Une nouvelle famille d'indices de dissimilarité pour la MDS

Roger Ngouenet

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Une Nouvelle Famille d'Indices de Dissimilarité Pour la MDS

Roger Ngouenet

N° 2087

Octobre

PROGRAMME 3

Intelligence artificielle,
systèmes cognitifs
et interaction homme-machine

ISSN 0249-6399

Une Nouvelle Famille d'Indices de Dissimilarité Pour la MDS

Roger Ngouenet

Programme 3 — Intelligence artificielle, systèmes cognitifs
et interaction homme-machine
Projet Repco

Rapport de recherche n° 2087 — Octobre — 38 pages

Abstract: Multidimensionnal scaling (MDS) seeks to build points in a metric space from a given proximity data. MDS analyses the proximity data in a way that displays the structure of the distance-like data as a geometrical picture. In this paper, we study the multidimensional scaling algorithm based on individual differences scaling and present two new ideas for transforming scales of dissimilarities. On the other hand and mainly, we evaluate a new family of dissimilarities based on a probabilistic approach through three data sets, and compare final configurations to the the results obtained with other types of dissimilarities.

Key-words: Multidimentionnal scaling (MDS), individual differences scaling, probabilistic approach, transforming.

(Résumé : tsvp)

Unité de recherche INRIA Rennes
IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex (France)
Téléphone : (33) 99 84 71 00 – Télécopie : (33) 99 38 38 32

An Evalution of a New Dissimilarities Family for MDS

Résumé : La Multidimensionnal scaling (MDS) est une méthodologie de l'analyse des données dont le but est une représentation géometrique de l'ensemble des unités de données telle que les distances interpoints sont fonction de la structure de dissimilarité. Dans ce travail, nous passons en revue l'approche exploratoire, en n'oubliant pas certains de ses aspects informatiques. Deux nouvelles idées de transformation de l'échelle des dissimilarités sont proposées et testées.Nous experimentons surtout une nouvelle famille d'indices de dissimilarité fondée sur l'approche probabiliste. Le bien fondé de cette famille est illustré par trois exemples d'applications sur de données recueillies en sciences naturelles et humaines. Les configurations obtenues sont comparés à celles des autres indices.

Mots-clé: Multidimentionnal scaling (MDS), approche exploratoire, approche probabiliste, transformation.

Contents

1 Introduction

Considérées par plusieurs auteurs comme méthodes - spatiales - de base de l'analyse des données - du fait qu'elles excellent dans le traitement des données d'un tableau individus × variables numériques $X = (x_{ij})\ i, j = 1, ..., n$ - les méthodes traditionnelles d'Analyses Factorielles ou en Composantes Principales voient leur application limitée dans de nombreux domaines, et ce pour deux raisons principales:

- La seule structure de condensation la plus adéquate des données expérimentales pour le problème posé par le spécialiste est une matrice de dissimilarités $D = (\delta_{ij})\ i, j = 1, ..., n$,non euclidienne , et ne peut résulter d'un tableau préalable tel que celui mentionné (e.g. matrice de confusion).
- Les variables descriptives sont qualitatives et l'expert specialiste des données souhaite que l'on respecte la représentation naturelle des données traitées. (e.g les séquences d'acides aminés decrivant les protéines)

De plus, l'objet premier de la MDS est le respect des distances mutuelles; par rapport auquel le critère peut être défini de la façon la plus adéquate. Alors que dans l'ACP, par exemple, le but principal consiste à définir par rapport au critère de la variance des axes indépendants et de forte variabilité. Ainsi, la première approche peut mieux convenir à la représentation des molécules en chimie ou à la construction d'une carte en géographie. Remarquons que cette exigence n'est compatible qu'avec un nombre relativement restreint de dimensions (idéalement inférieur ou égal à trois) alors que les méthodes classiques d'analyse factorielles ou en composantes principales aboutissent le plus souvent à un nombre théoriquement plus élevé de facteurs. Cependant, même pour ces dernières méthodes, l'interprétation d'un facteur de rang supérieur à 2 ou 3 devient vite délicate.

Dans ces conditions, le propre des méthodes basées sur les échelles multidimensionnelles - Multidimensional Scaling dans la littérature anglophone - est de proposer des solutions aux difficultés mentionnées ci-dessus en substituant aux contraintes linéaires - des méthodes factorielles - des relations monotones.

Ces méthodes se proposent d'obtenir une représentation condensée d'une structure métrique des données qui soit accessible à l'oeil, et ce , par la construction d'un espace affine de dimension finie $E = IR^m$ où les relations géométriques entre points $M_i, i \in I$ sont fonction de la structure des données traitées (D, I) , où I est l'ensemble des individus, objets ou stimuli.

Nous utiliserons le terme MDS (Multidimensional Scaling) afin d'éviter une éventuelle alimentation de la polémique des termes utilisés dans l'école française; Benzecri (1965) parle de l'Analyse des proximités, Y. Escouffier (1975) utilise le

terme Positionnement Multidimensionnel, d'Aubigny (1989) estime que le terme exact est Codage Multidimensionnel .

Les origines des méthodes de la MDS remontent aux travaux de Hayahsi (1952), R.N. Shepard (1962) et Kruskal [11, 1964]. Le domaine d'application visé est la psychophysique . Ils sont alors suivis par plusieurs articles qui fournissent les résultats faisant la preuve de la possibilité de résoudre le problème par des algorithmes. Citons en particuliers les travaux de J.B Kruskal (1972), Green et Rao (1972), Green ET Wind (1973), Greenacre (1978), Hartman (1979), Kruskal et Wish (1978), F.W Young(1980), V.E Mcgee (1971), J.C Lingoes(1973), L. Guttman (1968), Chang et Caroll(1970), de Leeuw et Heiser (1982), Shiffman et al [19, 1981], Meulman (1986), Desarbo et al [21, 1992].

Pour complèter cette liste, nous renvoyons le lecteur intéressé à l'article de MEAD [17, 1992]. Ces auteurs formulent généralement le problème en termes d'ajustement au sens des moindres carrés.

Plusieurs programmes informatiques appliquent ces méthodes. Citons en particulier - dans l'approche exploratoire - les programes MDSCAL, TORSCA, MINISSA et KYST dans le cas de la MDS non métrique, les progammes INDSCAL, ALSCAL, POLYCON, SMACOF et TSCALE dans le cas de la MDS métrique. Plus précisement si M_{il} et M_{jl} sont les coordonnées de deux points donnés M_i et M_j pour la dimension l de $E = IR^m$ espace des stimuli, et la distance entre ces deux points, lorsqu'on s'interresse uniquement aux représentations euclidiennes, définie par

$$d(M_i, M_j) = (\sum_{l=1}^{m} (M_{il} - M_{jl})^2)^{\frac{1}{2}}$$

alors

- Les méthodes de la MDS non métrique désirent que les distances dans E respectent au mieux la pré-ordonnance initiale au sens suivant
  $\forall\ (i,j)$ et $(i',j') \in I^2\ (i,j) \leq (i',j') \Longrightarrow d(M_i, M_j) \leq d(M_{i'}, M_{j'})$.

- Alors que les méthodes métriques apparient directement les dissimilarités aux distances - dans l'espace des stimuli E -. Ce qui consiste à avoir $d(M_i, M_j) \approx \delta_{ij}$, pour tout (i,j) de I^2 ; au mieux.

Nous nous intéresserons essentiellement à la deuxième approche. Notre première motivation résulte de la famille d'indices similarité probabiliste développée par I.C. Lerman dans le cadre de la méthode de classification par l'Analyse de la Vraissemblance des Liens

Notre travail se limite ici d'une part, à une phase expérimentale - vu que nous avons une base solide de comparaison des résultats (Méthodes factorielles, Classification) - et; d'autre part, aux aspects bibliographiques. Il est organisé

de la manière suivante:

Nous centrons la seconde partie autour de l'élaboration de la famille des indices dont nous nous proposons de tester le bien fondé de l'usage - via des tranformations adéquates - dans le contexte de la MDS. Ces transformations font l'objet d'une méthodologie globale consistant à associer à toute matrice de dissimilarités D , une matrice de dissimilarités euclidienne $\tilde{D}$. Nous présentons alors deux transformations nouvelles qui ont représenté une stimulation pour le développement de cette recherche.

Dans la troisième partie, nous présentons la méthode MDS utilisée et les difficultés rencontrées lors de l'application au cas de nos indices . Nous préciserons de plus quelques critères d' adéquation généralement adoptés dans la littérature de la MDS et les différentes étapes d'une mise en oeuvre de l'algorithme dans la pluspart des méthodes de la MDS. Il s'agit de

- Configuration initiale,
- Calcul des disparités,
- Fonction coût,
- Estimation de la configuration,
- Estimation des poids et tests d'arrêt.

La quatrième partie regroupe tous les exemples réels d'applications sur lesquels nous avons testé nos indices et transformations. Pour ces exemples, nous avons sans cesse établi des comparaisons, afin de nous faire une idée précise du comportement et de la sensibilité de nos indices.

En guise de conclusion, nous évoquons les perspectives d'études qu'il nous semble intéressant d'exploiter ou d'explorer pour développer davantage nos indices et nos transformations.

2 Indices de dissimilarité probabiliste

Toute analyse d'un tableau de données par les méthodes de la MDS passe par une construction préalable - lorsque le tableau mentionné n'en est pas une - d' une matrice de dissimilarités $D = (\delta_{ij})$ ou de similarités dans les cas des méthodes non métriques. L'utilisateur de ces méthodes est donc appelé à faire une "bonne" description des liens entre objets à étudier. La famille d'indice de dissimilarités qui fait l'objet de cet travail résulte de l'indice de similarité probabiliste introduit par I.C. Lerman dans la méthode de classification par l'Analyse de la Vraissemblance des Liens (AVL) . Nous allons maintenant préciser les grandes lignes de sa conception.

