| $a_{k,1}$ | 0.33 | 1.00 |
| $a_{k,2}$ | 0.00 | 1.00 |
| $a_{k,3}$ | 0.00 | 1.00 |
| $a_{k,4}$ | 0.00 | 1.00 |
| $a_{k,5}$ | 1.00 | 0.00 |
| $a_{k,6}$ | 1.00 | 0.50 |
| $a_{k,7}$ | 1.00 | 0.00 |
| $a_{k,8}$ | 1.00 | 0.00 |
| $a_{k,9}$ | 0.00 | 1.00 |
| $a_{k,10}$ | 0.25 | 1.00 |

Tableau 3 : estimations des paramètres des 2 composants.

La convergence est atteinte en 31 itérations et la valeur du critère est -16.2017.

Nous notons que pour les autres variantes du modèle associé aux données binaires les démonstrations sont identiques et se déduisent de celles développées dans le chapitre précédent.

**Remarques :**

*Dans le chapitre III, nous avons montré qu'en présence de données manquantes, la valeur $a_{k,j}^{(t+1)}$ (pour une variable $x^j$ et dans un composant k) dépendait de $a_{k,j}^{(t)}$, de $m_k^j$ et de $\xi$ où $\xi = (q_k^j - t_k^j)/(1- 2\varepsilon^{(t)})$. Le choix de $a_{k,j}^{(t+1)}$ est défini dans la proposition 2 de ce même chapitre.*

*Dans l'algorithme EM, lors de l'étape de maximisation nous pouvons montrer que pour la variante la plus simple, la valeur $a_{k,j}^{(t+1)}$ sera définie comme dans la proposition 2 du III, il suffit en effet de poser :*

$$\xi = \{ \sum_{i \in Aj} s_k^{(t)}(x_i)\, x_i^j - \sum_{i \in Bj} s_k^{(t)}(x_i)\, x_i^j \} / \sum_{i \in Cj} s_k^{(t)}(x_i)(1- 2\varepsilon^{(t)})$$

*où*

$A_j = \{ i \in I ; x_i^j = 0\}$, $B_j = \{ i \in I ; x_i^j = 1\}$ *et* $C_j = \{ i \in I ; x_i^j$ *n'est pas observée*$\}$.

*Les valeurs 1 et 0 apparaissent comme pondérées par les probabilités $s_k^{(t)}(x_i)$.*

*Rappelons que dans le premier chapitre, nous avons développé en absence de données manquantes, l'algorithme EM pour les trois variantes du modèle de mélanges de Bernoulli. Nous avons ainsi noté que la valeur $a_{k,j}^{(t+1)}$ calculée à l'itération (t+1), apparaissait comme la médiane binaire de l'ensemble suivant : $\{(x_i^j , s_k^{(t)}(x_i)), i \in I\}$. Nous pouvons retrouver ce résultat à partir de l'expression de $\xi$. En effet, dans ce cas le terme du dénominateur n'apparaîtra pas et $\xi$ désignera le signe de la différence des valeurs 0 et des valeurs 1 pondérées.*

## 2. EXPERIENCES NUMERIQUES

Nous disposons d'un tableau de données binaires simulées croisant 100 individus et 10 variables. Nous appliquons la méthode EMDM en demandant trois composants du mélange à partir du tableau initial qui a subi différents taux de destruction (T). Les

résultats sont présentés dans les tableaux 4 à 11. Ces résultats correspondent aux estimations des paramètres du modèle.

| T = 0% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.330 | 0.340 | 0.330 |
| $\alpha_{k,1}$ | 0.939 | 1.000 | 0.061 |
| $\alpha_{k,2}$ | 0.970 | 0.971 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.060 | 0.091 |
| $\alpha_{k,4}$ | 0.880 | 0.059 | 0.091 |
| $\alpha_{k,5}$ | 0.121 | 0.147 | 1.000 |
| $\alpha_{k,6}$ | 0.121 | 0.000 | 0.939 |
| $\alpha_{k,7}$ | 0.030 | 0.910 | 1.000 |
| $\alpha_{k,8}$ | 0.090 | 0.912 | 1.000 |
| $\alpha_{k,9}$ | 0.939 | 0.089 | 0.031 |
| $\alpha_{k,10}$ | 0.939 | 0.912 | 0.031 |

Tableau 4 : 0% de destruction

| T = 12% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.331 | 0.340 | 0.329 |
| $\alpha_{k,1}$ | 0.936 | 1.000 | 0.062 |
| $\alpha_{k,2}$ | 0.967 | 0.964 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.062 | 0.111 |
| $\alpha_{k,4}$ | 0.890 | 0.067 | 0.111 |
| $\alpha_{k,5}$ | 0.143 | 0.175 | 1.000 |
| $\alpha_{k,6}$ | 0.142 | 0.000 | 0.928 |
| $\alpha_{k,7}$ | 0.032 | 0.962 | 1.000 |
| $\alpha_{k,8}$ | 0.105 | 0.929 | 1.000 |
| $\alpha_{k,9}$ | 0.924 | 0.102 | 0.032 |
| $\alpha_{k,10}$ | 0.963 | 0.903 | 0.036 |

Tableau 5 : 12% de destruction

| T = 20% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.338 | 0.333 | 0.329 |
| $\alpha_{k,1}$ | 0.934 | 1.000 | 0.072 |
| $\alpha_{k,2}$ | 0.965 | 0.956 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.046 | 0.111 |
| $\alpha_{k,4}$ | 0.879 | 0.075 | 0.083 |
| $\alpha_{k,5}$ | 0.127 | 0.160 | 1.000 |
| $\alpha_{k,6}$ | 0.145 | 0.000 | 0.923 |
| $\alpha_{k,7}$ | 0.035 | 0.898 | 1.000 |
| $\alpha_{k,8}$ | 0.052 | 0.925 | 1.000 |
| $\alpha_{k,9}$ | 0.928 | 0.133 | 0.034 |
| $\alpha_{k,10}$ | 0.960 | 0.892 | 0.038 |

Tableau 6 : 20% de destruction

| T = 32% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.338 | 0.332 | 0.330 |
| $\alpha_{k,1}$ | 0.927 | 1.000 | 0.040 |
| $\alpha_{k,2}$ | 0.963 | 0.958 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.048 | 0.111 |
| $\alpha_{k,4}$ | 0.868 | 0.078 | 0.087 |
| $\alpha_{k,5}$ | 0.132 | 0.165 | 1.000 |
| $\alpha_{k,6}$ | 0.145 | 0.000 | 0.913 |
| $\alpha_{k,7}$ | 0.037 | 0.894 | 1.000 |
| $\alpha_{k,8}$ | 0.054 | 0.923 | 1.000 |
| $\alpha_{k,9}$ | 0.917 | 0.138 | 0.039 |
| $\alpha_{k,10}$ | 0.957 | 0.888 | 0.040 |

Tableau 7 : 32% de destruction

| T = 40% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.326 | 0.347 | 0.327 |
| $\alpha_{k,1}$ | 0.941 | 1.000 | 0.033 |
| $\alpha_{k,2}$ | 0.953 | 1.000 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.058 | 0.152 |
| $\alpha_{k,4}$ | 0.790 | 0.090 | 0.100 |
| $\alpha_{k,5}$ | 0.177 | 0.221 | 1.000 |
| $\alpha_{k,6}$ | 0.103 | 0.070 | 0.935 |
| $\alpha_{k,7}$ | 0.004 | 0.867 | 1.000 |
| $\alpha_{k,8}$ | 0.000 | 0.909 | 1.000 |
| $\alpha_{k,9}$ | 0.948 | 0.090 | 0.053 |
| $\alpha_{k,10}$ | 0.960 | 1.000 | 0.000 |

Tableau 8 : 40% de destruction

| T = 50% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.341 | 0.325 | 0.335 |
| $\alpha_{k,1}$ | 0.920 | 1.000 | 0.092 |
| $\alpha_{k,2}$ | 0.950 | 1.000 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.091 | 0.159 |
| $\alpha_{k,4}$ | 0.888 | 0.068 | 0.165 |
| $\alpha_{k,5}$ | 0.129 | 0.206 | 1.000 |
| $\alpha_{k,6}$ | 0.107 | 0.079 | 0.932 |
| $\alpha_{k,7}$ | 0.001 | 0.889 | 1.000 |
| $\alpha_{k,8}$ | 0.000 | 0.882 | 1.000 |
| $\alpha_{k,9}$ | 0.906 | 0.071 | 0.061 |
| $\alpha_{k,10}$ | 1.000 | 1.000 | 0.000 |

Tableau 9 : 50% de destruction

| T = 56% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.335 | 0.327 | 0.338 |
| $\alpha_{k,1}$ | 0.910 | 1.000 | 0.001 |
| $\alpha_{k,2}$ | 0.942 | 1.000 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.103 | 0.139 |
| $\alpha_{k,4}$ | 0.871 | 0.071 | 0.204 |
| $\alpha_{k,5}$ | 0.142 | 0.225 | 1.000 |
| $\alpha_{k,6}$ | 0.140 | 0.072 | 0.903 |
| $\alpha_{k,7}$ | 0.000 | 0.899 | 1.000 |
| $\alpha_{k,8}$ | 0.000 | 0.852 | 1.000 |
| $\alpha_{k,9}$ | 0.893 | 0.081 | 0.065 |
| $\alpha_{k,10}$ | 1.000 | 1.000 | 0.000 |

Tableau 10 : 56% de destruction

| T = 64% | k = 1 | k = 2 | k = 3 |
|---|---|---|---|
| $p_k$ | 0.311 | 0.346 | 0.344 |
| $\alpha_{k,1}$ | 0.883 | 1.000 | 0.001 |
| $\alpha_{k,2}$ | 0.923 | 1.000 | 0.000 |
| $\alpha_{k,3}$ | 1.000 | 0.000 | 0.168 |
| $\alpha_{k,4}$ | 0.746 | 0.103 | 0.103 |
| $\alpha_{k,5}$ | 0.123 | 0.264 | 1.000 |
| $\alpha_{k,6}$ | 0.327 | 0.000 | 0.878 |
| $\alpha_{k,7}$ | 0.000 | 1.000 | 1.000 |
| $\alpha_{k,8}$ | 0.000 | 0.615 | 1.000 |
| $\alpha_{k,9}$ | 1.000 | 0.120 | 0.082 |
| $\alpha_{k,10}$ | 1.000 | 1.000 | 0.000 |

Tableau 11 : 64% de destruction

Nous constatons que les estimations des paramètres du modèle sont assez similaires même lorsque le pourcentage des données manquantes est important.

## 3. LIENS AVEC L'APPROCHE CLASSIFICATION

Dans l'approche "estimation", le problème de la recherche des classes n'est qu'un sous-produit de l'estimation des paramètres du modèle. Cependant, pour l'approche classification auquelle nous nous intéressons dans ce paragraphe, les estimés de ces paramètres sont un sous-produit de la décomposition en classes de l'échantillon. Ainsi, à partir des paramètres estimés par l'algorithme EMDM, nous nous proposons de chercher une partition en K classes. Pour cela, nous nous servons des liens existant entre le modèle associé aux données binaires dans le cas le plus général (les paramètres du modèle dépendent à la fois des classes et des variables) et le modèle des classes latentes (voir chapitre I).

Nous disposons par le méthode EMDM de l'estimation des ($\alpha_k$ ; k = 1, ..., K). Ces derniers nous permettent de composer des noyaux binaires ($a_k$ ; k = 1, ..., K). L'étape de représentation est ainsi définie. Il nous reste à caractériser l'étape d'affectation. Pour cela, nous proposons d'utiliser la mesure de dissimilarité vue dans le chapitre II. Celle-ci s'écrit entre un vecteur $x_i$ qui peut être incomplet et un vecteur complet $a_k$ comme suit :

$$d(x_i, a_k) = d_{\varepsilon_k}(x_i, a_k) - A_k$$

ou encore :

$$d(x_i, a_k) = \sum_{j \in O_i} \text{Log} \frac{1-\varepsilon_k^j}{\varepsilon_k^j} |x^j - a_k^j| + \sum_{j \in J} \text{Log}(1-\varepsilon_k^j)$$

Cette mesure utilise uniquement les composantes observées simultanément. Le choix de cette mesure n'est pas au hasard. En effet, comme le processus est tel que les données manquantes de chaque variable sont reconstituées par les paramètres $\alpha_{k,j}$ correspondants, cela revient dans notre approche à reconstituer ces données par les noyaux binaires $a_k^j$ correspondants. Ainsi, l'influence des données non observées est nulle lors de l'étape d'affectation. Nous avons noté également dans le chapitre II que la méthode MNDM utilisant cette mesure a donné de bons résultats.

Ainsi, avec les différents taux de destruction réalisés, nous récupérons les estimations des paramètres du modèle par la méthode EMDM et nous utilisons la mesure d'adéquation citée ci-dessus pour caractériser l'étape d'affectation. Nous notons :

EMDM+ = EMDM + étape d'affectation.

Les résultats présentés dans le tableau 12 correspondent à l'évolution du pourcentage d'objets mal classés.

| % de destruction | % de mal classés | objets mal classés |
|---|---|---|
| 0 | 0 | Ø |
| 8 | 0 | Ø |
| 12 | 0 | Ø |
| 16 | 1 | {77} |
| 20 | 3 | {13, 77, 78} |
| 24 | 3 | {13, 77, 78} |
| 28 | 3 | {18, 77, 78} |
| 32 | 4 | {18, 38, 77, 78} |
| 36 | 4 | {18, 38, 77, 78} |
| 40 | 3 | {72, 77, 78} |
| 44 | 3 | {72, 77, 78} |
| 48 | 3 | {72, 77, 78} |
| 50 | 5 | {25, 63, 72, 77, 78} |
| 52 | 7 | {25, 63, 72, 77, 78, 82, 95} |
| 56 | 6 | {21, 63, 72, 77, 78, 82} |
| 64 | 12 | {4, 18, 25, 38, 43, 63, 71, 73, 78, 79, 82, 95} |

Tableau 12 : pourcentages d'objets mal classés et énumération de ces objets.

Les résultats sont très satisfaisants. Une comparaison avec la méthode MNDM3, lorsque $\alpha = 1/2$, est illustrée par la figure 1.

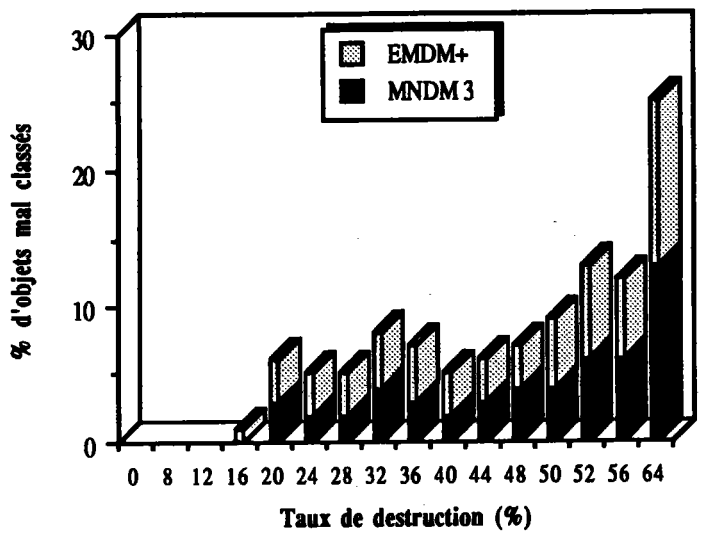



Figure 1 : comparaison des méthodes EMDM+ et MNDM3 lorsque $\alpha = 1/2$.

# CHAPITRE V

# METHODE MNDQAN ET DONNEES MANQUANTES

## INTRODUCTION

Dans le deuxième chapitre, nous avons vu que l'écriture du critère métrique est telle que le paramètre $\alpha$ semble substituer les données manquantes. Ce paramètre $\alpha$, qui peut être un réel ou un vecteur dépendant de chaque variable, est supposé connu. Ainsi, nous avons développé plusieurs variantes de la méthode des nuées dynamiques MNDM. Dans ce chapitre, en exploitant cette contatation, nous montrons comment utiliser la méthode des nuées dynamiques sur données quantitatives MNDQAN afin de trouver la meilleure partition.

Dans le premier paragraphe, nous rappelons la méthode de classification MNDM sous sa variante la plus simple ($\varepsilon$ paramètre fixe et $\alpha$ un réel). Dans le second paragraphe, nous montrons sous quelles conditions la méthode MNDQAN peut être appliquée et nous établissons les relations existant entre les méthodes MNDQAN et MNDM. Dans le troisième paragraphe, nous présentons une étude comparative des deux critères associés à ces deux méthodes et appliquons celles-ci sur des données binaires dans le dernier paragraphe.

## 1. METHODE MNDM

### 1.1 RAPPELS ET NOTATIONS

Dans les paragraphes faisant l'objet de ce chapitre, nous utilisons les mêmes notations que précédemment.

$\Omega$ étant un ensemble de n individus $\{x_1, ..., x_n\}$ mesurés par p variables binaires $\{x^1, ..., x^p\}$, nous cherchons une partition $P = (P_1, ..., P_K)$ de $\Omega$ en K classes "homogènes" sachant que certaines variables n'ont pas été relevées sur certains objets.

### 1.2 METHODE

Lorsque les variables suivent une distribution de Bernoulli de paramètre identique et lorsque les données manquantes suivent une distribution de paramètre $\alpha$, nous avons montré qu'en remplaçant le critère de vraisemblance classifiante par son espérance, le problème de la recherche d'une partition $P = (P_1, ..., P_K)$, où K est supposé connu, et où chaque classe $P_k$ est assimilable à un sous échantillon qui suit une loi $f(./a_k)$, revient à minimiser le critère de classification W(P, L) :

$$W(P, L) = \sum_{k=1}^{K} \sum_{i \in P_k} \left( \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} |\alpha - a_k^j| \right)$$

La mesure de dissimilarité apparaît comme une distance $L_1$ et les composantes non observées des vecteurs incomplets semblent être complétées par le paramètre $\alpha$. Le critère W(P, L) peut également s'écrire :

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

avec $x_i^j \in \{0, 1\}$ si $x_i^j$ est observée et $x_i^j = \alpha$ sinon.

Le critère étant ainsi exprimé, nous ne pouvons que constater la possibilité de transformer le tableau initial, où les données binaires côtoient des valeurs non observées, en un tableau dont les valeurs possibles appartiennent à $\{0, 1, \alpha\}$. Alors, les données manquantes sont toutes reconstituées par la valeur $\alpha$.

## 2. LIENS ENTRE LES METHODES MNDM ET MNDQAN

### 2.1 METHODE MNDQAN

L'échantillon étant connu, nous pouvons nous demander, quelle est la probabilité qu'une donnée du tableau soit égale à 1 ? Pour une variable $x^j$ la réponse est donnée dans le schéma suivant sachant que $\alpha$ est connu :

$$Pr(x_i^j = 1) = 1 \quad \text{si } x_i^j = 1$$

$$Pr(x_i^j = 1) = 0 \quad \text{si } x_i^j = 0$$

$$\text{et } Pr(x_i^j = 1) = \alpha \text{ si } x_i^j \text{ est non observée.}$$

Ainsi, nous pouvons donner une interprétation "probabiliste" des données du tableau initial malgré la présence de données manquantes. Notons que, dans le cas où toutes les données sont observées, aucune modification n'intervient sur ces données initiales.

**Exemple :**

Soit un tableau de données représentant 4 individus {**a, b, c, d**} décrits par 4 variables binaires {**1, 2, 3, 4**}. Le tableau initial, où "?" représente la valeur non observée, est présenté dans le tableau 1 et son interprétation probabiliste est présentée dans le tableau 2.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **a** | 1 | 0 | ? | 0 |
| **b** | 0 | 1 | 1 | ? |
| **c** | 1 | ? | 1 | 0 |
| **d** | 0 | 0 | 1 | ? |

Tableau 1 : tableau initial.

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| **a** | 1 | 0 | α | 0 |
| **b** | 0 | 1 | 1 | α |
| **c** | 1 | α | 1 | 0 |
| **d** | 0 | 0 | 1 | α |

Tableau 2 : interprétation probabiliste.

Ainsi, nous avons transformé notre tableau de données binaires en présence de données manquantes en un tableau de probabilités. Ce tableau peut être considéré comme un tableau de données quantitatives puisque ses valeurs appartiennent à $\{0, 1, \alpha\}$.

Soit P une partition en K classes de l'ensemble des individus. Le critère associé à la méthode MNDQAN lorsque les matrices variance des composants sont connues et identiques, s'exprime par :

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

où $d_e$ représente la distance euclidienne,
et $Lg = (g_1, ..., g_k)$ l'ensemble des centres de gravités.

Il s'agit du critère de la version la plus simple et la plus utilisée des Nuées Dynamiques (algorithmes des "centres mobiles" ou de "réallocation-recentrage").

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

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

En respectant les notations déjà utilisées, nous pouvons écrire :

$$\forall\, j = 1, p \qquad g_k^j = \frac{t_k^j + \alpha\, m_k^j}{n_k}$$

Ce critère peut s'écrire également :

$$\begin{aligned} Wg(P, Lg) &= \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (x_i^j - g_k^j)^2 \\ &= \sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} (x_i^j - g_k^j)^2 + \sum_{j \in M_i} (x_i^j - g_k^j)^2 \Big) \\ &= \sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} (x_i^j - g_k^j)^2 + \sum_{j \in M_i} (\alpha - g_k^j)^2 \Big) \end{aligned}$$

La méthode des nuées dynamiques repose essentiellement sur le choix de la nature du noyau. **Si nous imposons aux noyaux d'appartenir à $\{0, 1\}^p$ nous aurons**

$$Wg(P, Lg) = \sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} |x_i^j - g_k^j| + \sum_{j \in M_i} (\alpha - g_k^j)^2 \Big)$$

Nous allons chercher les liens existant entre les critères W(P, L) et $W_g(P, L_g)$ lorsque les noyaux sont des éléments de $\{0, 1\}^p$.

## 2.2 LIENS

**Proposition 1 :**

*Lorsque les noyaux des classes appartiennent à $\{0, 1\}^p$ nous avons :*

$$W(P, L) = W_g(P, L_g) + \text{constante},$$

*et les deux critères sont alors dits équivalents.*

**Preuve :**

$$\begin{aligned} W(P, L) &= \sum_{k=1}^{K} \sum_{i \in P_k} \Big( \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} |\alpha - a_k^j| \Big) \\ &= \sum_{k=1}^{K} \sum_{i \in P_k} \Big\{ \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} [\alpha(1 - a_k^j) + (1 - \alpha) a_k^j] \Big\} \\ &= \sum_{k=1}^{K} \sum_{i \in P_k} \Big\{ \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} [a_k^j (1 - 2\alpha) + \alpha] \Big\} \end{aligned}$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \{ \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} [a_k^j ((1-\alpha)^2 - \alpha^2) + \alpha^2 + (\alpha - \alpha^2)] \}$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} \{ \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} [(1-\alpha)^2 a_k^j + \alpha^2 (1 - a_k^j) + (\alpha - \alpha^2)] \}$$

$$= \sum_{k=1}^{K} \sum_{i \in P_k} ( \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} (\alpha - a_k^j)^2 ) + \sum_{j \in M_i} [(\alpha - \alpha^2)]$$

$$= W_g(P, L_g) + \sum_{j \in M_i} [(\alpha - \alpha^2)]$$

Les méthodes MNDM dans le cas le plus simple et MNDQAN sont équivalentes. En effet, sous la contrainte que les noyaux sont de même nature, les deux fonctions d'affectation et de représentation sont identiques.

## 3. ETUDE COMPARATIVE DES DEUX CRITERES

Dans ce paragraphe nous comparons les critères $W_g(P, L_g)$ où **les noyaux $g_k$ sont les centres de gravité** des classes et le critère $W(P, L)$ où **les noyaux $a_k$ appartiennent à $\{0,1\}^p$.**

**Proposition 2 :**

*Pour $k = 1, ..., K$, lorsque $a_k \in \{0, 1\}^p$ et $g_k \in R^p$ nous avons :*

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

**Preuve :**

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

$$= \sum_{k=1}^{K} \sum_{i \in P_k} ( \sum_{j \in O_i} |x_i^j - a_k^j| + \sum_{j \in M_i} (\alpha - a_k^j)^2) + A$$

où

$$A = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in M_i} ( |\alpha - a_k^j| - (\alpha - a_k^j)^2 )$$

Comme $a_k \in \{0, 1\}^p$ le terme A peut s'écrire :

$$A = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in M_i} \alpha(1-\alpha) = m\alpha(1-\alpha)$$

où m est le nombre total des données manquantes.

Le critère W(P, L) peut s'écrire également :

$$W(P, L) = \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} (x_i^j - a_k^j)^2 + m\alpha(1-\alpha)$$

où $x_i^j$ appartient à {0, 1} si est $x_i^j$ observée et égal à $\alpha$ sinon.

La relation d'Hyguens dans $R^p$ muni de la distance euclidienne permet d'écrire.

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

ou encore :

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

nous en déduisons :

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

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

**Remarques :**

*1 Lorsque* $\alpha = 1/2$ $W(P, L) = W_g(P, L_g) + \sum_{k=1}^{K} n_k d_e^2(a_k, g_k) + \frac{1}{4} m.$

*2 Lorsque* $\alpha$ *est un vecteur dépendant de chaque variable nous avons :*

$$W(P, L) = W_g(P, L_g) + \sum_{k=1}^{K} n_k d_e^2(a_k, g_k) + \sum_{j \in J} m^j \alpha^j (1 - \alpha^j)$$

**Exemple :**

Pour illustrer cette dernière proposition, nous proposons une application de ces deux méthodes à partir d'un tableau de données binaires (avec données manquantes) croisant 10 individus identifiés par les lettres **a à j**, et 10 variables identifiées par les nombres **1 à 10** ( tableau 3).

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

Tableau 3 : tableau de données.

Nous appliquons les deux méthodes MNDM et MNDQAN sur ce tableau en demandant 3 classes. Les tableaux réordonnés résultant sont respectivement présentés dans les tableaux 4 et 5.

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

Tableau 4 : tableau réordonné par MNDM.

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

Tableau 5 : tableau réordonné par MNDQAN.

Les noyaux correspondant à ces deux partitions sont respectivement présentés dans les tableaux 6 et 7.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{a_1}$ | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 |
| $\mathbf{a_2}$ | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 |
| $\mathbf{a_3}$ | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 |

Tableau 6 : tableau des noyaux binaires (MNDM).

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\mathbf{a_1}$ | 0.43 | 0.11 | 0.11 | 0.23 | 1.00 | 0.91 | 0.91 | 0.83 | 0.11 | 0.31 |
| $\mathbf{a_2}$ | 0.79 | 0.79 | 0.79 | 0.79 | 0.00 | 0.29 | 0.00 | 0.57 | 0.79 | 0.57 |
| $\mathbf{a_3}$ | 1.00 | 1.00 | 1.00 | 1.00 | 0.38 | 0.71 | 0.38 | 0.00 | 0.86 | 0.86 |

Tableau 7 : tableau des noyaux non binaires (MNDQAN).