2.1 conception de l'indice de la vraissemblance

Le schéma général - Lerman [13, 1981] - conduisant à l'expression de l'indice de similarité $P(\alpha, \beta)$ entre deux éléments ou structures α et β d'un ensemble quelconque se résume dans les étapes suivantes:

- 1) Représentation mathématique appropriée des structures α et β, basée sur une approche ensembliste;
- 2) Choix d'un indice brut de similarité s;
- 3) Définition d'une hypothèse d'absence de lien (h.a.l), qui au couple de structures observées (α, β) associe un couple de structures aléatoires (α^*, β^*) de même type et respectant les caractéristiques cardinales du couple (α, β) ;
- 4) Association à s sous h.a.l, d'une variable aléatoire S;
- 5) Réduction locale; formation de l'indice centre et réduit Q tel que

$$Q(\alpha, \beta) = \frac{s(\alpha, \beta) - E(S)}{\sqrt{V(S)}}$$

où $E(S)$ et $V(S)$ représentent respectivement l'espérance et la variance mathématique de la variable aléatoire S.

- 6) Réduction globale; pour tenir compte du contexte, on rapporte l'association entre α et β à l'ensemble des associations mutuelles. On construit ainsi $\tilde{Q}$ tel que:

$$\tilde{Q}(\alpha, \beta) = \frac{Q(\alpha, \beta) - moy(Q)}{\sqrt{var(Q)}}$$

où $moy(Q)$ et $var(Q)$ désignent respectivement, la moyenne et la variance de la distribution empirique de Q sur l'ensemble des paires des structures observées.

- 7) L'indice de similarité probabiliste définitif qui se réfère à l'échelle de probabilité de la loi normale centrée réduite $N(0, 1)$ est:

$$P(\alpha, \beta) = Pr[\tilde{Q}(\alpha^*, \beta^*) \leq \tilde{Q}(\alpha, \beta)/h.a.l] \approx \Phi(Q(\alpha, \beta))$$

où Φ est la fonction de répartition de $N(0, 1)$.

Outre son intérêt propre, la réduction globale présente un intérêt majeur dans la conception de l'indice puisqu'elle permet de préserver le pouvoir discriminant de l'indice P - Lerman [14, 1986] -. En effet, $Q(\alpha, \beta)$ peut, en valeur absolue, prendre des valeurs élevées et ne permet donc pas une discrimination

suffisante via P . La concentration de P serait soit en 0, soit en 1.

Mentionnons que la construction d'une telle famille dépend du probème posé qui peut bien sûr être celui de la construction des proximités de l'ensemble des objets (individus, concepts); mais qui peut également être celui de l'ensemble des variables de description ou stimuli. Elle dépend aussi du type mathématico-logique du tableau des données (individus ou concepts $\times$ variables) $X = (x_{ij})$ $i = 1, ..., n$, $j = 1, ..., p$, qui peut être quelconque.

2.2 Famille d'indices de dissimilarités probabiliste

Nous disposons au départ de la matrice de similarité probabiliste introduite dans la section précédente. La famille d'indices de dissimilarités probabiliste se conçoit clairement à partir de transformations de nature numérique. Ces dernières confèrent une interprétation propre intéressante aux indices transformés. En plus, elles doivent permettre de se rapprocher du caractère euclidien de la matrice de dissimilarités (travaux futurs).

Limitons nous à ce niveau aux transformations qui ont fait l'objet de nos expérimentations.

$$D_1(i,j) = 1 - P(i,j) \quad (1)$$

$$D_2(i,j) = -Log_2 P(i,j) \quad (2)$$

$$D_3(i,j) = (1 - 2P(i,j))^{\frac{1}{2}} \quad (3)$$

$$D_4(i,j) = \frac{1}{1 - P(i,j)} \quad (4)$$

$$D_5(i,j) = Log_2 \left(\frac{1 + P(i,j)}{1 - P(i,j)} \right), i \neq j \quad (5)$$

Cette famille peut être complètée par toute application transformant une similarité en dissimilarité, mais il est intéressant de noter que l'essentiel est la recherche des transformées euclidiennes . La notion de caractère euclidien d'une matrice des proximités $D = (\delta_{ij})$ est fondamentale dans la littérature de la MDS et fait alors l'objet d'une méthodologie dont nous présentons les grandes lignes dans le paragraphe suivant.

2.3 Caractère euclidien

2.3.1 Préambule

Lorsqu'on se limite aux représentations euclidiennes, la MDS vise à représenter tout individu i d'une structure de données (I, D) - I ensemble des objets et D

matrice des proximités - par un point M_i d'un espace euclidien $E = IR^m$ de telle sorte que la figure obtenue rende compte des liens entre individus, et ceci géométriquement. Ainsi, on dira de la structure des données (I, D) qu'elle est euclidienne si l'image euclidienne qui lui est associée, $\{M_i, i \in I\}$ vérifie :

$$\forall (i,j) \in I^2, D(i,j) = \delta_{ij} \quad (6)$$

$$= \| \overrightarrow{M_i M_j} \| \quad (7)$$

où δ_{ij} représente la dissimilarité entre les objets i et j .

Lorsque la matrice $D = (\delta_{ij})$ n'est pas euclidienne, la nécessité de faire des transformations - très souvent algébriques - s'impose . Bien que ceci soit la méthode généralement adoptée par plusieurs auteurs, Gower [18, 1968] analyse - dans le cadre de l'Analyse en Coordonnées Principales - les matrices non euclidiennes à l'aide de la technique d'ajout de points supplémentaires. Nous nous intéressons essentiellement aux approximations de nature algébrique . Pour estimer la qualité de la transformation utilisée, on se sert quelques fois du théorème de Young et Householder.

- Young-Householder (1938)
  La structure de données (D, I) admet une représentation euclidienne dans un espace euclidien de dimension finie m si et seulement si $W = -\frac{1}{2} J D^2 J$ est une matrice semi-définie positive de rang inferieur à m , où $D^2 = ({\delta_{ij}}^2)_{i,j}$ et $J = I_d - \frac{1}{n} 1^t 1$, où I_d désigne la matrice identité et 1 le vecteur dont toutes les coordonnées sont égales à l'unité.

Contentons-nous, à présent, des transformations fondamentales de la littérature MDS. Rappellons qu'on les retrouve dans la quasi-totalité des programmes informatiques destinées à la MDS.

2.3.2 Constantes additives

Le problème des constantes additives trouve son origine dans les travaux en psychométrie - Torgerson [20, 1952] -. La constante additive est destinée à assurer le caractère semi- défini de la matrice W introduite ci-dessus. Il est présenté en relation avec les contraintes d'échelle de mesure des dissimilarités observées. On définit la matrice

$$\tilde{D} = H + c(1^t 1 - I)$$

de terme générale

$$\tilde{\delta}_{ij} = h_{ij} + c(1 - {\delta_i}^j)$$

où ${\delta_i}^j$ représente le symbole de Kronecker.

Deux cas distincts de techniques de constante additive sont en général retenus dans la littérature de la MDS.

a) Type I On pose $h_{ij} = \delta_{ij}$ et le problème consiste à déterminer la valeur minimale de c de sorte que $\tilde{D}$ soit une dissimilarité euclidienne. La solution initiale proposée par Torgerson (1952) est de choisir la valeur minimale c^* pour laquelle $\tilde{\delta}_{ij}$ vérifie l'inégalité triangulaire. Cependant Schoneman (1971) montre que cette solution assure que $\tilde{D}$ est une distance dans le cas où D est une dissimilarités propre mais n'assure pas son caractère euclidien.

Cailliez [4, 1983] démontre que la constante additive c^* qui entraîne le caractère euclidien, est la plus grande valeur propre de la matrice $2n$-carrée définie par:

$$\begin{pmatrix} 0 & 2W \\ -I & -4W_0 \end{pmatrix}$$

où $W_0 = -\frac{1}{2}JDJ$. W et J sont définies dans la section précédente, n représente la dimension de la matrice des dissimilarités.

b) Type II Ce deuxième type attribué par Lingoes (1971), de Leeuw et Heiser (1982) à Guttman consiste à poser $h_{ij} = {\delta_{ij}}^2$, puis à déterminer la valeur minimale de c qui rende $\tilde{D}$ euclidienne.
Ces techniques de constantes additives ont l'avantage de préserver la préordonnance induite par D et de donner, lorsque la constante c est assez petite un bon apperçu sur la structure métrique de l'ensemble des individus. Notons néanmoins que, dans le cas des indices de dissimilarité sur les variables de présence-absence, l'application ces méthodes devient très delicate. En effet comme le font remarquer B. Fichet et G. Le Calve (1984), cela revient à changer la forme analytique de l'indice initial.
d'Aubigny [5, 1989] dans son modèle algebrique de la MDS considère le problème de la constante additive comme une heuristique.

Evoquons le dernier modèle de constante additive proposé par Hoang M. Thu (1978);

$$\tilde{\delta}_{ij} = {\delta_{ij}}^c + \delta_i^j$$

.

2.3.3 Autres transformations

Le lecteur sera alerté par le fait que dans cette section, les approximations n'ont pour le moment d'autre justification que les résultats obtenus.

a) Puissance euclidienne [Le Calve]
Il s'agit de déterminer $0 < \alpha < 1$ maximal tel que la dissimilarité définie par

$$\forall (i,j) \in I^2, \tilde{D}(i,j) = (1 - \delta_j^i) D(i,j)^\alpha$$

soit euclidienne.

Notons que pour certains indices de dissimilarité tels que ceux calculées sur les variables dichotomiques, on sait trouver une puissance euclidienne.

b) Transformation AVL [Lerman]
Cette transformation est - dans sa construction - conforme avec l'indice de similarité probabiliste introduit plus haut. L'auteur suggère de construire - avant toute éventuelle approximation par la technique de la constante additive - une nouvelle matrice des indices définie par

$$\tilde{D}(i,j) = -Log_2 \left[\phi \left(\frac{-D(i,j) + Moy(D)}{\sqrt{var(D)}} \right) \right]$$

où $Moy(D)$ et $var(D)$ désignent respectivement, la moyenne et la variance empiriques de la suite $\{ D(i,j) , (i,j) \in I^2 \ i \neq j \}$. Lerman signale par ailleurs la possibilité d'itérer une telle transformation .

c) Transformation Z [Ngouenet]
Nous appelons ainsi une approximation qui nous a paru assez intéressante pendant nos expérimentations. En effet la transformation Z consiste à approcher - avant toute éventuelle approximation par la technique de la constante additive - D par la matrice Z de terme génerale

$$Z_{ij}(D) = D^2(i,.) + D^2(.,j) - \frac{1}{\alpha} D^2(.,.)$$

où

$$D^2(i,.) = \frac{1}{n} \sum_{j=1}^{n} D^2(i,j)$$

$$D^2(.,j) = \frac{1}{n} \sum_{i=1}^{n} D^2(i,j)$$

$$D^2(.,.) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} D^2(i,j)$$

n désigne la taille de I .
et où α est un paramètre à fixer. Nous préconisons $\alpha = 2$.

Les deux idées nouvelles de transformation présentées feront, et ceci dans le cadre d'un travail futur, l'objet d'une étude destinée à préciser leurs intérêts respectifs à travers des applications plus approfondies. Il est intéressant de noter

que toutes ces transformations plus ou moins justifiées n'existent que grâce au rôle joué par l'ordinateur dans l'application des méthodes MDS. Ainsi dans la section suivante, nous présentons le programme informatique dont nous nous sommes servis pour la mise en oeuvre de la MDS. Nous commencerons par exprimer dans le préambule ses différentes étapes dans un organigramme. La suite sera consacrée à une description de façon plus précise et plus technique des parties de l'organigramme.

3 ALSCAL

3.1 Préambule

L'Alternating Least-Squares sCALing (ALSCAL) désigne un programme de traitement des données par MDS à l'aide d'un algorithme des moindres carrés alternés. Ses auteurs, Takane, Young et de Leeuw (1977) proposent un principe d'itération généralisant au cadre non linéaire, la procédure de Gauss Seidel dont la flexibilité et l'efficacité dans les problèmes d'estimation linéaire est fondée sur la partition de l'ensemble des paramètres à estimer, que l'on évalue de façon alternée. Pour mieux comprendre ALSCAL considérons l'organigramme suivant:

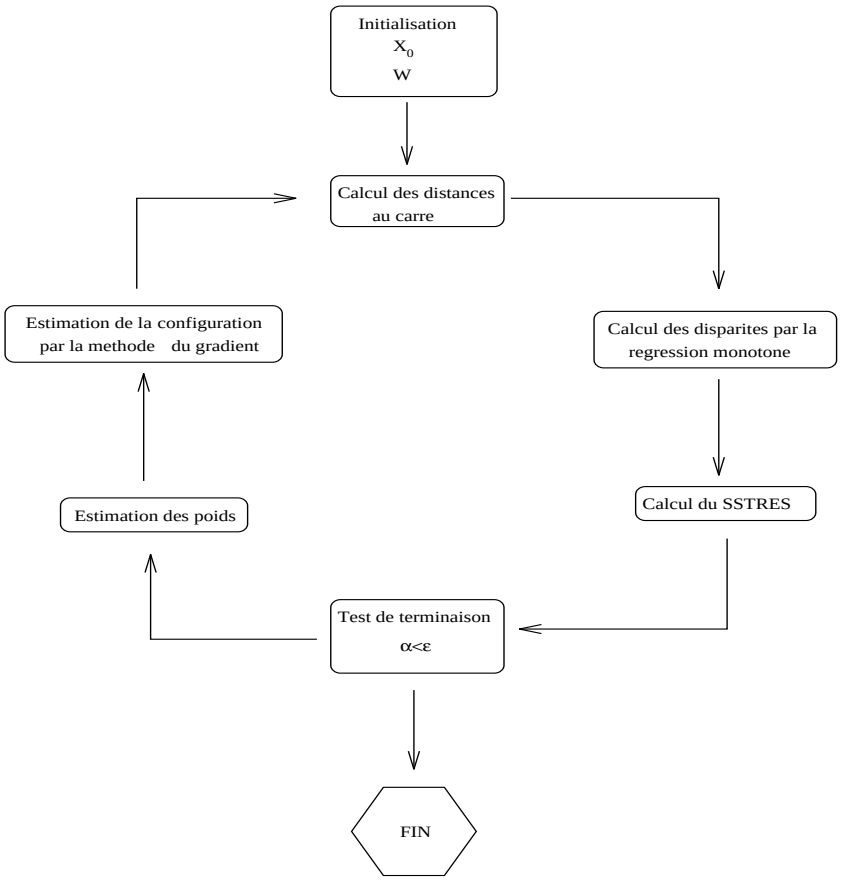

Figure 1: organigramme de ALSCAL

Nous examinerons dans la suite, chacune des étapes de l'agorithme ALSCAL tout en faisant intervenir les aspects intéressants des autres programmes MDS. Insistons une fois de plus sur le fait que ce premier travail restera essentiellement bibliographique et expérimental.

3.2 Configuration initiale

Le problème numérique que posent plusieurs algorithmes traitant de la MDS est celui de la minimisation d'une fonction perte dont - rappellons le - la résolution est faite par des méthodes utilisant des procédés itératifs nécessitant une initialisation.

L'utilisateur des programmes MDS est alors aménè à faire un " bon " choix de la configuration initiale. Ainsi on peut espérer la convergence de l'algorithme

utilisé vers un optimum globale. A cette fin, beaucoup de techniques de construction de la configuration initiale ont été développées. Contentons-nous de mentionner les plus utilisées dans la littérature de la MDS.

La toute première due à Kruskal [10, 1964], consiste - lorsque les données sont normalisées à réaliser un tirage aléatoire de n points selon une loi uniforme dans le cube $([-1,+1])^p$, où p designe la dimension de l'espace de représentation. Cette configuration - d 'après son auteur - permet d'éviter une quantité considérable des itérations inutiles.

La plus utilisée dans les programmes MDS métrique est celle faite à partir de la méthode de Torgerson [20, 1958]. Fondée sur le travail de Young et Householder (1938), elle consiste à effectuer au préalable, un double centrage de la matrice des carrés des indices de dissimilarités observées $\delta^2(i,j)$.

$$W_{ij}(\delta^2) = -\frac{1}{2}(\delta^2(i,j) - \delta^2(.,j) - \delta^2(i,.) + \delta^2(.,.))$$

où

$$\delta^2(i,.) = \frac{1}{n}\sum_{j=1}^{n}\delta^2(i,j)$$

$$\delta^2(.,.) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta^2(i,j)$$

avec n désignant le cardinal de I.

Ensuite, W - dont la forme matricielle est définie plus haut - est interprêté au coefficient (-1/2) près comme une matrice de pseudo produits scalaires. On calcule enfin les vecteurs propres qui sont alors considérés comme des points initiaux du processus de minimisation de la fonction coût.

On a alors $W = X_0^t X_0$ et la configuration initiale considérée est X_0. Rappellons au lecteur que ce W est le même que celui utilisé par Cailliez dans la détermination de la constante additive. Et Guttman (1967) de proposer, ceci dans le cas de la MDS non métrique, une configuration initiale X_0 basée sur les rangs . En effet, il suggère de prendre comme points initiaux au processus de minimisation de la fonction coût, les composantes principales de la matrice C de terme général $c_{i,j}$ suivant, en faisant tout simplement de l'analyse spectrale.

$$c_{i,j} = \begin{cases} 1 - \frac{1}{l}\sum_k \rho_{ik} & (i = j) \\ 1 - \frac{\rho_{ij}}{l} & (i \neq j) \end{cases}$$

où ρ_{ij} désigne le rang de δ_{ij} dans la matrice des dissimilarités initiale et l le rang maximun.

$$l = \frac{n(n-1)}{2}$$

d'Aubigny (1989) présente une technique de construction de configuration initiale inspirée des travaux de Benzecri (1964) et qui suppose l'existence d'une solution représentée par une matrice de configuration gaussienne dans IR^m. En effet, on transforme dans les c_{ij} de Guttman les ρ_{ik} selon la méthode préconisée par Benzecri en fractiles de la loi du χ^2 à $m-1$ d.d.l. (f_{ij}). La matrice de Guttman s'écrit alors

$$c_{i,j} = \begin{cases} 1 - \sum_k f_{ik} & (i = j) \\ 1 - f_{ij} & (i \neq j) \end{cases}$$

Machmouchi[16, 1992] propose une configuration intiale fondée sur l'application des méthodes de " Recuit Simulé" [Kirpatrick (1983)]. L'idée de base de cette méthode est de fournir un algorithme d'exploration probabiliste pas obligatoirement descendante au sens de la minimisation de la fonction objectif, et qui ne se termine pas au premier optimum local rencontré comme les méthodes de gradient.

Notons que Carroll[9, 1989] évalue cinq algorithmes proposés par différents auteurs pour générer la configuration initiale dans SINDSCAL. Il s'agit des programmes COSPA, SUMSCAL, FG, LINCINDS et DAVISON. Ce qui nous permet de constater que la configuration initiale dépend de chaque méthode. C'est ainsi que celle faite à partir de la méthode de Torgerson apparait dans plusieurs méthodes de traitement de la MDS. On la retrouve comme première étape dans POLYCON, dans ALSCAL et INDSCAL comme seconde étape après la constante additive. Dans le cas de nos indices, nous avons testé le bien fondé de la configuration issue de l'A.C.P et de l'A.F.C - lorsque la qualité de représentation est "bonne" - en tant que configuration initiale.

3.3 Disparités

Cette étape qui fait l'originalité de ALSCAL par rapport aux autres méthodes consiste en la recherche d'une transformation g - fonction des δ_{ij} - reproduisant l'information ordinale issue des indices de dissimilarités δ_{ij} et minimisant

$$\sigma_g(X^*) = \sum_{i>j}^{n} q_{ij}(g(\delta_{ij}) - d_{ij}(X^*))^2$$

où X^* désigne la configuration en cours, q_{ij} le poids de pondération du couple (i, j) et $n = \text{Card}(\ I\ )$.

On écrira indifféremment d_{ij} ou $d_{ij}(X)$ pour désigner $d(i, j)$.