**Remarque :**

*Cette méthode représente un inconvénient. En effet, la présentation du critère $W_g(P, L_g)$ explique clairement le fait, constaté expérimentalement, que cette méthode ait tendance à donner des classes sphériques de même volume. Friedman, Rubin (1967) et Govaert (1975) ont proposé un algorithme de type Nuées Dynamiques capable de reconnaître des classes ayant le même type de dispersion mais possédant des directions d'allongement inconnues. Celeux (1988) a montré que le critère utilisé par ces auteurs correspond à l'hypothèse d'une population issue d'un mélange de lois gaussiennes dont les matrices variance des composants sont identiques mais inconnues.*

## 4. APPLICATIONS

Nous appliquons la méthode MNDQAN dans le cas le plus simple avec les trois options qui dépendent du paramètre $\alpha$. Celui-ci est représenté, par la fréquence des 1 dans le tableau dans la première colonne, par un vecteur dont les composantes dépendent des variables dans la deuxième colonne et par la valeur 1/2 dans la troisième. Nous verrons ainsi l'évolution des pourcentages d'objets mal classés dans les différents cas suivant le taux de destruction T ( en %). Le tableau 8 illustre l'évolution des objets mal classés sur les données simulées.

| T | $\alpha$ | $\alpha_j$ | 1/2 |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 |
| 12 | 0 | 0 | 0 |
| 16 | 1 | 0 | 1 |
| 20 | 2 | 1 | 2 |
| 24 | 2 | 1 | 2 |
| 28 | 3 | 2 | 3 |
| 32 | 6 | 3 | 3 |
| 36 | 6 | 4 | 3 |
| 40 | 6 | 5 | 4 |
| 44 | 8 | 3 | 7 |
| 48 | 9 | 7 | 7 |
| 50 | 9 | 6 | 7 |
| 52 | 9 | 6 | 7 |
| 56 | 11 | 6 | 8 |
| 64 | 17 | 17 | 14 |

Tableau 8 : pourcentages d'objets mal classés.

En comparant avec les résultats obtenus dans le deuxième chapitre, nous remarquons que la méthode MNDQAN donne de meilleurs résultats que la méthode MNDM. Dans les figures 1, 2 et 3 nous présentons respectivement l'évolution du pourcentage des objets mal classés suivant les deux méthodes MNDM et MNDQAN lorsque $\alpha$ est un réel, $\alpha$ est un vecteur puis $\alpha = 1/2$.

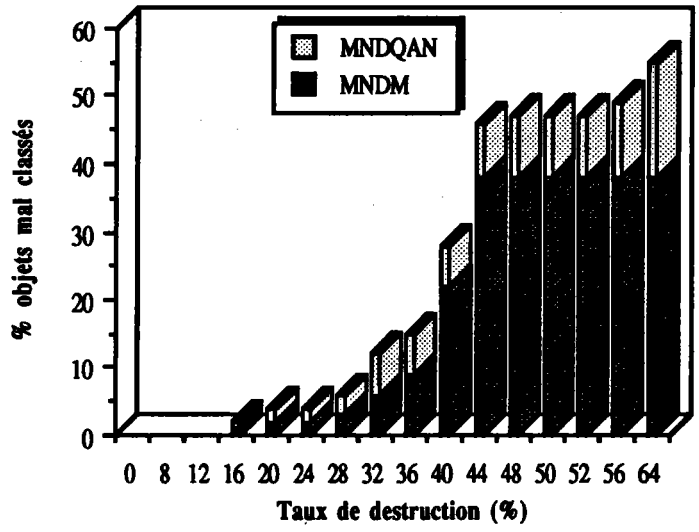



Figure 1 : comparaison de MNDM et MNDQAN ($\alpha$ réel).

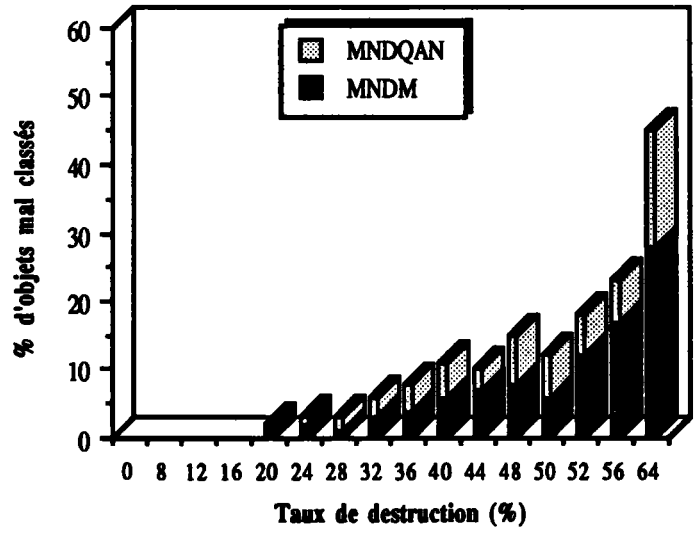



Figure 2 : comparaison de MNDM et MNDQAN ($\alpha$ vecteur).

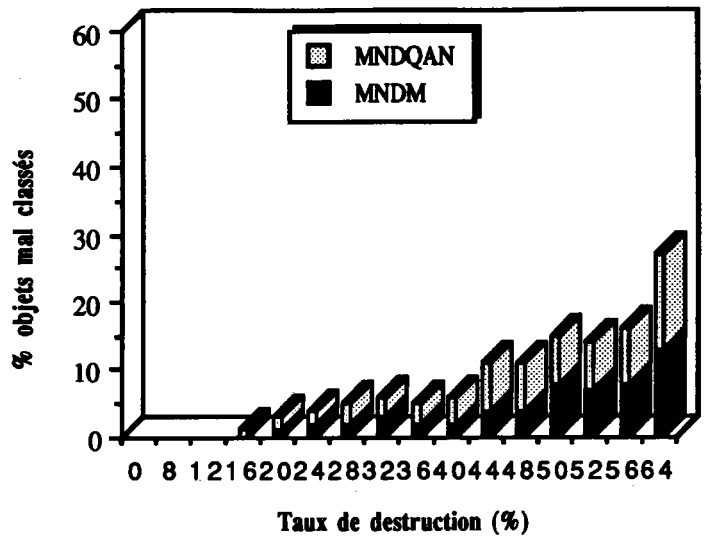



Figure 3 : comparaison de MNDM et MNDQAN ($\alpha$ =1/2).

Lorsque $\alpha$ = 1/2, nous constatons que les pourcentages d'objets mal classés sont sensiblement égaux. Nous avons appliqué les deux méthodes MNDQAN et MNDM sur le fichier de données **méro** après avoir généré des données manquantes avec des taux de destruction atteignant 64%, nous avons remarqué que lorsque $\alpha \neq 1/2$ la méthode MNDQAN donne toujours de meilleurs résultats. En effet, à partir de 50% de données manquantes seul l'individu 25 quitte sa classe.

# DONNEES QUALITATIVES

# CHAPITRE VI

# EXTENSION DES METHODES MNDM, MNDMIN ET EMDM

## INTRODUCTION

Nous avons proposé précédemment des modèles associés aux données binaires avec données manquantes. Dans le premier chapitre, les données non observées de chaque variable suivaient une distribution de Bernoulli dont le paramètre est identique pour toutes les variables. Nous avons généralisé ce modèle en prenant des paramètres qui dépendaient des variables. Ces paramètres étaient supposés connus et nous avons proposé dans ce cas la méthode de classification **MNDM**. Dans le deuxième chapitre, les paramètres associés aux données manquantes dépendaient du modèle et nous avons proposé la méthode de classification **MNDMIN**. Dans le troisième chapitre, nous nous sommes intéressés davantage au problème de l'estimation des paramètres du modèle, et nous avons proposé la méthode **EMDM** qui est une extension de l'algorithme EM. En suivant la même démarche, nous proposons dans ce chapitre une extension de chacune de ces méthodes lorsque les données sont qualitatives nominales.

Dans le premier paragraphe, nous rappelons brièvement le modèle des classes latentes lorsque les données sont qualitatives. Dans le deuxième paragraphe, nous proposons un modèle dans le cas où les données qualitatives comportent des données manquantes qui dépendent de paramètres connus et indépendants du modèle. Dans le troisième paragraphe, nous traitons le cas où ces paramètres sont inconnus et dépendent du modèle.

## 1. APPROCHE CLASSIFICATION ET MODELE DES CLASSES LATENTES

Dans ce paragraphe, nous supposons que toutes les valeurs du tableau de données dont nous disposons sont observées.

## 1.1 NOTATIONS

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 nominales. Nous notons :

$$Z(I, Q) = [z_i^q]$$

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

A chaque variable q correspond l'ensemble de modalités $J_q = \{1, 2, ..., m_q\}$. Nous définissons alors l'ensemble E comme le produit $J_1xJ_2x...xJ_p$.

Notre objectif est de chercher une partition $P = (P_1, ..., P_K)$ de Z(I, Q) en K classes "homogènes".

## 1.2 MODELE DES CLASSES LATENTES

Pour obtenir une partition en K classes d'une population de n individus décrits par des variables qualitatives, une méthode classique consiste à raisonner à partir du tableau de données mis sous forme disjonctive complète et à chercher des individus qui, croisés avec les variables, maximisent le critère $\chi^2$ de contingence. Cela revient à rechercher la partition d'inertie interclasse maximum pour la métrique du $\chi^2$. Cette méthode peut s'interpréter comme une méthode d'identification du modèle des classes latentes (Celeux 1988).

Les paramètres de ce modèle sont les fréquences relatives ($p_k$, k = 1, K) des K modalités de la variable cachée ou latente et les probabilités ( $a_k^{q(j)}$ , où q(j) correspond à la modalité j de la variable q. q(j) est l'indice de J = {1, ..., m} où m est le nombre total de modalités ) que $x_i^{q(j)}=1$ sachant que l'individu présente la modalité k de la variable latente.

Nous avons bien sûr $\sum_{j \in Jq} x_i^{q(j)} = 1$ pour tout j = 1, ..., p et pour tout k = 1, ..., K.
Ce modèle revient à supposer (Everitt 1981) que les n vecteurs binaires à m coordonnées décrivant les individus sont un échantillon du mélange de densités :

$$f(x_i) = \sum_{k=1}^{K} p_k f(x_i, a_k) \qquad \text{avec} \quad f(x_i, a_k) = \prod_{q \in Q} \prod_{j \in Jq} (a_k^{q(j)})^{x_i^{q(j)}}$$

où $a_k = (a_k^{q(j)}$ ; q(j) est l'indice de J correspondant à la modalité j de q) avec $\sum_{j \in Jq} a_k^{q(j)} = 1$.

$f(x_i, a_k)$ est la densité d'une loi multinomiale multivariée de paramètre $a_k$.

## 2. MODELES ASSOCIES AUX DONNEES QUALITATIVES AVEC DONNEES MANQUANTES

### 2.1 NOTATIONS

Nous gardons les mêmes notations que précédemment mais cette fois-ci notre tableau Z(I, Q) comporte des données manquantes.

$O_i$ est l'ensemble des indices q pour lesquels les valeurs du vecteur $z_i$ sont observées. $M_i$ est l'ensemble des indices q pour lesquels les valeurs du vecteur $z_i$ sont manquantes.

Nous avons pour tout $z_i$ appartenant à $\Omega$ : $O_i \cup M_i = Q$

Notons : d : nombre total des données manquantes.

Soit $X(I, J)_{(n,m)}$ le tableau de codage disjonctif "complet" (Jambu 1978)(**nous effectuons le codage sur les valeurs observées seulement**) du tableau de modalités $Z(I,Q)_{(n,p)}$.

$$\forall\ j \in Jq \qquad \begin{cases} x_i^{q(j)} = 1 \text{ si l'individu possède la modalité j} \\ x_i^{q(j)} = 0 \text{ sinon} \end{cases}$$

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

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

**Exemple** :

Soit le tableau représentant 5 individus {1, 2, 3, 4, 5} décrits par deux variables qualitatives $v_1$ à 3 modalités $\{m_1^1, m_2^1, m_3^1\}$ et $v_2$ à 3 modalités $\{m_1^2, m_2^2, m_3^2\}$. Le tableau 1 représente le tableau initial et le tableau 2 représente le tableau de codage.

| | $v_1$ | $v_2$ |
|---|---|---|
| 1 | 1 | 2 |
| 2 | 1 | 3 |
| 3 | 3 | ? |
| 4 | 2 | 1 |
| 5 | ? | 2 |

Tableau 1 : tableau initial.

| | $m_1^1$ | $m_2^1$ | $m_3^1$ | $m_1^2$ | $m_2^2$ | $m_3^2$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| 2 | 1 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 1 | ? | ? | ? |
| 4 | 0 | 1 | 0 | 1 | 0 | 0 |
| 5 | ? | ? | ? | 0 | 1 | 0 |

Tableau 2 : tableau de codage.

Soit $x_i$ un élément de X(I, J). Il est convenable d'écrire $x_i = (x_i^o, x_i^d)$ où

$x_i^o$ représente les valeurs observées.

$x_i^d$ représente les valeurs non observées.

## 2.2 HYPOTHESES SUR LES DONNEES MANQUANTES

Dans les cas des données manquantes, le calcul de la vraisemblance est très difficile. Il y a $b = \prod_{q \in Q} (m_q)^{d_q}$ (où $d_q$ correspond au nombre de valeurs non observées par la variable q) façons de reconstituer les données manquantes, il en découle alors tout un ensemble de tableaux de données possibles $X_\theta$, noté H.

$$\mathbf{H} = \{X_\theta \,/\, \theta \in \{1, ..., b\}$$

Nous supposons que les données manquantes sont de type DMH (Données Manquantes au Hasard) et indépendantes. Chaque valeur non observée du tableau X ne pouvant appartenir qu'à {0, 1} avec la contrainte que la somme sur l'ensemble des modalités soit égale à 1, nous supposons que les données manquantes de chaque variable q suivent une distribution multinomiale de paramètre $\alpha^{q(j)}$ qui est supposé **connu** à priori.

Nous avons alors : $f(x_i^{q(j)} = 1) = \alpha^{q(j)} \quad \forall\ j \in Jq \quad (q \in M_i)$

Nous notons : $\alpha = (\alpha^{q(j)};\ j \in Jq;\ q \in Q)$

Le critère de vraisemblance classifiante W(p, a) peut être considéré comme une variable aléatoire de **H** dans **R** et son espérance existe. Comme précédemment, nous proposons alors de remplacer W(P, a) par son espérance E[ W( P, X, a, $\alpha$)].