La transformation g est un élément d'une classe de fonction IF déterminée par les contraintes de mesurages liées soit au processus de collecte des données (on suppose, par exemple, que g est une fonction monotone par rapport à l'ordre induit par les indices de dissimilarités observés [cf. Kruskal(1964)]), soit au souci d'assurer la stabilité des résultats en retenant des fonctions g dépendant d'un petit nombre de paramètres (on suppose par exemple que g est une fonction polynomiale de degré fixé, [cf. Kruskal (1964)]), soit enfin par un cumul des deux objectifs (on suppose par exemple que f est une fonction spline cubique monotone, cf. Winsberg et Ramsay (1983)). En fait, on peut dire qu'on recherche g^* tel que

$$g^* = Min\{\sigma_g(X^*), .g \in IF\}$$

Dans ALSCAL, les disparités sont obtenues par régression monotone à l'aide de l'algorithme "de blocs" proposé par Kruskal qui procède de manière suivante:

- On ordonne les couples (i, j) suivant les valeurs croissantes (ou décroissantes) des dissimilarités δ_{ij}, $i > j$. Si l'on obtient une suite de distances d_{ij} ordonnée selon l'ordre de dissimilarités, alors il existe une fonction monotone stricte respectant cet ordre. En général, ce n'est pas le cas et l'on doit construire une fonction monotone au sens large. Lorsque l'ordre est violé, on remplace les nombres concernés par leur moyenne. Pour bien comprendre la régression monotone, nous avons construit l'algorithme .

Algorithme MATLAB correspondant

```
function y=disparit(d) /*distances  $d_{ij}$ rang\'es*/

n = length(d);
i = 1;
while i < n,
  if d(i) > d(i+1)
    sum = d(i);
    jd = i+1;
    jg = i;
    j = jd;
    continue = 1;
    while continue   /* recherche des mal classes*/
      sum = sum + d(j);
      moy = sum/(jd-jg+1);
      if     jd < n & moy > d(jd+1)
        jd = jd+1; j = jd;
      elseif jg > 1 & moy < d(jg-1)
```

```
        jg = jg-1; j = jg;
      else
        for k = jg:jd
          d(k) = moy;
        end;
        continue = 0;
      end;
    end;
    i = jd-1;
  end;
  i = i+1;
end;
```

Quant aux méthodes non métriques, les disparités sont généralement obtenues par l'algorithme du rang de Guttman basé sur le même principe que celui de Kruskal.

3.4 Fonction coût: SSTRESS

Lorsqu'on formalise le problème de la MDS en termes d'approximation au sens des moindres carrés, les méthodes concernées se singularisent par le critère d' adéquation - SSTRESS - qui est optimisé. A cette fin, ALSCAL utilise essentiellement les distances au carré - contrairement à l'idée initiale de Kruskal - et sa fonction coût s'écrit:

$$SSTRESS = \left(\frac{\sum_i \sum_j q_{ij} (d_{ij}^2 - d_{ij}^{*2})^2}{\sum_i \sum_j q_{ij} d_{ij}^{*4}} \right)^{\frac{1}{2}}$$

où d_{ij}^* est la disparité correspondant à δ_{ij}.

Pour mieux comprendre la fonction ci-dessus écrite par Takane, Young et de Leeuw revenons à Hayashi (1952) qui, dans sa méthode IV de quantification, propose de construire la solution réalisant le maximun de la fonction critère

$$\theta(X) = \sum_{i<j}^{n} q_{ij} \delta_{ij} d_{ij}(X)^2$$

La maximisation de cette fonction est équivalente à la minimisation - d'Aubigny (1989) - de la fonction

$$HSTRESS = \frac{\sum_i \sum_j q_{ij} (\delta_{ij} - d_{ij}^2(X))^2}{\sum_i \sum_j q_{ij} d_{ij}^4}$$

Le dénominateur de $HSTRESS$ s'interprète comme une contrainte portant sur la variance empirique de la variable $d^2(X)$. Il conduit à neutraliser les effets de variance interpoints lors de la comparaison des configurations admissibles X et doit donc être rapprochée du constat fait par SHEPARD (1962), qui préconise d'accroître la variance des distances interpoints afin de diminuer la dimension de l'espace affine dans lequel est réalisé la représentation euclidienne.

Dans les méthodes se proposant de minimiser les résidus du modèle

$$\delta_{ij} = d_{ij}(X) + e_{ij}$$

au sens d'un critère de moindres carrés pondérés, ALSCAL nous semble le plus approprié.

L'origine de cette formulation du problème de la MDS remonte aux travaux de Kruskal [11, 1964] qui donne une formulation mathématique du problème en termes d'optimisation et un algorithme résolutif convaincant, destinés à rationaliser l'approche heuristique proposée par Shepard (1962).

Cet auteur retient comme mesure de la qualité de l'approximation - STRESS -, la moyenne quadratique des résidus e_{ij} . Cette approche constitue la base des programmes KYST et POLYCON qui traite le problème de la MDS.

$$STRESS = \left(\frac{\sum_i \sum_j q_{ij}(d_{ij} - d^*_{ij})^2}{\sum_i \sum_j q_{ij} d^2_{ij}} \right)^{\frac{1}{2}}$$

où d^*_{ij} représente la disparité correspondant à δ_{ij}.

Originellement, Kruskal [11, 1964] traite le cas d'une pondération triviale q_{ij} = 1 pour tout couple d'objets de I et Guttman (1968) utilise la pondération q_{ij} = 0 si la donnée δ_{ij} est manquante et 1 dans le cas contraire.

Caroll et Chang généralisent la fonction coût proposée par Kruskal et suggèrent de considérer la fonction perte suivante que l'on retrouve dans INDSCAL:

$$STRESS = \left(\frac{\sum_i \sum_j q_{ij}(d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_j q_{ij}(d_{ij} - \bar{d})^2} \right)^{\frac{1}{2}}$$

où $\hat{d}_{ij}$ et $\bar{d}$ désignent respectivement $\delta_{ij} + c^*$(constante additive) et leur moyenne empirique.

Dans l'objectif de faire valoir leur idée, selon laquelle l'utilisation de méthodes de gradient préconisée par de très nombreux auteurs dont Kruskal [11, 1964] n'est pas satisfaisante pour minimiser la fonction STRESS en fonction de

X puisque cette fonction n'est pas différentiable aux points X pour lesquels il existe au moins un couple de points (i , j) - i distinct de j - tel que ${d_{ij}}^2(X) = 0$, Heiser et de Leeuw [8, 1986] proposent dans l'algorithme SMACOF, la fonction coût

$$\sigma(X) = \sum_{i>j}^{n} q_{ij}(\delta_{ij} - d_{ij}(X))^2$$

où X désigne la configuration en cours, et

$$d_{ij}(X) = \left(\sum_{s=1}^{m} (x_{is} - x_{js})^2 \right)^{\frac{1}{2}}$$

m représente la dimension de l'espace.

Dans une méthode récente [21, 1992] basée sur le modèle de Tversky(1977) - dont l'idée est de mesurer la proximité entre deux stimuli en tenant compte à la fois de leur ressemblance et de leur dissemblance - une gamme de fonctions coûts destinée à estimer l' adéquation de la configuration obtenue est proposée. On retrouve ces fonctions dans le programme TSCALE.

La première, Root-Mean-Square (RMS) a pour expression:

$$RMS = \left(\frac{\sum_i^N \sum_j^N \sum_{r=1}^R (\delta_{ijr} - \hat{\delta}_{ijr})}{RN(N-1)} \right)^{\frac{1}{2}}$$

Celle qui me semble la plus compliquée - question de forme -, " Sum-of-Square-Accouted-For (SSAF) mesure " s'écrit:

$$SSAF = \frac{\left(\sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R \delta_{ijr}\hat{\delta}_{ijr} \right)^2}{\sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R \delta_{ijr}^2 \sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R \hat{\delta}_{ijr}^2}$$

La troisième; " Variance-Accounted-For (VAF) mesure " est:

$$VAF = 1 - \left(\frac{\sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R \left(\delta_{ijr} - \hat{\delta}_{ijr} \right)^2}{\sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R \delta_{ijr}^2 \sum_i^N \sum_{j,j\neq i}^N \sum_{r=1}^R (\delta_{ijr} - \delta_{..r})^2} \right)$$

où

$$\hat{\delta}_{ijr} = \frac{\sum_{s=1}^{P} \alpha_r(x_{is} - x_{js}) + \sum_{s=1}^{P} \beta_r(x_{is} - x_{js})}{\sum_{s=1}^{P} \alpha_r(x_{is} - x_{js}) + \sum_{s=1}^{P} \beta_r(x_{is} - x_{js}) - \sum_{s=1}^{P} \theta_r min(x_{is}, x_{js})}$$

R désigne le nombre de juges. Les différents paramètres:

- α_r représente l'impact des traits propres à i dans la paire de simuli $\{ i, j \}$ pour le juge r.
- β_r désigne l'impact des traits propres à j dans la paire de stimuli $\{ i, j \}$ pour le juge r .
- θ_r représente l'impact des traits communs aux stimuli i et j pour le juge r .

Notons que les paramètres $\alpha_r, \beta_r, \theta_r$ sont tous positifs. Pour complèter cette liste, nous renvoyons le lecteur à de Leeuw et Meulman [6, 1986].

3.5 Estimation de la configuration X

Lorsqu'elles ne se singularisent pas par la fonction coût, les méthodes - exploratoires - de la MDS qui se proposent de faire un ajustement au sens des moindres carrés, peuvent se distinguer les unes des autres par l'algorithme itératif mis en oeuvre pour la construction de la configuration X. Dans ALSCAL, le problème de minimisation de la fonction perte est résolu par - comme dans la majorité des programmes cités à l'introduction - la méthode du gradient. de Leeuw et Heiser (1977) montrent que l'utilisation des méthodes de gradient, préconisée par de très nombreux auteurs dont Kruskal [11, 1964] n'est pas très satisfaisante pour minimiser la fonction STRESS en fonction de X [cf. 3.4] . Ils généralisent, en conséquence, la démarche usuelle au cadre de l'optimisation convexe non différentiable et proposent dans SMACOF [8, 1986], un algorithme appartenant à la classe des méthodes de sous-gradient qui consistent à choisir une configuration initiale et à construire une suite

- $\{X^1, X^2, ..., X^m\}$ avec $m \in IN$ telle que la suite des valeurs $\{ STRESS(X^1), STRESS(X^2), ..., STRESS(X^m) \}$ soit décroissante et bornée.

Les algorithmes de sous-gradient se caractérisent par une vitesse de convergence lente et les auteurs de SMACOF présentent alors un procédé permettant de réduire le nombre d' itérations de la première version . C'est dans ce contexte que Browne [3, 1985] propose dans ELEGANT, une approche algorithmique dans laquelle les pondérations sont de signe quelconque.

En effet, si de Leeuw et Heiser considèrent la pondération de signe positive

$$q_{ij}(X) = q_{ij}\frac{\delta_{ij}}{d_{ij}(X)}$$

Browne considère la pondération de signe quelconque

$$q'_{ij}(X) = q_{ij}(d^2_{ij}(X) - \delta^2_{ij})$$

Notons que cet auteur présente de plus un algorithme de type Newton-Raphson qui semble préférable au premier, compte tenu de son efficacité numérique. Browne [3, 1985] fait une série d' expérimentations assez convaincantes. Machmouchi [16, 1992] propose un algorithme d'exploration probabiliste basée sur les méthodes de Recuit Simulé. L'avantage de cette approche est que l'algorithme ne se termine pas au premier optimum local rencontré. L'inconvénient majeur est l'élaboration d'un critère efficace permettant l'arrêt du processus.

Nos idées algorithmiques que nous exposerons dans le cadre d'un travail futur, permettent de déterminer d'une façon très convaicante la configuration X.

3.6 Estimation des poids et Test d'arrêt

Dans ALSCAL, le calcul des poids q_{ij} intervient au moment du calcul des disparités. En effet, l'idée de pondération est liée aux préoccupations d'échantillonage et de traitement des données manquantes, lorsqu'on se place dans un cadre statistique. Il réside aussi dans la possibilité de donner un poids subjectif à chaque comparaison par couple, par exemple, dans un cadre psychométrique lorsqu'il y a plusieurs juges. Dans nos expérimentations, tous les poids sont pris égaux à 1. Quant au test d'arrêt, il en existe plusieurs dans la littérature de la MDS. Contentons nous des critères les plus utilisés.

Soit $\epsilon > 0$,

1. $max \left| \frac{\partial \sigma(X^*)}{\partial x_{ij}} \right| < \epsilon$
2. $\sum_i \sum_j (\frac{\partial \sigma(X^*)}{\partial x_{ij}})^2 < \epsilon$
3. $\left|\sigma(X^{k+1}) - \sigma(X^k)\right| < \epsilon$

où X^k désigne la configuration à l' itération k.

4 Applications aux données réelles

4.1 Préambule

Afin de mettre en évidence l' intérêt et l'utilité des divers indices de dissimilarité proposés dans la deuxième partie, nous allons maintenant considérer quelques

applications sur des données recueillies par des chercheurs en siences naturelles et humaines. En réalité, ce sont des problèmes concrets qui sont à l'origine de la recherche d'une méthodologie adaptée aux types de données à traiter.

Notre objectif est de construire une méthode dont on poura effectuer l'application sur les données du génome où chaque protéine est caractérisée par une séquence d'acides aminés. Toutes les protéines n'ayant pas la même longueur.

Nous avons retenu pour illustrer nos propos, des exemples qui traitent des trois situations suivantes:

1. Catégories Socio-professionnelles
2. Agriculture régionale Française
3. Cytochrome-C

Les deux premières expériences portent sur des données dont les méthodes factorielles et la classification fournissent de bons résultats. Ce qui constitue une bonne base de comparaison lors des interprétations de nos résultats.

Les sorties de la MDS sont celles du programme ALSCAL. L'identification d'un même élément se fait par un code unitaire (1,...,9,A,...,Z); ce qui ne permet pas une lecture assez explicite. Le tableau que nous avons associé à gauche de chaque graphique permet de porter en clair le sens de chacun des codes à partir des coordonnées des points représentatifs.

4.2 Expérience I : Catégories Socio-professionnelles

Nous reprenons ici une étude d'un petit tableau 8×8 assez connue en Analyse des données, résultant de la répartition des dépenses sur 8 denrées alimentaires de differentes catégories socio-professionnelles. Il s'agit ici de comparer les résultats des méthodes factorielles à ceux des méthodes de la MDS afin de les utiliser si possible comme configuration initiale dans l'algorithme de la MDS. Nous nous intéressons essentiellement aux catégories socio-professionnelles à ce niveau, tout en sachant que nos indices sont valables aussi bien pour les variables que pour les individus.

4.2.1 Les Données

Le tableau de données utilisé est numérique et concerne huit catégories socio-professionnelles - ouvrier, agriculteur, cadre supérieur, etc... - décrites par leur consommation annuelle (1972) de huit denrées alimentaires: viandes, pains, etc... .

	pao	paa	vio	via	pdt	lec	rai	plp
agri	167	1	163	23	41	8	6	6
saag	162	2	141	12	40	12	4	15
prin	119	6	69	56	39	5	13	41
csup	87	11	63	111	27	3	18	39
cmoy	103	5	68	77	32	4	11	30
empl	111	4	72	66	34	6	10	28
ouvr	130	3	76	52	43	7	7	16
inac	138	7	117	74	53	8	12	20

Table 1: La consommation alimentaire des Français, collection de l'INSEE, M 34

4.2.2 Les Résultats

Dans un premier temps, nous avons pratiqué l'A.C.P. normée [cf. figure 2] (resp. l'A.F.C. [cf. figure 3]) sur le tableau objets × variables par la procédure ancomp de ADDAD . Les résultats à consulter en premier sont les valeurs propres, les pourcentages d'inertie et surtout la qualité de représentation de chaque point .

En examinant ces résultats, on remarque aussi bien dans le cas de l'A.C.P que de l'A.F.C., que le support du nuage est bien un espace E à 7 dimensions car la somme des valeurs propres vaut $m = 7$, trace de la matrice des corrélations. Nous considérons dans les deux cas, un sous-espace de représentation à deux dimensions compte tenu de la valeur des deux premiers pourcentages de variance assez élevés (88,5 %).
La table [cf. table 3] (resp. [cf. table 4]) donne les différentes valeurs de la qualité de représentation du premier factoriel de l'A.C.P. [cf. figure 2] (resp. l'A.F.C. [cf. figure 3]). Dans l'ensemble, tous les points sont bien représentés dans les deux plans factoriels. On peut cependant signaler la petitesse de la valeur de la qualité de représentation de PRIN - professions in dépendantes - dans le resultat de l'A.C.P. [cf. table 3].

Nous avons ensuite appliqué la MDS, premièrement, directement sur le tableau objets × variables $X = (x_{ij})$ $i = 1, ..., n$, $j = 1, ..., p$ [cf. figure 5] par le programme informatique ALSCAL. La configuration initiale est alors celle donnée par la méthode de TORGERSON. L'indice utilisé est la métrique euclidienne. Si l'on considère deux individus x_i et x_j, le calcul se fait par:

$$d(x_i, x_j) = \left(\sum_{l=1}^{p} (x_{il} - x_{jl})^2 \right)^{\frac{1}{2}}$$

Une interprétation en termes d'aggrégats donnerait trois groupes:

1. { Agriculteurs, Salariés agricoles }
2. { Inactifs }
3. { Cadres supérieurs, Cadres moyens, Employés, Professions libérales, Ouvriers }

Deuxièmement, la MDS est appliquée sur une matrice de dissimilarités construite à partir des catégories socio-professionnelles avec nos indices. Dans un premier temps, la configuration initiale est restée celle fournie par la méthode de Torgerson. La figure 7 montre précisément les résultats obtenus à partir de l'indice de dissimilarité $D(i,j) = -Log_2 P(i,j)$. Dans le calcul de $P(i,j)$, l'indice $Q(i,j)$ [voir 2.1] est ici de type corrélationnel conformément à une représentation géométrique du tableau de contingence, telle qu'elle est fournie dans l'A.F.C. [cf. Tallur [1]]. Plus précisement, il s'agit de:

$$Q(i,j) = \frac{\sum_{l=1}^{p} \frac{f_{il} f_{jl}}{f_{.l}} - f_{i.} f_{j.}}{\left(\{ \sum_{l=1}^{p} \frac{f_{il}^2}{f_{.l}} - f_{i.}^2 \} \{ \sum_{l=1}^{p} \frac{f_{jl}^2}{f_{.l}} - f_{.j}^2 \} \right)^{\frac{1}{2}}}$$

où

$$f_{ij} = \frac{x_{ij}}{\sum_{l=1}^{n} \sum_{s=1}^{p} x_{ls}}$$

$$f_{i.} = \sum_{j} f_{ij}$$

$$f_{.j} = \sum_{i} f_{ij}$$

désignent respectivement, les fréquences relatives associées à la case (i,j), à la case i de la marge colonne et à la j de la marge ligne. L'interprétation en termes d'aggrégats donne un résultat tout à fait remarquable; le nombre de groupes est maintenu, mais les ouvriers passent aux côtés des agriculteurs.

1. { Agriculteurs, Salariés agricoles , Ouvriers. }
2. { Inactifs }
3. { Cadres supérieurs, Cadres moyens, Employés, Professions libérales.}

Tout en gardant la matrice obtenue à partir de notre indice, nous avons appliqué la MDS en considerant des configurations initiales autres que celle de Torgerson.
La figure 6 (resp. figure 7) donne le résultat obtenu à partir d'une configuration

initiale fournie par l'A.C.P (resp. A.F.C) .On observe une très grande stabilité de la disposition relative des points. Plus précisement, les interprétations dimensionnelles des résultats de la MDS correspondent exactement à celles des représentations obtenues par l'A.C.P et l'A.F.C . Ce qui fait que les résultats de l'A.C.P et de l'A.F.C correspondront pour nous à un point de référence, lorsque l'inertie expliquée par les deux premiers axes est assez forte.

Fort de cette constatation, nous avons orienté nos expérimentations vers un tableau de données plus important du point du vue dimensions.

4.3 Expérience II : Agriculture régionale française

Notre objectif ici est de comparer les résultats de la MDS à ceux de la classification A.V.L. afin de mesurer l'importance des groupes de stimuli observés sur une représentation graphique de la MDS.