## 2.3 EXPRESSION DU CRITERE

La proposition suivante donne l'expression du critère dans notre cas.

**proposition 1:**

$$E[W(P, X, a, \alpha)] = \sum_{i=1}^{K} \sum_{i \in P_k} \{ \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} Log(a_k^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} Log(a_k^{q(j)}) \}$$

**preuve :**

$$\begin{aligned}
E[W(P, X, a, \alpha)] &= E[\sum_{k=1}^{K} Log \prod_{i \in P_k} f(x_i ; a_k)] \\
&= E[\sum_{k=1}^{K} \sum_{i \in P_k} (Log\ f(x_i ; a_k))] \\
&= \sum_{k=1}^{K} \sum_{i \in P_k} E[Log\ f(x_i ; a_k)] \\
&= \sum_{k=1}^{K} \sum_{i \in P_k} E[Log(f(x_i^o, x_i^d ; a_k))] \\
&= \sum_{k=1}^{K} \sum_{i \in P_k} E[Log(f(x_i^o ; a_k) f(x_i^d ; a_k))] \\
&= \sum_{k=1}^{K} \sum_{i \in P_k} (Log\ f(x_i^o ; a_k) + E[Log\ f(x_i^d ; a_k)])
\end{aligned}$$

Explicitons le terme $E[Log\ f(x_i^d ; a_k)]$

$$E[Log\ f(x_i^d ; a_k)] = E[Log \prod_{q \in M_i} \prod_{j \in Jq} f(x_i^{q(j)} ; a_k^{q(j)})]$$

$$= \sum_{q \in M_i} \sum_{j \in Jq} E[Log\ f(x_i^{q(j)} ; a_k^{q(j)})]$$

Explicitons maintenant le terme : $E[Log\ f(x_i^{q(j)} ; a_k^{q(j)})]$

$$E[Log\ f(x_i^{q(j)} ; a_k^j)] = f(x_i^{q(j)} = 1) Log\ f(x_i^{q(j)} = 1 ; a_k^{q(j)}) + f(x_i^{q(j)} = 0) Log\ f(x_i^{q(j)} = 0 ; a_k^{q(j)})$$

$$= \alpha^{q(j)} Log(a_k^{q(j)}) \qquad \text{car } f(x_i^{q(j)} = 0 ; a_k^{q(j)}) = 1$$

Nous en déduisons la forme générale de $E[\text{Log } f(x_i^d ; a_k)]$ :

$$E[\text{Log } f(x_i^d ; a_k)] = \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} \text{Log}(a_k^{q(j)})$$

Explicitons le terme $\text{Log } f(x_i^o ; a_k)$.

Nous avons : $$\text{Log}f(x_i^o ; a_k) = \text{Log}[\prod_{q \in O_i} \prod_{j \in Jq} (a_k^{q(j)})^{x_i^{q(j)}}]$$

$$= \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \text{Log}(a_k^{q(j)})$$

Nous en déduisons l'expression du critère :

$$E[W(P,X,a,\alpha)] = \sum_{i=1}^{K} \sum_{i \in P_k} \{ \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \text{Log}(a_k^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} \text{Log}(a_k^{q(j)}) \} \quad (2.3.1)$$

**Remarque :**

*Dans le cas où toutes les valeurs sont observées, nous avons alors :*

$$\sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} Log(a_k^{q(j)}) = 0 \text{ et } E[W(P, X, a, \alpha)] = W(P,a)$$

*où* $$W(p, a) = \sum_{i=1}^{K} \sum_{i \in P_k} \sum_{q \in Q} \sum_{j \in Jq} x_i^{q(j)} Log(a_k^{q(j)})$$ *(Celeux 1988).*

## 2.4 RECHERCHE DES NOYAUX

Nous allons chercher à calculer les $(a_k ; k = 1, ..., K)$ qui maximisent (2.3.1).

**proposition 2 :**

*Les composantes de $q_k$ élément de $R^m$ qui maximisent $E[W(P, X, a, \alpha)]$ sont définies par : pour tout $k = 1, ..., K$ et $q(j) = 1, ..., m_q$*

$$a_k^{q(j)} = \frac{x_k^{q(j)} + d_k^q \alpha^{q(j)}}{n_k}$$

**Preuve :**

$$E[W(P, X, a, \alpha)] = \sum_{i=1}^{K} \sum_{i \in P_k} \{ \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \text{Log}(a_k^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} \text{Log}(a_k^{q(j)}) \}$$

$$= \sum_{i=1}^{K} \{ \sum_{q \in O_i} \sum_{j \in Jq} x_k^{q(j)} \text{Log}(a_k^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} d_k^q \alpha^{q(j)} \text{Log}(a_k^{q(j)}) \}$$

où $\begin{cases} x_k^{q(j)} = \sum_{i \in P_k} x_i^{q(j)} \\ d_k^q = \text{nombre de données manquantes pour q dans la classe k} \end{cases}$

A partition fixé, la recherche des $a_k$ maximisant $E[W(P, X, a, \alpha)]$ se ramène, du fait que les p lois multinomiales sont mutuellement indépendantes, à la recherche des paramètres : $a_k^q = (a_k^{q(1)}, ..., a_k^{q(m_q)})$ qui maximisent chacun des termes $L_k^q$ où :

$$L_k^q = \sum_{j \in Jq} x_i^{q(j)} \text{Log}(a_k^{q(j)}) + \sum_{j \in Jq} d_k^q \alpha^{q(j)} \text{Log}(a_k^{q(j)}) \qquad \text{sous contrainte } \sum_{j \in Jq} a_k^{q(j)} = 1$$

Ce terme peut s'écrire également d'une manière plus simple :

$$L_k^q = \sum_{j \in Jq} (x_k^{q(j)} + d_k^q \alpha^{q(j)}) \text{Log}(a_k^{q(j)}) \qquad \text{sous contrainte } \sum_{j \in Jq} a_k^{q(j)} = 1$$

Le lagrangien de ce problème s'écrit alors :

$$\text{Lag} = L_k^q - \lambda( \sum_{j \in Jq} a_k^{q(j)} - 1)$$

$$\frac{\partial \text{Lag}}{\partial a_k^{q(j)}} = \frac{x_k^{q(j)} + d_k^q \alpha^{q(j)}}{a_k^{q(j)}} - \lambda$$

$$\frac{\partial \text{Lag}}{\partial a_k^{q(j)}} = 0 \text{ alors } \lambda = \frac{x_k^{q(j)} + d_k^q \alpha^{q(j)}}{a_k^{q(j)}} \text{ pour tout j appartenant à } J_q.$$

De $\sum_{j \in Jq} a_k^{q(j)} = 1$ et $\sum_{j \in Jq} \alpha_k^{q(j)} = 1$ On tire $\lambda = n_k$ où $n_k$ est l'effectif de la classe k.

Et par suite, nous en déduisons que :

$$a_k^{q(j)} = \frac{x_k^{q(j)} + d_k^q \alpha^{q(j)}}{n_k} \qquad \#$$

**Remarques :**

1 *Dans le cas où toutes les valeurs sont observées* $a_k^{q(j)} = \frac{x_k^{q(j)}}{n_k}$ *autrement dit* $a_k$ *est le centre de gravité de* $P_k$.

2 *Les valeurs non observées du tableau de codage disjonctif complet semblent être reconstituées par les paramètres de la distribution associée aux données manquantes.*

## 2.5 METHODE MNDM

La méthode que nous allons proposer n'est autre qu'une version de MNDM dans le cas le plus général appliqué sur un tableau de codage (lorsque $\alpha$ est un vecteur et les paramètres du modèle dépendent des variables et des classes).

**Etape de représentation :** *(recherche des noyaux)*

Nous associons à chaque classe $P_k$ le vecteur $a_k$ défini par :

$$a_k^{q(j)} = \frac{x_k^{q(j)} + d_k^q \alpha^{q(j)}}{n_k} \quad \text{pour tout } j \text{ appartenant à } J_q.$$

**Etape d'affectation :** *(recherche des classes)*

Nous associons chaque individu $x_i$ à la classe $P_k$ dont le noyau est le plus proche. Les classes sont alors définies comme suit :
$\forall\ k = 1, \ldots, K$

$$P_k = \{ x_i / \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \mathrm{Log}(a_k^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} \mathrm{Log}(a_k^{q(j)}) \geq$$

$$\sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \mathrm{Log}(a_m^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha^{q(j)} \mathrm{Log}(a_m^{q(j)}) \quad \forall\ m \neq k \}$$

## 3. LORSQUE LES DONNEES MANQUANTES SUIVENT LE MODELE

Dans ce paragraphe nous rejoignons le chapitre **III** dans lequel les données manquantes suivent le modèle. Ainsi les paramètres $\alpha$ sont inconnus et dépendent des classes et des modalités. Plus formellement nous avons en respectant les notations précédentes :

$$\forall k \in \{1, ..., K\} \quad \forall j \in J_q \quad a_k^{q(j)} = \alpha_k^{q(j)}$$

Nous avons constaté que le problème de la maximisation de l'espérance de la vraisemblance classifiante était difficile à accomplir. Ainsi, nous avons proposé la méthode MNDMIN qui, à partir d'une valeur initiale de $\alpha$, organise une procédure cyclique jusqu'à la convergence. L'extension de cette méthode au cas qualitatif est très simple. En effet, en considérant le modèle des classes latentes, chaque modalité joue le rôle d'une variable binaire et l'estimation des paramètres $\alpha$ dans chaque classe est identique. Elle correspond pour chaque variable au centre de gravité calculé sur les données observées et les données manquantes reconstituées par une estimation de $\alpha$ calculée dans l'étape précédente. Nous nous passerons des démonstrations qui sont exactement celles réalisées dans le chapitre **II**. Dans le paragraphe suivant, nous décrivons rapidement les étapes qui caractérisent la méthode MNDMIN.

### 3.1 METHODE MNDMIN

Elle est initialisée par une partition initiale $P^{(0)}$ et par les centres de gravité $\{ ( \alpha_{k,1}^{(0)}, ..., \alpha_{k,p}^{(0)} ) ; k = 1, ..., K \}$ calculés sur les valeurs observées. L'étape de maximisation est décrite à l'itération (t+1).

**Etape d'estimation :**

En respectant les notations du troisième chapitre, l'espérance de la vraisemblance classifiante conditionnelle s'écrit :

$$E[W(P, X, \alpha) / P^{(t)}, \alpha^{(t)}] = \sum_{k=1}^{K} \sum_{i \in P_k} \Big\{ \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \, \mathrm{Log}\,(\alpha_{k,q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha_{k,q(j)}^{(t)} \, \mathrm{Log}(\alpha_{k,q(j)}) \Big\}$$

Nous allons décrire maintenant les étapes d'affectation et de représentation caractérisant l'étape de maximisation. Cette description se fera à l'itération (t+1).

**Etape de maximisation :**

*Recherche des noyaux*

Nous associons à chaque classe $P_k^{(t)}$ le vecteur $\alpha_k^{(t+1)}$ défini par :

$$\alpha_{k,q(j)}^{(t+1)} = \frac{x_k^{q(j)} + d_k^q \alpha_{k,q(j)}^{(t)}}{n_k} \quad \text{pour tout } j \text{ appartenant à } J_q.$$

*Recherche des classes*

Nous associons chaque individu $x_i$ à la classe $P_k^{(t+1)}$ dont le noyau est le plus proche. Les classes sont alors définies comme suit :

$\forall\ k = 1, \ldots, K$

$$P_k^{(t+1)} = \{ x_i / \sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \mathrm{Log}(\alpha_{k,q(j)}^{(t+1)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha_{k,q(j)}^{(t)} \mathrm{Log}(\alpha_{k,q(j)}^{(t+1)}) \geq$$

$$\sum_{q \in O_i} \sum_{j \in Jq} x_i^{q(j)} \mathrm{Log}(\alpha_m^{q(j)}) + \sum_{q \in M_i} \sum_{j \in Jq} \alpha_k^{q(j)} \mathrm{Log}(\alpha_m^{q(j)}) \quad \forall\ m \neq k \}$$

## 4. METHODE EMDM

Les deux méthodes décrites précédemment consistent en la recherche des classes. Lorsque nous nous intéressons uniquement à l'estimation des paramètres du modèle (les $p_k$ et les $\alpha_k$ ; k = 1, K), la méthode EMDM s'étend facilement au cas des données qualitatives et la description y est identique (voir chapitre IV). Nous rappelons que la recherche des classes, n'est qu'un sous produit de ce problème.

## 5. REMARQUES

Nous avons vu que l'extension de ces méthodes est très simple. Et comme précédemment tout se passe comme si les données manquantes dans le tableau de codage sont reconstituées par le paramètre $\alpha$, supposé connu dans la méthode MNDM et par le paramètre $\alpha$, calculé dans l'étape précédente dans les deux méthodes MNDMIN et EMDM.

Lorsque le paramètre $\alpha$ est supposé connu, nous avons vu dans le chapitre V, comment transformer les données du tableau initial en données définissant la probabilité d'avoir 1. Ainsi, nous avons appliqué la méthode MNDQAN (méthode des nuées dynamiques sur données quantitatives) qui a donné de bons résultats. Nous pouvons procéder de la même manière dans le cas des données qualitatives nominales, en considérant le tableau de codage comme un tableau binaire comportant des valeurs non observées. De cette façon, les données manquantes pour une modalité j de la variable q sont remplacées par $\alpha^{q(j)}$ et, à partir de ce tableau, nous pouvons appliquer la méthode MNDQAN.

Dans ce chapitre nous n'avons pas proposé une extension de la méthode MNDMRE pour la simple raison que nous ne disposons pas d'un critère numérique défini par une mesure adaptative et par conséquent d'un modèle probabiliste. Le chapitre VII est consacré à ce problème.