4.3.1 Les Données

Les données utilisées se présentent sous forme d'un tableau de 89 individus - départements français de la métropole en 1972 cf. [1] - dont chacun est décrit par 6 variables représentant les six principales modalités d'occupation des sols:

1. les surfaces toujours en herbes (S.T.H), avec les prairies et herbages
2. les cultures fourragères (FOUR)
3. les céréales (CER)
4. les cultures permanentes (CP)
5. les cultures maraîchères et potagères (CMP)
6. les plantes sarclées (PS)

Nous nous intéressons uniquement aux individus, (les départements). Originellement, les départements sont décrits par 17 modalités dont celles concernant l'importance du cheptel et la structure d'exploitation agricole qui n'influent vraiment pas sur les résultats d'après les travaux de B. Tallur [1].

4.3.2 Les Résultats

Les figures 13 et 14 présentent l'arbre obtenu par la classification A.V.L . On observe une partition de l'ensemble des départements en quatre grandes classes, au niveau 86 où la statistique globale atteint son maximum absolu.

Lorsqu'on observe la figure 15, on voit clairement une partition en trois grandes classes conforme aux partitions données par la classification A.V.L. [cf.

figures 13 et 14]. Cette représentation est obtenue par application de la MDS sur la matrice de dissimilarités $D(i,j) = -Log_2 P(i,j)$ construite sur l'ensemble des départements. Cette fois le calcul de $P(i,j)$ est de nature différente de celui proposé dans l'Expérience I. Il revêt une notion de proximité entre points objets, toujours conformément à la représentation géométrique d'un tableau de contingence donnée par l'A.F.C .
Plus précisément, on introduit [cf. [15]] la notion de la contribution brute de la variable l à la comparaison des points objets i et j;

$$s_l(i,j) = \frac{1}{p} - \frac{1}{2}(\psi(i,l) - \psi(j,l))^2$$

où

$$\psi(i,l) = \frac{(\frac{f_{il}}{f_{i.}} - f_{.l})/\sqrt{f_{.l}}}{\sqrt{\sum_{h=1}^{p}(\frac{f_{ih}}{f_{.h}} - f_{.h})^2/f_{.h}}}$$

La normalisation statistique de $s_l(i,j)$, conformément à la distribution $\{ p_{i.} \times p_{j.}$ tq. $(i,j) \in I \times$ I $\}$ donne l'indice $S_l(i,j)$. D'où l'indice brut de similarité [cf. 2.1]

$$S(i,j) = \sum_{l=1}^{p} S_l(i,j)$$

Ce résultat satisfaisant nous permet de considérer désormais les groupes formés par la représentation MDS faite à partir de nos indices. Alors que les graphes de l'a.c.p [cf. figure 16] et de la MDS directe sur le tableau objets × variables [cf. figure 17] ne permettent aucune exploitation de notre tableau en termes d'aggrégats.

C'est ainsi que nous avons alors confronté notre démarche aux données délicates liées à la phylogénie des éléments du règne animal et végétal .

4.4 Expérience III :Cytochrome-C

Cet exemple est étudié par Hartigan [7, 1974] en classification et repris par Beninel [2, 1987] dans le cadre des méthodes factorielles en tant que tableau des distances. Notre idée est de pouvoir identifier par la représentation MDS, les différents regroupements de protéines qui soient conformes au règne animal et végétal.

4.4.1 Les Données

Les données se présentent en matrice de confusion [7, 1974] estimée par comparaison des séquences d'acides aminés de 20 protéines dont celles de l'homme et du singe.

ESPECES	HOMM	SING	CHIE	CHEV	ANE.	PORC	LAPI	KANG	CANA	PIGE	POUL	MANC	TORT	SERP	THON	MOUC	PHAL	MOIS	LEVU	CHAM
HOMM	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
SING	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
CHIE	13	12	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
CHEV	17	16	10	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ANE.	16	15	8	1	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
PORC	13	12	4	5	4	0	.	.	.	.	.	.	.	.	.	.	.	.	.	.
LAPI	12	11	6	11	10	6	0	.	.	.	.	.	.	.	.	.	.	.	.	.
KANG	12	13	7	11	12	7	7	0	.	.	.	.	.	.	.	.	.	.	.	.
CANA	17	16	12	16	15	13	10	14	0	.	.	.	.	.	.	.	.	.	.	.
PIGE	16	15	12	16	15	13	8	14	3	0	.	.	.	.	.	.	.	.	.	.
POUL	18	17	14	16	15	13	11	15	3	4	0	.	.	.	.	.	.	.	.	.
MANC	18	17	14	17	16	14	11	13	3	4	2	0	.	.	.	.	.	.	.	.
TORT	19	18	13	16	15	13	11	14	7	8	8	8	0	.	.	.	.	.	.	.
SERP	20	21	30	32	31	30	25	30	24	24	28	28	30	0	.	.	.	.	.	.
THON	31	32	29	27	26	25	26	27	27	27	26	27	27	38	0	.	.	.	.	.
MOUC	33	32	24	24	25	26	23	26	26	26	26	28	30	40	34	0	.	.	.	.
PHAL	36	35	28	33	32	31	29	31	30	30	31	30	33	41	41	16	0	.	.	.
MOIS	63	62	64	64	64	64	62	66	59	59	61	62	65	61	72	58	59	0	.	.
LEVU	56	57	61	60	59	59	59	58	62	62	62	61	64	61	66	63	60	57	0	.
CHAM	66	65	66	68	67	67	67	68	66	66	66	65	67	69	69	65	61	61	41	0

Table 2: Mutation Distances, From Fitch and Margoliash, Science (1967)

4.4.2 Les Résultats

Nou avons faits plusieurs traitements MDS, donnant des représentations graphiques [cf. figures 8 à 12] satisfaisantes par rapport aux méthodes factorielles - voir Beninel [2, 1987] -. Nous laissons donc le soin au lecteur d'apprécier les résultats qui - disons le - sont loin de résoudre le problème de la reconnaissance des protéines à travers des séquences d'acides aminés. Si la représentation obtenue directement par ALSCAL [cf. figure 10] montre une bonne cohésion entre les protistes (moisissure, levure et champignon), elle a du mal à regrouper les oiseaux. Le pigeon se trouve opposé à la poule.

Nous avons testé les différentes transformations citées au [2.3.3]. On observe tout d'abord qu'à l'inverse de la MDS directe,les protistes ont du mal à se regrouper. La figure 12 montre le résultat obtenu avec la transformation A.V.L. La proximité du chien et de la poule nous semble peu plausible. Par contre, la transformation Z de la figure 11 permet un regroupement remarquable des oiseaux. Quant à la transformation puissance [cf. fig. 8 pour $\alpha = 0.5$ et fig. 9 pour $\alpha = 0.2$], tous les α testés donnent une représentation qui laisse voir un rapprochement entre le lapin et le manchot.

Nous attirons l'attention du lecteur sur le fait que la matrice de confusion [cf. table 2] utilisé dans cette expérience décrit grossièrement les relations entre protéines; elle considère comme dissimilarité entre deux protéines le nombre de sites où on ne retrouve pas le même acide aminé.

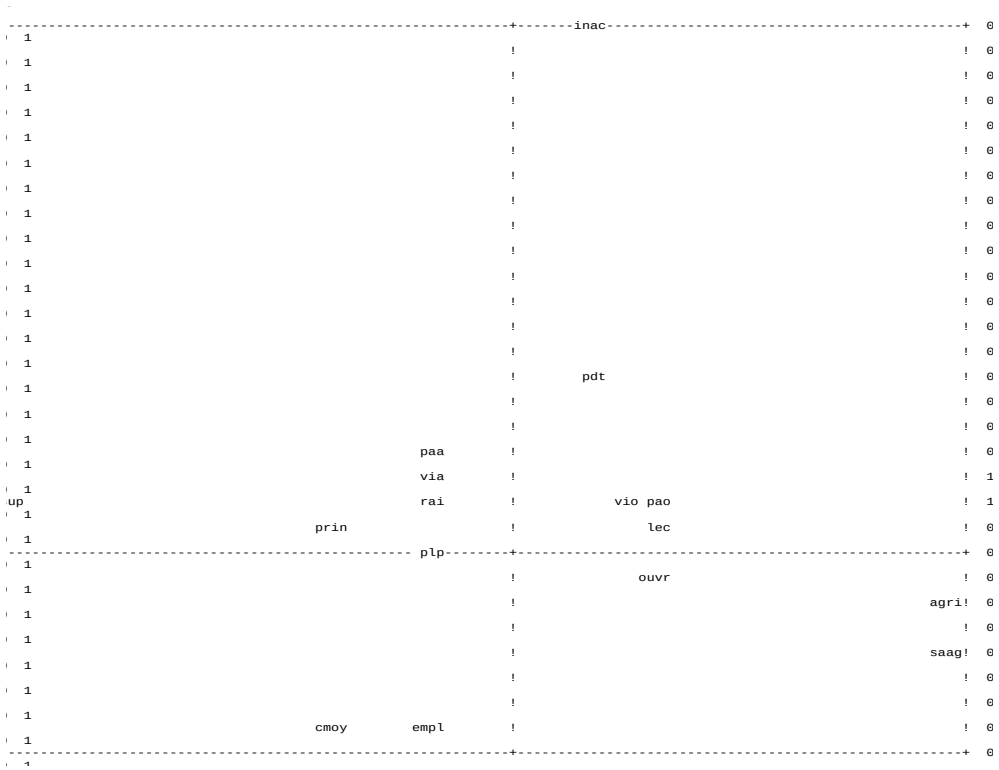

Figure 2: a.c.p. - exp. 1, ancomp (ADDAD)

	I1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR	4#F	COR	CTR
1	agri	974	125	201	3372	884	229	-246	5	9	-840	55	212	-622	30	158
2	saag	982	125	216	3522	898	250	-447	14	28	-352	9	37	916	61	342
3	prin	850	125	59	-1472	575	44	59	1	0	553	81	92	854	194	298
4	csup	996	125	315	-4359	942	383	176	2	4	-1029	53	318	15	0	0
5	cmoy	991	125	61	-1718	753	59	-857	187	104	175	8	9	-412	43	69
6	empl	939	125	24	-807	428	13	-809	430	93	345	78	36	-69	3	2
7	ouvr	937	125	35	899	361	16	-183	15	5	978	426	287	-551	135	124
8	inac	996	125	89	563	56	6	2307	932	756	170	5	9	-132	3	7

Table 3: qualité du graphe de l'a.c.p.

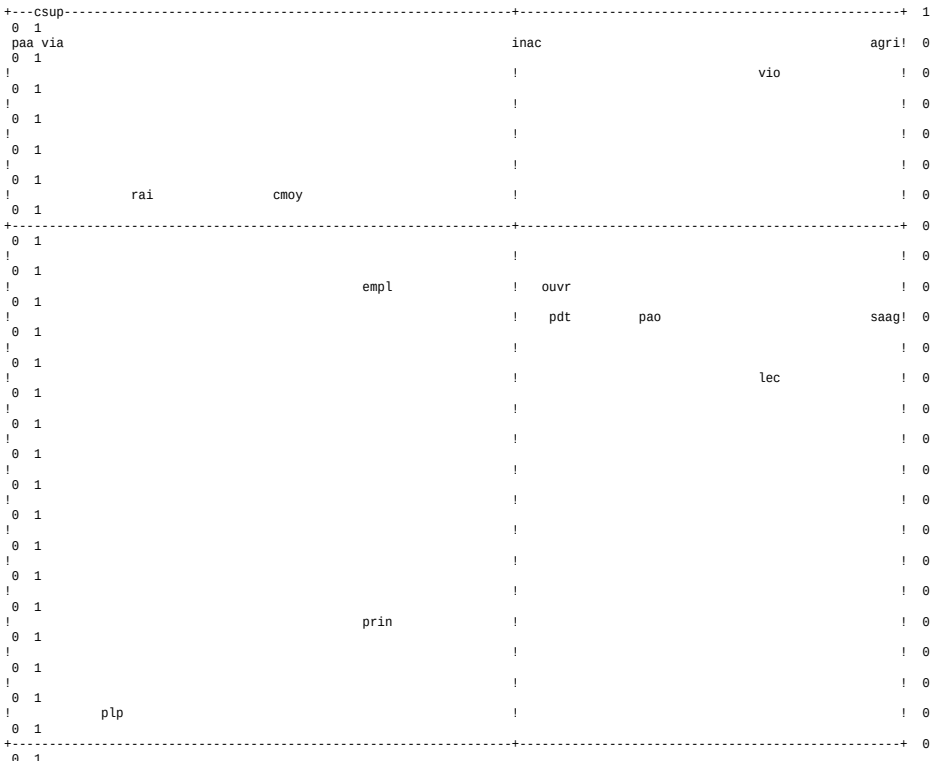

Figure 3: a.f.c. - exp. 1, ancorr (ADDAD)

	I1	QLT	POID	INR	1#F	COR	CTR	2#F	COR	CTR	3#F	COR	CTR
1	agri	993	141	258	435	948	278	81	33	129	-44	10	68
2	saag	989	132	236	433	962	258	-53	14	50	-50	13	79
3	prin	985	119	66	-159	416	31	-179	527	525	-41	28	49
4	csup	998	122	312	-515	950	338	90	29	136	-72	18	153
5	cmoy	993	112	63	-243	960	69	5	0	0	7	1	1
6	empl	976	113	23	-139	850	23	-32	46	16	22	22	13
7	ouvr	977	114	25	48	98	3	-27	30	11	141	841	551
8	inac	984	146	17	5	2	0	81	508	132	49	186	85

Table 4: qualité du graphe de l'a.f.c.

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	AGRI	1	1.5956	-.4787
2	SAAG	2	1.5956	-.8532
3	PRIN	3	-.9574	.7978
4	CSUP	4	-1.8084	-.8532
5	CMOY	5	-.9574	-.9842
6	EMPL	6	-.1064	-.9842
7	OUVR	7	.7446	-.8532
8	INAC	8	-.1064	1.6488

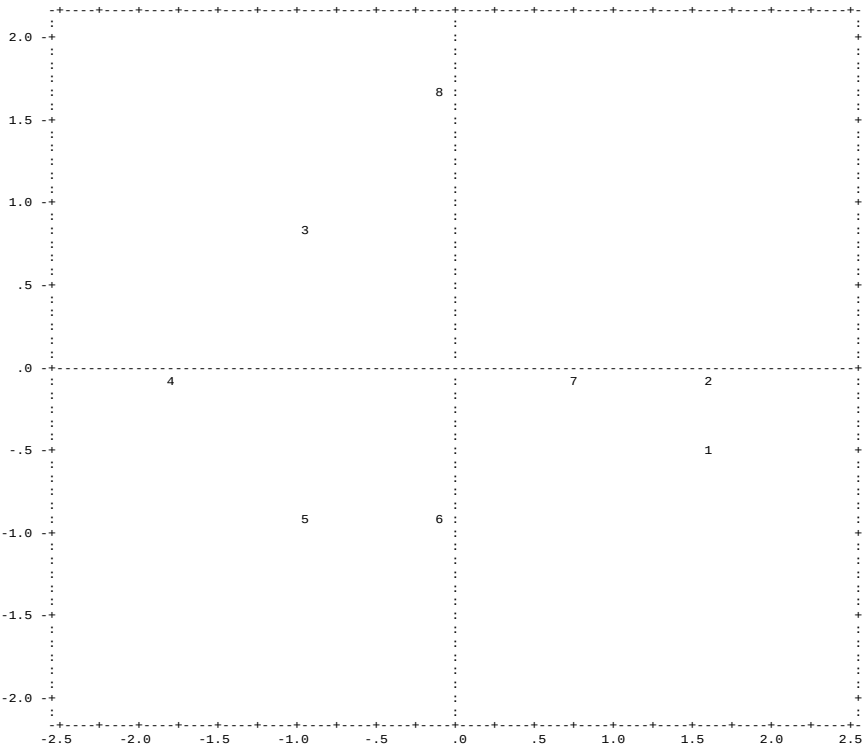

Figure 4: m.d.s. - exp.1, conf. init. = a.c.p, indice = diss. prob.

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	AGRI	1	1.5047	.6520
2	SAAG	2	1.5047	-.1585
3	PRIN	3	-.9028	-1.7555
4	CSUP	4	-1.7053	.6520
5	CMOY	5	-.9028	-.1585
6	EMPL	6	-.1003	-.5517
7	OUVR	7	.7022	-.1585
8	INAC	8	-.1003	1.4546

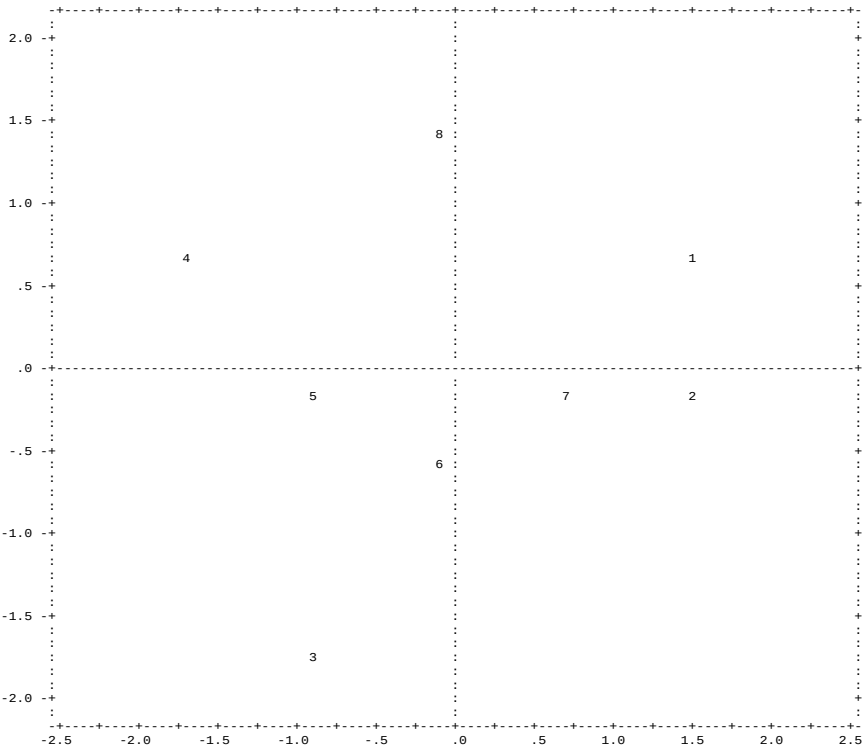

Figure 5: m.d.s. - exp. 1,conf.init. = a.f.p, indice = diss. prob.

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
COLUMN				
1	PAO	1	3.1371	1.2392
2	PAA	2	-.4604	.0522
3	VIO	3	2.5042	-1.2612
4	VIA	4	-1.8781	-.9143
5	PDT	5	-.2447	1.0793
6	LEC	6	-.4249	.0711
7	RAI	7	-.5147	.0431
8	PLP	8	-.8539	.1648

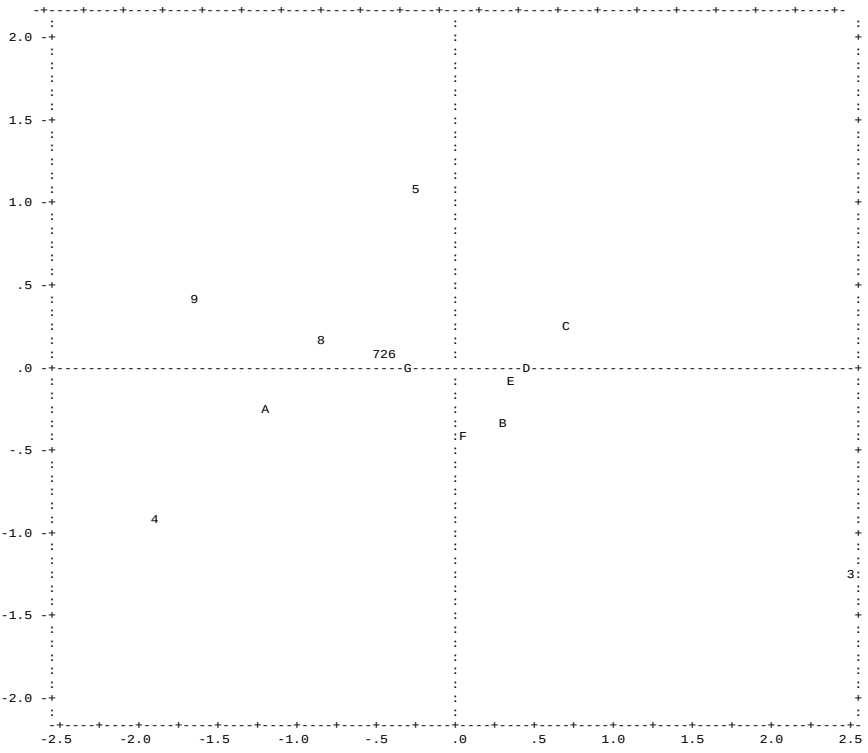

Figure 6: m.d.s. - exp. 1, conf.init. = Torg.,indice = dist.euclid.