# CHAPITRE VII

# EXTENSION DE LA METHODE MNDMRE

## INTRODUCTION

Dans le chapitre précédent, nous avons constaté que dans la méthode MNDMIN le processus est tel que les données manquantes sont reconstituées à chaque itération. Dans le deuxième chapitre, les liens existant entre le modèle des classes latentes et le modèle associé aux données binaires dans le cas le plus général, nous ont permis d'élaborer la méthode de classification avec reconstitution des données manquantes (MNDMRE). Le critère utilisé est défini par une mesure de dissimilarité variable et dépendant de chaque classe. Cette mesure définit un critère métrique obtenu à partir du modèle de Bernoulli dans le cas le plus général. Dans ce chapitre notre démarche est réciproque, elle consiste à trouver une mesure variable et dépendant de chaque classe qui définit un critère métrique associé à un critère probabiliste. A partir de cette mesure nous proposons une extension de la méthode MNDMRE.

Pour ceci, nous nous basons exclusivement sur les travaux de Govaret (1989) qui montre sous quelles conditions un critère métrique est associé un critère probabiliste dans le cas continu. Ces travaux sont rappelés et étendus dans le cas discret dans le premier paragraphe. Dans le deuxième paragraphe, nous proposons un modèle associé aux données qualitatives. Dans le troisième paragraphe, nous décrivons l'extension de la méthode MNDMRE.

## 1. CRITERES METRIQUES ET PROBABILITES DANS LE CAS DISCRET

Tout d'abord nous définissons les notions de critères métriques et probabilistes et établissons les liens qui existent entre les deux types de critères. Nous ne donnons ici que les principaux résultats. Nous utilisons le tableau de modalités Z(I, Q).

### 1.1 CRITERE METRIQUE

Les algorithmes de type nuées dynamiques que nous avons élaboré ont été destinés à minimiser le critère de la forme :

$$\sum_{k=1}^{K} \sum_{i \in Pk} D(z_i, a_k)$$

Ce critère qui dépend de la mesure de dissimilarité **D** et l'ensemble des noyaux **L**, est appelé critère métrique. La mesure **D** étant définie de **ΩxL** dans **R** et nous notons ce critère **CM(Ω, L, D)**.

### 1.2 CRITERES METRIQUES EQUIVALENTS

**Définition 1 :**

*Deux critères métriques sont dits équivalents si et seulement s'ils sont définis sur les mêmes ensembles* $\Omega$ *et* $L$ *et s'il existe une bijection* $\phi$ *sur* $R$ *strictement croissante vérifiant*

$$CM(\Omega, L, D_1) = \phi o\, CM(\Omega, L, D_2)$$

*où* $D_1$ *et* $D_2$ *sont les mesures de dissimilarité associées aux deux critères.*

A partir de cette définition nous pouvons déduire immédiatement la proposition suivante :

**Proposition 1 :**

$\forall\ \alpha \in R^+$ *et* $\beta \in R$, *les critères* $CM(\Omega, L, D)$ *et* $CM(\Omega, L, \alpha D + \beta)$ *sont équivalents.*

### 1.3 CRITERE PROBABILISTE

Il s'agit alors de maximiser le critère de vraisemblance classifiante :

$$W(P, a) = \sum_{k=1}^{K} \text{Log } L(P_k, a_k)$$

où a est le p-uplet $(a_1, ..., a_K)$ et $L(P_k, a_k)$ est la vraisemblance du sous-échantillon $P_k$ suivant la loi $f(., a_k)$. Ce critère qui dépend de la famille **F** des distributions de probabilités définies sur **Ω** est appelé critère probabiliste et noté **CP(Ω, L)**.

A partir de ces deux définitions nous pouvons en déduire une relation évidente entre ces deux critères.

### 1.4 CRITERE METRIQUE ASSOCIE A UN CRITERE PROBABILISTE

**proposition 2 :**

$$CP(\Omega, F) = -CM(\Omega, L, D)$$

*où L est l'ensemble de définition des paramètres de la famille F et D est définie par :*

$$\forall z_i \in \Omega, \forall a \in L \quad D(z_i, a) = -Log\, f(z_i, a)$$

Le critère métrique ainsi défini à partir d'un critère probabiliste est appelé critère métrique associé.

### 1.5 CRITERES PROBABILISTES ET METRIQUES EQUIVALENTS

**Définition 2 :**

*Deux critères probabilistes sont équivalents si leurs critères métriques associés sont équivalents.*
*Un critère probabiliste $CP_1$ et un critère métrique $CM_2$ sont équivalents si le critère métrique $CM_1$ associé à $CP_1$ est équivalent au critère métrique $CM_2$.*

Ces propriétés ont été utilisées dans les chapitres précédents. Dans la proposition suivante nous établissons la condition nécessaire et suffisante pour qu'un critère métrique soit associé à un critère probabiliste.

### 1.6 CONDITION POUR QU'UN CRITERE METRIQUE SOIT ASSOCIE A UN CRITERE PROBABILISTE

**Proposition 3 :**

*Un critère métrique $CM(\Omega, L, D)$ est associé à un critère probabiliste si et seulement si*

$$\forall a \in L \qquad \sum_{z_i \in \Omega} e^{-D(z_i, a)} = 1$$

**Preuve :**

En effet nous avons : $D(z_i, a) = - \text{Log } f(z_i, a)$ et $f(z_i, a) = \exp(-D(z_i, a))$. Comme nous avons $\sum_{z_i \in \Omega} f(z_i, a) = 1$ alors :

$$\sum_{z_i \in \Omega} e^{- D(z_i, a)} = 1. \ \#$$

Dans ce qui suit, l'ensemble $\Omega$ est supposé être le produit cartésien d'ensembles $\Omega_1, ..., \Omega_p$ et la distance D peut se "décomposer " en une somme de p termes correspondant aux p ensembles $\Omega_q$.

$$D(z_i, a) = \sum_{q \in Q} D_q(z_i^q, a^q)$$

**proposition 4 :**

*S'il existe un réel positif r tel que pour tout j = 1, ..., p ; les quantités :*

$$s_q = \sum_{z_i^q \in \Omega_q} r^{-Dq(z_i^q, a^q)}$$

*soient indépendantes de $a^q$ alors il existe un critère probabiliste équivalent au critère métrique défini par la fonction D. De plus, ce critère probabiliste correspond à un produit de distributions sur les $\Omega_q$ définies par :*

$$f(z_i^q, a^q) = \frac{1}{s_q} r^{-Dq(z_i^q, a^q)} .$$

*Enfin si $s_q = 1$, alors le critère métrique est associé au critère probabiliste.*

**Preuve :**

En effet, le critère métrique associé à la famille proposée est défini par la fonction D' :

$$\begin{aligned} D'(z_i, a) &= - \text{Log } f(z_i, a) \\ &= - \text{Log} \{ \prod_{q \in Q} f(z_i^q, a^q) \} \\ &= - \sum_{q \in Q} \text{Log } f(z_i^q, a^q) \end{aligned}$$

$$= - \sum_{q \in Q} \text{Log} \frac{1}{s_q} r^{-D_q(z_i^q, a^q)}$$

$$= r \sum_{q \in Q} D_q(z_i^q, a^q) + \sum_{q \in Q} \text{Log}(s_q)$$

$$= r\, D(z_i, a) + \sum_{q \in Q} \text{Log}(s_q)$$

La proposition **1** permet d'affirmer que les critères métriques associés à D et D' sont équivalents.

Lorsque $s_q = 1$ nous avons $f(z_i^q, a^q) = r^{-D_q(z_i^q, a^q)}$ ou $f(z_i^q, a^q) = \exp(-D_q(z_i^q, a^q)\text{Log}(r))$.

Cela entraîne que $\text{Log}(r) D_q(z_i^q, a^q) = -\text{Log}[f(z_i^q, a^q)]$ et comme $r > 1$ alors le critère métrique défini à partir de $D_q(z_i^q, a^q)$ est associé à un critère probabiliste.

## 2. MODELE ASSOCIE AUX DONNEES QUALITATIVES

### 2.1 METHODE MNDDIJ

Cette méthode de classification proposée par Marchetti (1989), respecte le principe d'homogénéité, les noyaux sont des modalités. La distance utilisée D, est égale au nombre de composantes différentes entre deux individus considérés. Son expression sur $\Omega$ est la suivante :

$$\forall (x, y) \in \Omega^2 \qquad D(x, y) = \sum_{q \in Q} \delta^q(x^q, y^q)$$

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

Nous définissons l'espace **Y** réduit aux seuls vecteurs binaires de modalités. Les éléments de **Y** résultent du codage des vecteurs de modalités de l'espace $\Omega$. En munissant l'espace **Y** de la distance en valeurs absolues **d**, nous constatons un lien entre d et D. En effet, en considérant deux individus $z_1$ et $z_2$ de l'espace $\Omega$, de codage respectifs $x_1$ et $x_2$ nous avons :

$$d(x_1, x_2) = 2D(z_1, z_2)$$

Dans ce qui suit, nous utilisons la distance D qui nécessite pas le codage du tableau de modalités. Nous savons d'après ce qui précède que les critères métriques associés aux

distances d et D sont équivalents et par conséquent leurs critères probabilistes sont équivalents.

Rappelons que notre but est de proposer une mesure de dissimilarité dont les coefficients de pondérations dépendent des classes et des variables.

## 2.2 MESURE VARIABLE ET DEPENDANT DE CHAQUE CLASSE

Nous supposons que la métrique dépend de chaque classe.

$$D(z_i, a_k) = \sum_{q \in Q} \lambda_k^q \delta^q(z_i^q, a_k^q) + \beta_k$$

Pour que ce critère soit associé à un critère probabiliste, nous devons avoir la condition de la proposition 3 c'est-à-dire $\sum_{z_i \in \Omega} e^{-D(z_i, a)} = 1$.

$$\begin{aligned}
1 &= \sum_{z_i \in \Omega} e^{-D(z_i, a_k)} \\
&= \exp(-\beta_k) \sum_{z_i \in \Omega} \exp\left(- \sum_{q \in Q} \lambda_k^q \delta^q(z_i^q, a_k^q)\right) \\
&= \exp(-\beta_k) \sum_{z_i \in \Omega} \prod_{q \in Q} \exp\left(- \lambda_k^q \delta^q(z_i^q, a_k^q)\right) \\
&= \exp(-\beta_k) \prod_{q \in Q} \sum_{z_i^q \in \Omega_q} \exp\left(- \lambda_k^q \delta^q(z_i^q, a_k^q)\right) \\
&= \exp(-\beta_k) \prod_{q \in Q} \{1 + (m_q - 1) \exp(- \lambda_k^q)\}
\end{aligned}$$

Nous en déduisons que $\beta_k = \sum_{q \in Q} \text{Log}\{1 + (m_q - 1) \exp(- \lambda_k^q)\}$.

La proposition 4 permet d'affirmer que le critère métrique défini à partir de D est associé au critère probabiliste correspondant à un mélange où pour chaque composant, les variables sont indépendantes.

Nous pouvons écrire que :

$$D(z_i, a_k) = \sum_{q \in Q} \lambda_k^q \delta^q(z_i^q, a_k^q) + \sum_{q \in Q} \text{Log}\{1 + (m_q - 1) \exp(- \lambda_k^q)\}$$

Soit $\varepsilon_k^q$ appartenant à ]0, 1/2[. En posant $\varepsilon_k^q = \dfrac{m_q - 1}{\exp(\lambda_k^q) + m_q - 1}$, (les $\varepsilon_k^q$ ne sont pas fixes et dépendent de chaque classe) la distance D s'écrit :

$$D(z_i, a_k) = \sum_{q \in Q} \text{Log}\{ \frac{1-\varepsilon_k^q}{\varepsilon_k^q} (m_q - 1)\} \delta^q(z_i^q, a_k^q) - \sum_{q \in Q} \text{Log}(1 - \varepsilon_k^q)$$

$$= d_{\varepsilon_k}(z_i, a_k) - A_k^q$$

où $d_{\varepsilon_k}(z_i, a_k) = \sum_{q \in Q} \text{Log}\{ \frac{1-\varepsilon_k^q}{\varepsilon_k^q} (m_q - 1)\} \delta^q(z_i^q, a_k^q)$ et $A_k^q = \sum_{q \in Q} \text{Log}(1 - \varepsilon_k^q)$.

**Remarques :**

*1- Dans le cas binaire, $m_q = 2$ et nous retrouvons la mesure de dissimilarité dans le cas le plus général :* $D(z_i, a_k) = \sum_{q \in Q} Log( \frac{1-\varepsilon_k^q}{\varepsilon_k^q} ) |z_i^q - a_k^q| - \sum_{q \in Q} Log(1 - \varepsilon_k^q)$.

*2- Si nous notons $z_i$ et $a_k$ deux individus de l'espace $\Omega$, de codage respectifs $x_i$ et $A_k$ dans l'espace Y nous avons :*

$$D(z_i, a_k) = 2d(x_i, A_k) + (1 - 2m_q) Log(1 - \varepsilon_k^q)$$

*Nous pouvons affirmer que les deux critères métriques sont équivalents.*

Nous pouvons montrer que le modèle correspond, pour chaque variable, à une distribution où les probabilités des $m_q$ modalités de la variable q sont les valeurs :

$$\{1-\varepsilon_k^q, \frac{\varepsilon_k^q}{m_q - 1}, \ldots, \frac{\varepsilon_k^q}{m_q - 1}\}$$

prises dans un ordre quelconque. La modalité prenant la probabilité $1 - \varepsilon_k^q$ étant celle qui est prise par le noyau $a_k^q$. En effet, à partir de la proposition 4 nous avons :

$$p(z_i^q, a_k^q) = \exp(-\delta^q(z_i^q, a_k^q)).$$

Et si $z_i^q = a_k^q$ alors $p(z_i^q, a_k^q) = 1 - \varepsilon_k^q$, sinon $p(z_i^q, a_k^q) = \frac{\varepsilon_k^q}{m_q - 1}$.

## 3. EXTENSION DE LA METHODE MNDMRE

Nous nous proposons ici d'élaborer un algorithme de classification et de reconstitution. Nous avons vu dans le chapitre précédent que tout se passe comme si les données manquantes étaient reconstituées à chaque étape par les paramètres de localisation correspondants calculés dans l'étape précédente. L'algorithme que nous allons décrire est une extension de la méthode MNDMRE dans le cas qualitatif. Le critère à maximiser s'écrit dans ce cas :

$$W(P, a, \varepsilon) = \sum_{k=1}^{K} \sum_{i \in P_k} \{ - \sum_{q \in O_i} \mathrm{Log}(\frac{1-\varepsilon_k^q}{\varepsilon_k^q}(m_q - 1))\delta^q(z_i^q, a_k^q) + \sum_{q \in Q} \mathrm{Log}(1-\varepsilon_k^q)\} \qquad (3.1)$$

**Etape de représentation :** *(recherche des $a_k^q$ et $\varepsilon_k^q$)*

Quelles que soient les valeurs $\varepsilon_k^q$, les $a_k^q$ sont nécessairement les valeurs majoritaires de chaque classe pour chaque variable. Par ailleurs, les $(\varepsilon_k^q,\ k \in \{1, \ldots, K\},\ q \in Q)$ maximisant (3.1) sont les valeurs $\frac{e_k^q}{n_k}$ où $e_k^q = \sum_{i \in P_k} \delta^q(z_i^q, a_k^q)$ exprime le nombre de fois où la valeur majoritaire n'a pas été prise dans la classe k.

**Etape d'affectation :** *(recherche des classes)*

Lors de cette étape le terme $A_k$ n'est pas constant. Nous affectons $z_i$ à la classe $P_k$ qui minimise $d_{\varepsilon_K}(z_i, a_k) - A_k$.

**Etape de reconstitution :** *(reconstitution des données manquantes)*

Le but de cette étape est de reconstituer les données manquantes de chaque variable par les modalités que présente cette variable en maximisant le critère. Il est clair qu'une donnée non observée pour une variable q, de l'individu $z_i$ et se trouvant dans la classe k sera reconstituée par $a_k^q$. En effet, le critère $W(P, a, \varepsilon)$ peut s'écrire :

$$W(P, a, \varepsilon) =$$

$$\sum_{k=1}^{K} \sum_{i \in P_k} \{ - \sum_{q \in O_i} \text{Log}\left(\frac{1-\varepsilon_k^q}{\varepsilon_k^q}(m_q - 1)\right)\delta^q(z_i^q, a_k^q) - \sum_{q \in M_i} \text{Log}\left(\frac{1-\varepsilon_k^q}{\varepsilon_k^q}(m_q - 1)\right)\delta^q(z_i^q, a_k^q)$$

$$+ \sum_{q \in Q} \text{Log}(1-\varepsilon_k^q)\}$$

et les données manquantes qui maximisent $W(P, a, \varepsilon)$ sont celles qui annulent :

$$\sum_{q \in M_i} \text{Log}\left(\frac{1-\varepsilon_k^q}{\varepsilon_k^q}(m_q - 1)\right) \delta^q(z_i^q, a_k^q).$$

Ainsi la décroissance du critère augmente. Cette convergence se démontre de la façon suivante :

Notons **DM** l'ensemble des données manquantes. A l'itération (t) nous avons :

$$W(P^t, L^t, DM^t) \leq W(P^t, L^t, DM^{t-1}).$$

et par suite, à l'itération (t+1) nous avons :

$$W(P^{t+1}, L^{t+1}, DM^t) \leq W(P^t, L^t, DM^t)$$

d'où la convergence de l'algorithme.

**Remarque :**

*Pour retrouver la méthode MNDMRE dans le cas binaire, il suffit de poser* $m_q = 2$.

# CHAPITRE VIII

# EXTENSION DE LA METHODE MNDDIK

## INTRODUCTION

Dans le cas où toutes les valeurs du tableau des modalités sont observées, Marchetti (1989) a proposé une méthode de classification sur les profils des individus en utilisant la métrique du Khi2 et en respectant le principe d'homogénéité c'est-à-dire que les noyaux ont la même structure que les individus. Par ailleurs, Ralambondrainy (1988) a développé la méthode MNDQAL dans laquelle les noyaux des classes sont des centres de gravité.

Dans ce chapitre, à partir de la distance du Khi2 légèrement modifiée, nous proposons un critère métrique associé à un critère probabiliste. Nous étendons ces résultats au cas où certaines valeurs du tableau de données ne sont pas relevées.

Dans le premier paragraphe, nous rappelons la Méthode des Nuées Dynamiques sur un tableau DIsjonctif complet utilisant la métrique du Khi2 (MNDDIK). En nous basant sur les propositions du chapitre précédent, nous montrons que le critère métrique de cette méthode n'est pas associé à un critère probabiliste et nous proposons alors une modification sur la métrique permettant l'association de ces deux critères.

Dans le chapitre VI, le codage des modalités s'est effectué uniquement sur les données observées. Dans le deuxième paragraphe, les données manquantes sont codées par un vecteur dont toutes les composantes sont égales à 0. Le tableau résultant est dit disjonctif "incomplet" et la distance du Khi2 n'a alors plus aucun sens. Ainsi, nous proposons d'utiliser une distance du Khi2 avec "marge imposée" (Escofier 1981) et nous obtenons alors une variante de la méthode MNDDIK où l'une des marges du tableau de données est remplacée par une marge imposée. Appliquée au problème posé, cette variante permet une analyse qui possède la plupart des propriétés de la classification des tableaux disjonctifs complets. En modifiant légèrement cette distance, nous montrons que le critère métrique correspondant est associé à un critère probabiliste.

# 1. METHODE MNDDIK

Dans ce qui suit, sauf indication nous gardons les mêmes notations.

## 1.1 RAPPELS ET NOTATIONS

A partir du tableau disjonctif complet X(I, J) nous définissons les valeurs suivantes :

$$f_{ij} = \frac{x_i^j}{np}$$

$$f_{i.} = \sum_{j \in J} f_{ij} = \frac{1}{np} \sum_{j \in J} x_i^j = \frac{1}{n} \qquad f_{.j} = \sum_{i \in I} f_{ij} = \frac{1}{np} \sum_{i \in I} x_i^j = \frac{n^j}{np} \qquad f = \sum_{i \in I} \sum_{j \in J} f_{ij} = 1$$

où $n^j = \sum_{i \in I} x_i^j$

Nous définissons un ensemble de profils. Un individu i est représenté dans cet ensemble par son profil ligne $\{\frac{f_{ij}}{f_{i.}}, j \in J\}$. Le nuage des individus N(I) est l'ensemble des profils lignes affectés des poids $p_i = f_{i.}$.

## 1.2 DISTANCE DU KHI2

La méthode de classification MNDDIK utilise la distance du Khi2 entre deux profils associés à deux individus i et i'.

$$D^2_{\chi 2}(i, i') = \sum_{j \in J} \frac{1}{f_{.j}} \left( \frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

Cette distance apparaît comme une distance $L_1$, en effet nous pouvons écrire :

$$D^2_{\chi 2}(i, i') = \frac{n}{p} \sum_{j \in J} \frac{1}{n^j} | x_i^j - x_{i'}^j |$$

Dans ce qui suit nous notons D cette distance.

## 1.3 PROBLEME ET METHODE

Il s'agit de déterminer une partition de l'ensemble I des individus en K classes, K étant fixé a priori. Dans cette méthode nous respectons le principe d'homogénéité en imposant aux noyaux d'avoir la même structure que les individus. De cette façon, une

variable est caractérisée, dans une classe, par une modalité. Chaque classe est résumée par un vecteur de modalités. Notons **B** l'espace des vecteurs binaires de modalités.

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* **B** *tels que le critère :*

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

*soit minimum.*

Il nous reste à construire les fonctions d'affectation et de représentation caractérisant l'algorithme MNDDIK.

**Etape d'affectation :**

Nous affectons chaque individu i à la classe $P_k$ dont il est le plus proche au sens de cette distance.

**Etape de représentation :**

La solution est de choisir pour composante égale à 1, celle correspondant à la modalité s de la variable q (notée q(s)) et minimisant la quantité :

$$\frac{n_k - 2n_k^{q(s)}}{n^{q(s)}}$$

où

$n_k = \text{Card}(P_k)$

$n_k^{q(s)} = \sum_{i \in P_k} n_i^{q(s)}$ nombre d'individus dans la classe k présentant la modalité s.

$n^{q(s)} = \sum_{i \in I} n_i^{q(s)}$ nombre d'individus présentant la modalité s.

## 1.4 CRITERE PROBABILISTE

Dans ce paragraphe, l'ensemble $\Omega$ est supposé être le produit cartésien fini $\Omega_1, \ldots, \Omega_p$ et la distance que nous notons dorénavant **D**, peut se "décomposer" en une somme de

p termes correspondant aux p ensembles $\Omega_q$. Chaque ensemble $\Omega_q$ est formé des $m_q$ uplets où $m_q$ est le nombre de modalités de la variable qualitative nominale q.

**Exemple :**

Soit **q** une variable nominale à 3 modalités dans {1, 2, 3}.

$$\Omega_q = \{(1, 0, 0) ; (0, 1, 0) ; (0, 0, 1)\} \qquad ; \qquad m_q = 3$$

La distance du Khi2 peut s'écrire :

$$D(i, a_k) = \sum_{j \in J} \alpha^j \mid x_i^j - x_i^j \mid \quad \text{avec } \alpha^j = \frac{n}{pn^j}$$

ou encore

$$D(i, a_k) = \sum_{q \in Q} \sum_{j \in Jq} \alpha^{q(j)} \mid x_i^{q(j)} - a_k^{q(j)} \mid$$

où q(j) est l'indice de J correspondant à la modalité j.

nous pouvons écrire aussi :

$$D(i, a_k) = \sum_{q \in Q} D_q(x_i^q, a_k^q) \quad \text{où } D_q(x_i^q, a_k^q) = \sum_{j \in Jq} \alpha^{q(j)} \mid x_i^{q(j)} - a_k^{q(j)} \mid$$

posons $s_q = \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q))$

$s_q$ dépend de $a_k^q$ et les conditions de la proposition 2 dans IV ne sont pas vérifiées. Le critère métrique correspondant à cette métrique n'est donc pas équivalent à un critère probabiliste.

En utilisant la proposition ci-dessous nous pouvons proposer une nouvelle métrique voisine de la précédente qui elle sera associée à un critère probabiliste.

**proposition 1 :**

*Etant donné D de $\Omega$xL dans R, le critère métrique CM($\Omega$, L, D') où D' vérifie*

$$D'(i, a_k) = \sum_{q \in Q} \left\{ D(x_i^q, a_k^q) + Log\left( \sum_{x_i^q \in \Omega_q} exp(-D_q(x_i^q, a_k^q)) \right) \right\}$$

*est associé au critère probabiliste défini par la distribution* $p(i, a_k) = exp(-D'(i, a_k))$

**Preuve :**

Le terme ajouté à la distance D, qui n'intervient que lorsque nous affectons un point à une classe, joue un rôle de normalisation des différentes distributions définies à partir des K noyaux. Ce terme aura un effet correcteur et de cette façon la proposition 1 dans chapitre VII est vérifié. En effet, nous aurons pour tout i $D'(i, a_k) = -\text{Log } p(i, a_k)$. #

Nous pouvons proposer ainsi une nouvelle distance voisine de la précédente qui elle par contre, sera associée à un critère probabiliste.

$\forall\ i \in \Omega, \forall\ a_k \in L$

$$D'(i, a_k) = \sum_{q \in Q} \left\{ D_q(x_i^q, a_k^q) + \text{Log} \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q)) \right\}$$

$$D'(i, a_k) = \sum_{q \in Q} D_q(x_i^q, a_k^q) + \zeta_k^q$$

$$\text{où } \zeta_k^q = \sum_{q \in Q} \text{Log} \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q))$$

Le critère métrique défini à partir de cette nouvelle distance D' est associé au critère probabiliste défini par la distribution $p(i, a) = \exp(-D'(i, a))$.

$$p(i, a_k) = \exp(-D(i, a_k)) \left\{ \prod_{q \in Q} \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q)) \right\}^{-1}$$

**Exemple :**

Prenons le cas d'une variable à 3 modalités. Si nous notons $\beta$ l'expression suivante :

$$\beta = \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q)),$$

les distributions de probabilités qui correspondent aux trois vecteurs de modalités possibles de $a_k^q$ sont les suivantes. Par commodité d'écriture nous notons les vecteurs de modalités par leur modalité correspondante.

$a_k^q = 1$

$p(1,1)=1/\beta$ $\quad p(2,1)=e^{-(\alpha^{q(1)}+\alpha^{q(2)})}/\beta$ $\quad p(3,1)=e^{-(\alpha^{q(1)}+\alpha^{q(3)})}/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(1)}+ \alpha^{q(2)})} + e^{-(\alpha^{q(1)}+ \alpha^{q(3)})}$$

$a_k^q = 2$

$p(1,2)=e^{-(\alpha^{q(1)}+\alpha^{q(2)})}/\beta$ $\quad p(2,2)=1/\beta$ $\quad p(3,2)=e^{-(\alpha^{q(2)}+\alpha^{q(3)})}/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(1)}+ \alpha^{q(2)})} + e^{-(\alpha^{q(2)}+ \alpha^{q(3)})}$$

$a_k^q = 3$

$p(1,3)=e^{-(\alpha^{q(1)}+\alpha^{q(3)})}/\beta$ $\quad p(2,3)=e^{-(\alpha^{q(2)}+\alpha^{q(3)})}/\beta$ $\quad p(3,3)=1/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(2)}+ \alpha^{q(3)})} + e^{-(\alpha^{q(1)}+ \alpha^{q(3)})}$$

Nous pouvons dire que le modèle correspond pour chaque variable à une distribution où les probabilités des $m_q$ modalités de la variable q sont définies par :

$$p(x_i^q, a_k^q) = \exp(-D_q(x_i^q, a_k^q)) / \sum_{x_i^q \in \Omega_q} \exp(-D_q(x_i^q, a_k^q))$$

## 2. DONNEES MANQUANTES

### 2.1 PROBLEME DES NON REPONSES

Soit $Z'(I, Q)$ le tableau de modalités avec données manquantes. $X'(I, J)_{(n, m)}$ est le tableau de codage disjonctif "incomplet" du tableau de modalités $Z'(I, Q)_{(n, p)}$. Les données manquantes associées à la variable q sont codées **zéro** sur l'ensemble des modalités $J_q$ de cette question.

**Exemple :**

Soit I un ensemble de 5 individus décrits par une variable nominale **q** à 3 modalités dans $\{m_1, m_2, m_3\}$. Le tableau des modalités comportant des données manquantes est

présenté dans le tableau 1. Le tableau de codage disjonctif "incomplet" associé est présenté dans le tableau 2.

| | q |
|---|---|
| 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | ? |
| 5 | ? |

Tableau 1 : tableau initial.

| | m1 | m2 | m3 |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 0 | 1 |
| 4 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 |

Tableau 2 : tableau de codage.

## 2.2 INCONVENIENT DE LA DISTANCE DU KHI2

Dans un tableau disjonctif incomplet, la distance du Khi2 n'a plus aucun sens car si deux individus i et i' n'ont pas donné le même nombre de réponses ( $f_{i.} \neq f_{i'.}$) ; une modalité j choisie simultanément par ces individus augmente leur distance car le terme $(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}})$ n'est pas nul ; ce qui pose un réel problème d'interprétation. Cette métrique est donc inadaptée à l'étude du tableau disjonctif incomplet. Cependant, nous pouvons y remédier en remplaçant la marge $f_{i.}$ par la marge constante 1/n partout où elle intervient : profil et poids. Le profil ligne de l'individu devient $\{nf_{ij}, j \in J\}$.

## 2.3 DISTANCE DU KHI2 AVEC MARGE IMPOSEE

la distance du Khi2 avec marge imposée entre deux profils associés à deux individus entre i et i' s'exprime de la façon suivante :

$$D^2_{\chi 2}(i, i') = \sum_{j \in J} \frac{1}{f_{.j}} ( nf_{ij} - nf_{i'j} )^2$$

avec $f_{ij} = \frac{x_i^j}{np\text{-}r}$ où r est le nombre total des données manquantes.

Cette distance peut s'écrire alors :

$$D^2_{\chi 2}(i, i') = n^2 \sum_{j \in J} \frac{np\text{-}r}{n^j} ( \frac{x_i^j}{np\text{-}r} - \frac{x_{i'}^j}{np\text{-}r} )^2$$

ou

$$D^2_{\chi 2}(i, i') = \frac{n^2}{np\text{-}r} \sum_{j \in J} \frac{1}{n^j} | x_i^j - x_{i'}^j |$$

La distance du Khi2 avec marge imposée entre deux profils, apparaît alors comme une distance $L_1$ pondérée. Dans ce qui suit, nous notons cette distance **d.**

## 2.4 PROBLEME ET METHODE

Si **B** est l'espace des vecteurs binaires de modalités, les profils des noyaux devront être de la forme $\frac{n}{np\text{-}r}$ a où a est un vecteur binaire de modalités de **B.**

Le problème à résoudre est alors 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 B tels que le critère :*

$$W(P, L_a) = \sum_{k=1}^{K} \sum_{i \in P_k} p_i d(i, a_k) \text{ soit minimum.}$$

L'expression du critère à minimiser est alors la suivante :

$$W(P, L_a) = \frac{n}{np\text{-}r} \sum_{k=1}^{K} \sum_{i \in P_k} \sum_{j \in J} \frac{1}{n^j} | x_i^j - a_k^j |$$

Il nous reste à construire les fonctions d'affectation et de représentation caractérisant cette variante de méthode MNDDIK que nous nommerons MNDKIDM.

**Etape d'affectation :**
Elle est définie à partir de la distance du Khi2 avec marge imposée.

**Etape de représentation :**
Comme dans le cas précédent La solution est de choisir pour composante égale à 1, celle correspondant à la modalité s de la variable q (notée q(s)) et minimisant la quantité :

$$\frac{n_k - 2n_k^{q(s)}}{n^{q(s)}}$$

où $n_k = \text{Card}(P_k)$

$n_k^{q(s)} = \sum_{i \in P_k} n_i^{q(s)}$ nombre d'individus dans la classe k présentant la modalité s.

$n^{q(s)} = \sum_{i \in I} n_i^{q(s)}$ nombre d'individus présentant la modalité s.

Dans l'expression du critère il apparaît que seules les composantes différentes de celles du noyau ont une contribution non nulle au critère. Nous notons que les données manquantes ont toujours une contribution non nulle.

**Exemple :**

Soit un tableau de données croisant 10 individus {**1**, ..., **10**} et 5 variables qualitatives nominales {**a**, ..., **e**} (tableau 3). Sur ce tableau nous appliquons la méthode MNDDIK en demandant 3 classes et la partition obtenue est présentée dans le tableau 4.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 2 | 2 | 2 |
| **2** | 1 | 1 | 2 | 1 | 2 |
| **3** | 2 | 1 | 1 | 1 | 1 |
| **4** | 3 | 3 | 1 | 1 | 2 |
| **5** | 2 | 3 | 3 | 1 | 2 |
| **6** | 1 | 2 | 2 | 2 | 2 |
| **7** | 2 | 2 | 2 | 1 | 1 |
| **8** | 1 | 2 | 2 | 2 | 1 |
| **9** | 1 | 3 | 1 | 3 | 2 |
| **10** | 3 | 3 | 3 | 3 | 2 |

Tableau 3 : tableau initial.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| **1** | 1 | 1 | 2 | 2 | 2 |
| **2** | 1 | 1 | 2 | 1 | 2 |
| **6** | 1 | 2 | 2 | 2 | 2 |
| **3** | 2 | 1 | 1 | 1 | 1 |
| **7** | 2 | 2 | 2 | 1 | 1 |
| **8** | 1 | 2 | 2 | 2 | 1 |
| **4** | 3 | 3 | 1 | 1 | 2 |
| **5** | 2 | 3 | 3 | 1 | 2 |
| **9** | 1 | 3 | 1 | 3 | 2 |
| **10** | 3 | 3 | 3 | 3 | 2 |

Tableau 4 : partition obtenue.

Nous détruisons au hasard 22% de données (tableau 5). Dans le tableau 6 nous présentons la partition obtenue par la méthode MNDKIDM.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| **1** | 1 | 1 | ? | 2 | 2 |
| **2** | 1 | ? | 2 | ? | 2 |
| **3** | 2 | ? | ? | 1 | 1 |
| **4** | 3 | 3 | 1 | ? | 2 |
| **5** | 2 | 3 | 3 | 1 | ? |
| **6** | 1 | 2 | 2 | 2 | 2 |
| **7** | 2 | 2 | 2 | 1 | ? |
| **8** | 1 | 2 | 2 | ? | 1 |
| **9** | ? | 3 | 1 | 3 | 2 |
| **10** | 3 | 3 | ? | 3 | 2 |

Tableau 5 : tableau initial.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| **1** | 1 | 1 | ? | 2 | 2 |
| **2** | 1 | ? | 2 | ? | 2 |
| **6** | 1 | 2 | 2 | 2 | 2 |
| **3** | 2 | ? | ? | 1 | 1 |
| **5** | 2 | 3 | 3 | 1 | ? |
| **7** | 2 | 2 | 2 | 1 | ? |
| **8** | 1 | 2 | 2 | ? | 1 |
| **4** | 3 | 3 | 1 | ? | 2 |
| **9** | ? | 3 | 1 | 3 | 2 |
| **10** | 3 | 3 | ? | 3 | 2 |

Tableau 6 : partition obtenue.

Nous constatons que seul l'individu 5 a quitté sa classe.

**Remarque :**

*Dans la méthode MNDKIDM, nous notons que les données manquantes jouent un rôle passif. Nous pouvons comparer cette méthode à la méthode MNDM lorsque $\alpha = 1/2$. En effet, la distance utilisée peut être remplacée par une "pseudo-distance" du Khi2. Cette "pseudo-distance" entre deux vecteurs exprime des écarts pondérés calculés sur les composantes relevées simultanément. Lors de l'étape de représentation, nous avons noté que les noyaux sont calculés uniquement sur les données observées.*

### 2.5 CRITERE PROBABILISTE

Dans ce cas chaque ensemble $\Omega_q$ est formé des $m_q+1$ uplets. En effet, pour une variable nominale à 3 modalités qui présente des données manquantes l'ensemble $\Omega_q$ :

$$\Omega_q = \{(1, 0, 0) ; (0, 1, 0) ; (0, 0, 1) ; (0, 0, 0)\}$$

Rappelons que la distance entre deux profils s'exprime de la façon suivante :

$$\mathbf{d(i, i')} = \frac{n^2}{np\text{-}r} \sum_{j \in J} \frac{1}{n^j} | x_i^j - x_{i'}^j |$$

$$= \sum_{j \in J} \alpha^j | x_i^j - x_{i'}^j | \qquad \text{avec } \alpha^j = \frac{n^2}{(np\text{-}r)n^j}$$

Nous sommes exactement dans la situation où toutes les données sont observées. En effet, seul l'ensemble $\Omega_q$ change. Le critère métrique correspondant à cette métrique n'est donc pas équivalent à un critère probabiliste. En utilisant la proposition 1, nous pouvons proposer comme précédemment une métrique voisine de la précédente dont le critère métrique correspondant sera associé à un critère probabiliste.

$$\mathbf{d(i, a_k)} = \sum_{q \in Q} d_q(x_i^q, a_k^q) + \zeta_k^q \qquad \text{avec} \qquad \zeta_k^q = \sum_{q \in Q} \text{Log} \sum_{x_i^q \in \Omega_q} \exp(-d_q(x_i^q, a_k^q))$$

**Exemple :**

Prenons le cas d'une variable à 3 modalités présentant au moins une donnée manquante. Si nous notons $\beta$ l'expression suivante :

$$\beta = \sum_{x_i^q \in \Omega_q} \exp(-d_q(x_i^q, a_k^q)) ,$$

les distributions de probabilités qui correspondent aux trois vecteurs de modalités possibles et au vecteur (0, 0, 0) qui définit qu'aucune modalité n'est observée, sont les suivantes.
Par commodité d'écriture nous notons les vecteurs de modalités par leur modalité correspondante et le vecteur (0, 0, 0) par "*".

$a_k^q = 1$

$p(1,1)=1/\beta \quad p(2,1)=e^{-(\alpha^{q(1)}+\alpha^{q(2)})}/\beta \quad p(3,1)=e^{-(\alpha^{q(1)}+\alpha^{q(3)})}/\beta \quad p(*,1)=e^{-\alpha^{q(1)}}/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(1)}+ \alpha^{q(2)})} + e^{-(\alpha^{q(1)}+ \alpha^{q(3)})} + e^{-\alpha^{q(1)}}$$

$a_k^q = 2$

$p(1,2)=e^{-(\alpha^{q(1)}+\alpha^{q(2)})}/\beta \quad p(2,2)=1/\beta \quad p(3,2)=e^{-(\alpha^{q(2)}+\alpha^{q(3)})}/\beta \quad p(*,2)=e^{-\alpha^{q(2)}}/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(1)}+ \alpha^{q(2)})} + e^{-(\alpha^{q(2)}+ \alpha^{q(3)})} + e^{-\alpha^{q(2)}}$$

$a_k^q = 3$

$p(1,3)=e^{-(\alpha^{q(1)}+\alpha^{q(3)})}/\beta \quad p(2,3)=e^{-(\alpha^{q(2)}+\alpha^{q(3)})}/\beta \quad p(3,3)=1/\beta \quad p(*,3)=e^{-\alpha^{q(3)}}/\beta$

$$\text{avec } \beta = 1+ e^{-(\alpha^{q(1)}+ \alpha^{q(3)})} + e^{-(\alpha^{q(2)}+ \alpha^{q(3)})} + e^{-\alpha^{q(3)}}$$

Nous pouvons dire que le modèle correspond pour chaque variable à une distribution où les probabilités des $m_q$ modalités et la donnée manquante (qui pourra être considérée comme une modalité définissant la non présence d'aucune modalité de la variable q) sont définies par :

$$p(x_i^q, a_k^q) = \exp(-d_q(x_i^q, a_k^q)) / \sum_{x_i^q \in \Omega_q} \exp(-d_q(x_i^q, a_k^q))$$

**Remarque :**

*Nous constatons que la probabilité associée à la donnée non observée est la plus proche de celle associée à la modalité choisie. Nous pouvons penser à réaliser une variante de cette méthode qui comporte une étape de reconstitution. Celle-ci consistera à reconstituer les vecteurs binaires non observés par les vecteurs binaires des noyaux correspondants.*

# CONCLUSION

# CONCLUSION

Dans ce travail, nous avons traité du problème de l'estimation des paramètres du modèle de mélanges finis en présence de données manquantes sous l'approche estimation et sous l'approche classification. Ainsi, pour estimer ces paramètres, nous avons utilisé respectivement des algorithmes de type EM et des algorithmes de type nuées dynamiques. Notons toutefois que la méthode MNDMIN développée dans le chapitre III, bien qu'elle soit basée essentiellement sur le processus de l'algorithme EM, peut être considérée comme une méthode des nuées dynamiques. En effet, nous reconnaissons les deux fonctions d'affectation et de représentation dans l'étape de maximisation. Ainsi, cette méthode ne s'attaque pas directement à l'estimation des paramètres du mélange.

En faisant une hypothèse sur la distribution des données manquantes, nous avons constaté que la méthode MNDM avec ses trois variantes et lorsque $\alpha = 1/2$ donne des résultats très satisfaisants et que la convergence de l'algorithme est rapide. Il est en de même pour la méthode MNDKIDM dans le cas des données qualitatives nominales. Nous avons aussi montré comment transformer un tableau de données binaires comportant des données manquantes en un tableau complet de probabilités. Ainsi, nous avons appliqué la méthode MNDQAN et avons constaté que les résultats sont meilleurs que ceux obtenus par la méthode MNDM. Notons toutefois que cette méthode doit être utilisée avec prudence car elle a tendance à donner des classes de même volume.

Sous l'hypothèse que les données manquantes suivent le modèle, nous avons remarqué que les méthodes MNDMIN et EMDM donnent de bons résultats. Dans le cadre de la reconstitution des données non observées, la méthode MNDMRE se comporte de manière satisfaisante tant que le pourcentage des données manquantes est raisonnable.

Ces méthodes ont été appliquées sur des données simulées et réelles qui ont subi des destructions de façon aléatoire et progressive. Ainsi, nous avons pu juger leur qualité en étudiant l'évolution des pourcentages d'objets mal classés. Nous avons constaté que ces objets présentent souvent des données manquantes pour les variables les plus

discriminantes mais montrent une faible proportion de données manquantes pour les variables les moins significatives.

Nous avons également noté que toutes ces méthodes pouvaient s'étendre facilement au cas des données qualitatives nominales. De plus, nous pouvons noter que les critères numériques des méthodes proposées sont associés à des critères probabilistes permettant ainsi d'apporter une interprétation de ces méthodes de classification.

Ainsi, ces nouvelles méthodes, bien que ne prétendant pas être les méthodes absolues dans le cadre des données binaires et qualitatives nominales en présence de données manquantes, permettent cependant de classifier l'ensemble des individus de manière stable avec ou sans reconstitution des données non observées. Cette stabilité reste à étudier rigoureusement comme le fit Benali en 1985 pour l'analyse factorielle des correspondances multiples en présence de données manquantes.

Toutes les méthodes proposées dans notre travail présentent deux limitations. En effet, le nombre de composants est supposé connu et la solution obtenue dépend de la position initiale de l'algorithme. L'algorithme que nous pourrons envisager d'utiliser est l'algorithme SEM qui répond en grande partie à ces deux limitations. Cet algorithme repose sur un principe de tirage aléatoire des données manquantes suivant une loi conditionnelle aux observations. Pratiquement, il se comporte plus efficacement que EM, et évite notamment les convergences vers des points "selle" ou des "méplats" de la vraisemblance en raison de l'effet apporté par la perturbation aléatoire.

Dans notre travail, nous avons supposé que les données manquantes étaient de type DMH. Il serait intéressant d'étudier le cas où les données manquantes ne sont pas de type DMH. Nous pourrons éventuellement étudier le cas des données censurées pour lequel le mécanisme des données manquantes n'est pas ignoré. Enfin, en utilisant les mélanges gaussiens, une extension aux données quantitatives serait nécessaire.

# BIBLIOGRAPHIE

# BIBLIOGRAPHIE


Afifi A.and Elashoff R.M. 1966. Missing observations in Multivariate statistics I-Review of the litterature. *J. Amer. Statis. Assoc.* **61** : pp 595-604.

Anderson P.O. 1975. Stepwise with missing values. Proceeding of the statistical computing section, pp 78-80.

Anderson T.W. 1957. Maximum likelihood estimates for a multivariate noramal distribution when some observations ara missing. *J. Amer. Statis. Assoc.* **52** : pp 200-203.

Beale E.M.L. and Little R.J.A. 1975. Missing values in multivariate analysis. *J.R.S.S.B.* **37** : pp 129-145.

Benali H. 1985. Stabilité de l'analyse en composantes principales et de l'analyse des correspondances multiples en présence de certains types de perturbations - Méthodes de dépouillement d'enquêtes. Thèse de 3ème cycle. Université de Rennes 1.

Bilght, B.J.N 1970. Estimation from a censored sample for the exponantial family. *Biometrika*, **57** : pp 389-395

Buck, S.F. 1960. Amethod of estimation of missing values in multivariate data suitable for use with an electronic computer. In "*Journal of the royal statistical society* ", series B, **22** : pp 302-307.

Celeux G. 1988. Classification et modèles. *R.S.A.* **4** : pp 43-58.

Celeux G. 1988. Le traitement des données manquantes dans le logiciel Sicla. Rapport de recherche INRIA n° 102.

Celeux G. et Diebolt J. 1985. The SEM algorithm : A probalistic teacher algorithm derived from the mixture problem. C.S.Q vol 2, Issue 1 : pp 73-82.

Celeux G. et Diebolt J. 1986. Etude du comportement asymptotique d'un algorithme d'apprentissage probabiliste pour les mélanges de lois de probabilité. Rapport de recherche INRIA n° 563.

Celeux G. et Diebolt J. 1988. A probalistic teacher algorithm for iterative maximum likelihood estimation. In "*classification and related methods of data Analysis*", Bock H.H., ed. Elsevier Science Publishers B.V.(North-Holland) : pp 617-623.

Celeux G. et Diebolt J. 1988. The EM and the SEM algorithms for mixtures : numérical and statistical aspects. Proceeding of the 7th Franco-Belgium meeting on statistic. Publication des Facultés Universitaires St Louis. Bruxelles.

Celeux G. et Govaert G. 1989. Clustering criteria for discrete data and latent class models. Rapport de recherche INRIA n° 1122.

Celeux G. et Govaert G. 1991. A classification EM algorithm for clustering and two stochastic versions. Rapport de recherche INRIA n° 1364.

Celeux G., Diday E., Govaert G., Lechevallier Y. et Ralambondrainy H. 1989. Classification automatique des données. Dunod.

Christofferson A. 1974. The One Component Model with Incomplete Data. Faculty of Social Sciences. University of Upp sala.

Day N.E. 1969. Estimating the components of a mixture of normal distributions. *Biometrika* **56** : pp 464-474.

Dear R.E. 1975. A principal-Component Missing data Method for Multiple Regression Models. SP-86 System Development Corporation, Santa Monica, California.

Dempster A.P., Laird N.M. et Rubin D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm . *J.R.S.S.B.* **39** : pp 1-38.

Der Megreditchian G. 1988. Problèmes engendrés par les données manquantes dans la pratique statistique. Rapport EERM n° 208.

Diday E. 1975. Classification automatique séquentielle sur grands tableaux. *RAIRO - B1* : pp 29-61.

Diday E. et Govaert G. 1977. Classification avec distances adaptatives. *RAIRO*, **vol 11, n°4** : pp 329-349.

Dixon W.J. 1983. In "*BMDP statistical Sofware*". Dixon W.J., ed. Berkeley : University of California Press.

Efron B. and Morris C. 1975. data analysis using Stein's estimator and its generalizations. *J. Amer. Statis. Assoc.* **70** : pp 311-319.

Escoffier B. 1981. Traitement de questionnaires avec non-réponses et analyse des correspondances avec marge modifiée et analyse multicanonique avec contrainte. Pubication interne IRISA n° 146.

Everitt B. 1981. An introduction to latent variable models. Chapman and Hall.

Facy F.et Lechevallier Y. 1978. Traitement des non réponses et des données manquantes pour des variables qualitatives après classification automatique. *R.S.A* **vol 26, n°4** : pp 39-53.

Fèvre P.Ph. 1980. Nouvelles méthodes de traitement de données quantitatives incomplètes. Thèse de doctorat de 3ème cycle. Université Paris IX Dauphine.

Frane J.W. 1975. Anew BMDP program for the description and estimation of missing data. In "*Proceedings of the statistical computing section of the American statistical association* " pp 110-113.

Friedman H. et Rubin J. 1967. On some invariant criterion for grouping data. JASA. n° 62.

Fuchs C. 1982. Maximum likelihood estimation and model selection in contingency tables with missing data. *J. Amer. Statis. Assoc.* **77** : 270-278.

Gleason T.C. and Stealin R. 1975. A proposal for Handling Mssing data. *Psycometrika.* **40** : pp 229-252.

Goodman L. 1974. Exploratory latent structure models using both identifiable and unidentifiable models. *Biometrika.* **61.**

Govaert G. 1975. Classification Adaptative. Thèse de 3ème cycle. Université Paris VI.

Govaert G. 1989. Classification binaire et modèles. *R.S.A.* **37** : pp 67-81.

Govaert G. 1989. Clustering model and metric with continuous data. In "*Data analysis, learning symbolic and numeric knowledge*". Diday E., ed. New-york : Nova Science Publishers, pp 95-102.

Grundy P.M. 1952. The fitting of grouped truncated and grouped censored normal distributions. *Biometrika.* **39** : pp 252-259.

Haitovsky Y. 1968. Missing data in regression analysis. In "*Journal of the royal statistical society* ", series B, **30** : pp 67-82.

Hartigan J.A 1975. Clusterings algorithms. Wiley.

Hartley H.O. 1958. Maximum likelihood estimation from incomplete data. *Biometrics.* **14** : pp 174-194.

Irwin J.O. 1959. On the estimation of the mean of Poisson distribution with the zero class missing. *Biometrics.* **15** : pp 324-326.

Irwin J.O. 1963. The place of mathematics in medical and biological statistics. *J. R. Statist. Soc.*, A, **126** : pp 1-45.

Jambu M. 1978. Thechniques de classification automatique appliquées à des données de sciences humaines. Thèse de doctorat de 3ème cycle. Université de Paris.

Kalbfleisch J.D. and prentice R.L. 1980. The statistical analysis of failure time data. Wiley.

Lacourly N. 1974. Problèmes Statistiques posés par le dépouillement d'enquêtes alimentaires. Thèse de 3ème cycle. Université Paris VI : pp 223-231.

Lebart L., Morineau A. and Tabard N. 1977. Techniques de la description statistique. Dunod.

Little R.J.A. 1988. A test of missing completely at random for multivariate data with misssing values. *J. Amer. Statis. Assoc.* **vol 83, n° 404** : pp 1198-1202.

Little R.J.A. and Rubin D.B. 1987. Statistical analysis with missing data. Wiley.

Marchetti F. 1989. Contribution à la classification de données binaires et qualitatives. Thèse de doctorat de l'Université. Université de Metz.

Marini M.M, Olsen A.R et Rubin D.B 1980. Maximum likelihood estimation in panel studies with missing data. In *Socioligical Methodology*. San francisco : Jossey-Bass.

McKendrick A.G. 1926. Applications of mathematics to medical problems. *Proc. Edin. Math. Soc.* **44** : pp 98-130.

Nora-chouteau C. 1974. Une méthode de reconstitution et d'analyse de données incomplètes. Thèse de 3ème cycle. Université Paris VI.

Ok-Sakun Y. 1975. Analyse factorielle typologique et lissage typologique. Thèse de 3ème cycle. Université Paris VI.

Orchard T. et Woodbury M.A. 1972. Missing information principle : Theory and Applications. In "*Proceedings of the sixth Berkeley Symposium on Mathematical Statistics and Probability*". Berkeley : University of California Press. **vol 1** : pp 697-715.

Pearson K. 1894. Contribution to the mathematic theory of evolution. *Philo. Trans. Soc.* n° 185.

Ralambondrainy H. 1988. Etude des données qualitatives par les méthodes typologiques. Actes au congrés de l'association française de Marketting. Montpelier.

Rao C.R 1972. Linear statistical Inference and its applications. New york : Wiley.

Redner R.A. et Walker H.F. 1984. Mixture densities, maximum likelihood and the EM algorithm. *SIAM review.* **Vol 26, n° 2** : pp 195-239.

Rubin D.B. 1976. Inference and missing data. *Biometrika* **63** : pp 581-592.

Schroeder A. 1976. Analyse d'un mélange de distribution de probabilité de même type. *R.S.A.* **vol 24, n°1** : pp 39-62 .

Scott A.J. et Symons M.J. 1971. Clustering methods based on likelihood ratio criteria. *Biometrics* **27** : pp 387-397.

Turnbull B.W. 1974. Nonparametric estimation of survivorship function with doubly censored data. *J. Amer. Statist. Assoc.*, **69** : pp 169-173.

Turnbull B.W. 1976. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J. Amer. Statist. Soc.*, B, **38** : pp 290-295.

Van der Heijden P.G.M. et Escofier B. 1988. Multiple corresponding analysis with missing data. Rapport de recherche INRIA n° 902.

Wolfe J.H. 1970. Pattern clustering by multivariate mixture analysis. *Multivar. Behavior. Res.* **5** : pp 329-350.

Wu C.F. 1983. On the convergence of the EM algorithm. *Ann. Statist.* **Vol 11, n° 1** : pp 93-103.

Face à un problème pratique de traitements de données, il arrive souvent qu'un certain nombre desdites données se trouve manquer, et dont l'absence peut être imputable à diverses raisons comme une erreur de saisie ou d'expérimentation ou un refus de réponses. Notre travail a consisté à classifier un ensemble d'individus décrits par des variables binaires ou qualitatives nominales sachant que certaines de ces variables n'ont pas été relevées. Les modèles probabilistes étant notre principal outil pour étudier et proposer des solutions au problème de la classification automatique en présence des données manquantes, nous commençons par rappeler comment la classification peut être vue comme une solution à un problème d'estimation de paramètres d'un modèle de mélanges et comment associer à l'algorithme EM (Estimation, Maximisation) un algorithme CEM (Classification, Estimation, Maximisation). En nous appuyant sur les modèles de Bernoulli et en faisant une hypothèse sur la distribution des données manquantes, nous retenons comme critère, l'espérance de la vraisemblance classifiante. Ensuite, nous utilisons le processus de l'algorithme EM en supposant que les données manquantes suivent le modèle de Bernoulli choisi. De plus l'extension de cet algorithme est étudiée dans ce travail. Nous proposons également une autre approche qui consiste à transformer un tableau de données "incomplet" en un tableau de probabilités "complet". Ainsi, nous pouvons utiliser la méthode des nuées dynamiques sur des données quantitatives. Nous nous sommes aussi intéressés à la reconstitution des données non observées. Toutes les méthodes proposées dans cette thèse ont été programmées et intégrées au logiciel d'analyse de données SICLA (Système Interactif de Classification Automatique, INRIA) et ont été appliquées sur des données simulées et réelles.

**MOTS CLES**

Classification, données binaires, données qualitatives nominales, données manquantes, modèle de mélanges, algorithme EM, algorithme CEM, méthode des nuées dynamiques.