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	AGRI	1	1.4583	.2646
2	SAAG	2	1.6697	.7484
3	PRIN	3	-1.3634	.0402
4	CSUP	4	-1.4793	-.4995
5	CMOY	5	-1.2993	-.1260
6	EMPL	6	-.6731	.1465
7	OUVR	7	1.6175	.4944
8	INAC	8	.0696	-1.0685

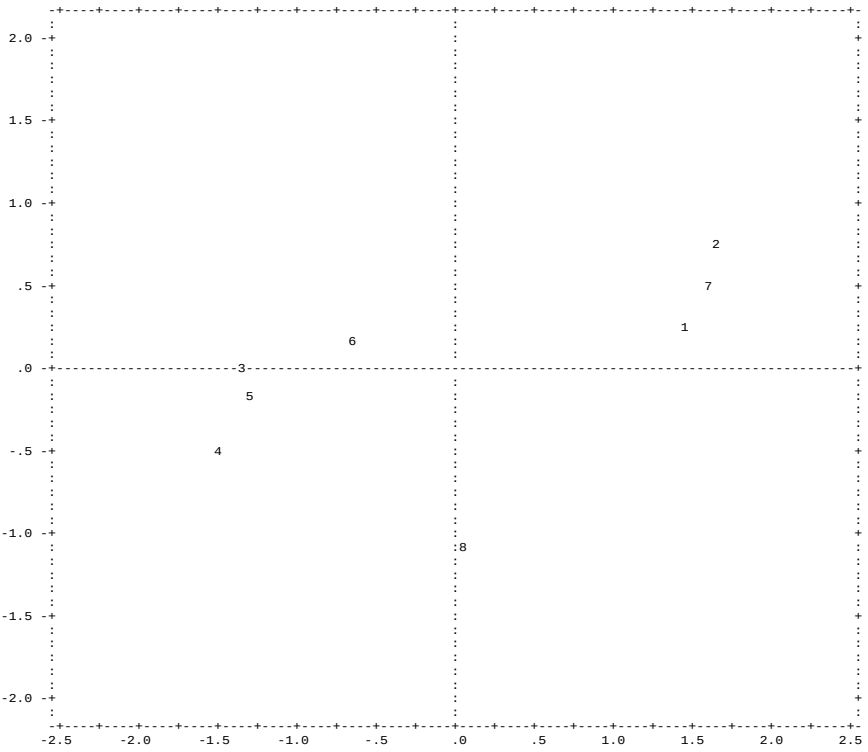

Figure 7: m.d.s. - exp. 1, conf.init. = Torg.,indice = diss. prob.

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	HOMM	1	.2974	.4837
2	SING	2	.2525	.2156
3	CHIE	3	.7243	.2702
4	CHEV	4	.6373	.2848
5	ANE.	5	.5519	.4153
6	PORC	6	.5435	.4982
7	LAPI	7	.4479	-.1866
8	KANG	8	.5029	.6672
9	CANA	9	.3748	-.7316
10	PIGE	A	.3731	-.7389
11	POUL	B	.5698	-.4441
12	MANC	C	.5553	.0821
13	TORT	D	.9518	.2525
14	SERP	E	.5011	-.2043
15	THON	F	.9750	.1413
16	MOUC	G	.3949	-.5943
17	PHAL	H	-.0039	-.1434
18	MOIS	I	-2.4566	-2.0971
19	LEVU	J	-2.6686	.8621
20	CHAM	K	-3.5323	.9693

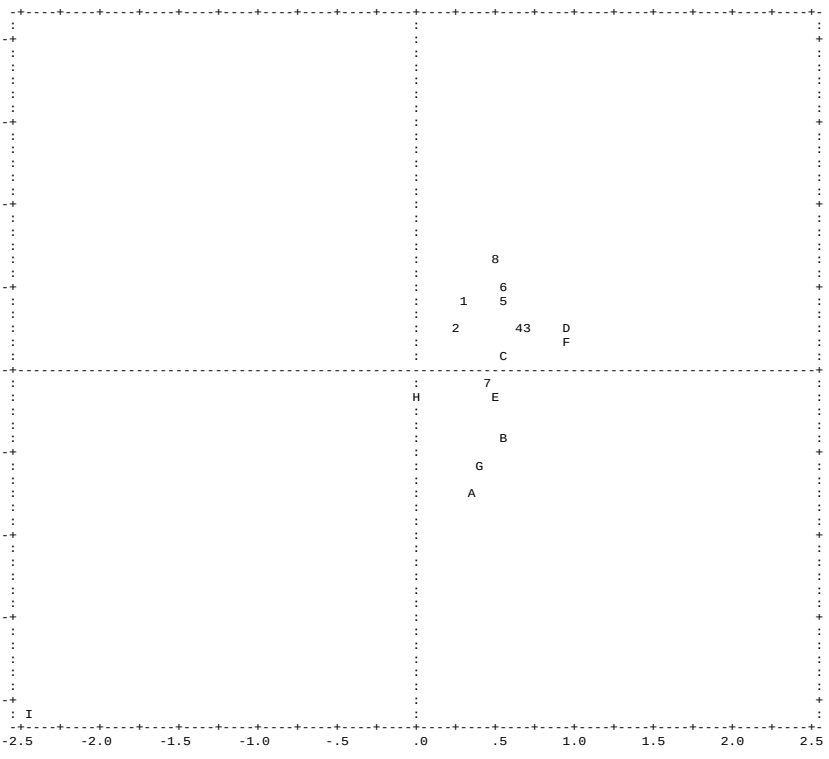

Figure 8: m.d.s. - exp. 3, conf.init. = Torg., transfor. puiss. 0.2

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	HOMM	1	.4195	-.1985
2	SING	2	.3862	-.1182
3	CHIE	3	.5792	-.0757
4	CHEV	4	.6034	-.0465
5	ANE.	5	.5520	-.1151
6	PORC	6	.5524	-.1199
7	LAPI	7	.4855	.0626
8	KANG	8	.6065	-.2594
9	CANA	9	.4262	.3453
10	PIGE	A	.4255	.3486
11	POUL	B	.5049	.1895
12	MANC	C	.4905	.0345
13	TORT	D	.7128	-.0483
14	SERP	E	.5502	.2871
15	THON	F	.9453	-.3701
16	MOUC	G	.3767	.4310
17	PHAL	H	.1623	.2795
18	MOIS	I	-2.6519	2.1155
19	LEVU	J	-2.8621	-1.2919

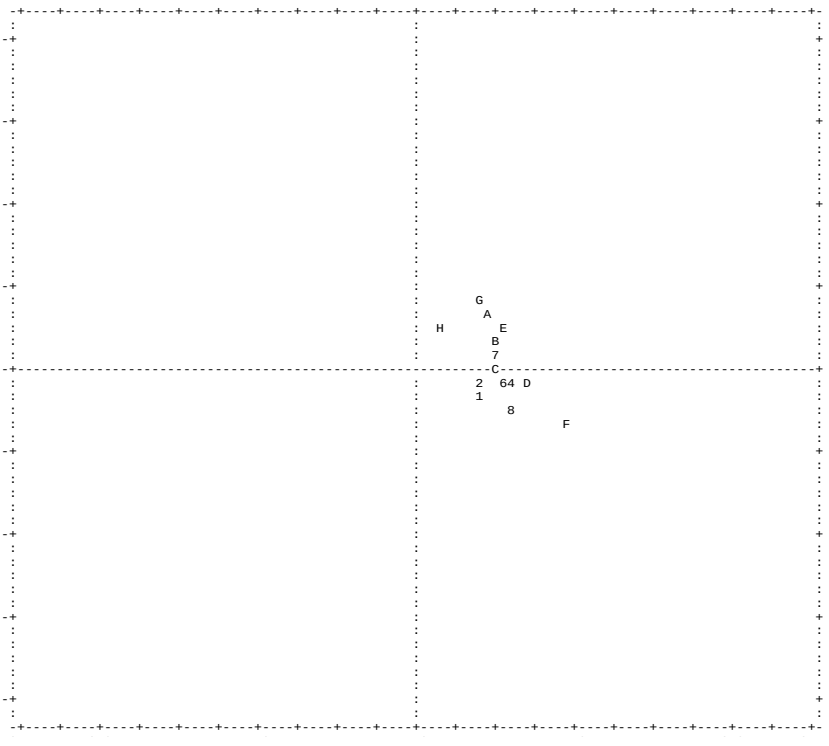

Figure 9: m.d.s. - exp. 3, conf.init.= Torg., transfor. puiss. 0.5

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	HOMM	1	-.3421	.2467
2	SING	2	-.3715	.2962
3	CHIE	3	-1.1165	1.3724
4	CHEV	4	-1.3868	-.3568
5	ANE.	5	-1.3263	-.5621
6	PORC	6	2.5579	.7033
7	LAPI	7	-.4584	-2.1067
8	KANG	8	-.8781	.6276
9	CANA	9	-.5228	.1088
10	PIGE	A	.0880	1.8458
11	POUL	B	-.5175	-1.2108
12	MANC	C	-.4678	-1.2338
13	TORT	D	.1176	1.0975
14	SERP	E	-.2772	.4427
15	THON	F	-.2683	.3882
16	MOUC	G	-.1265	.2078
17	PHAL	H	.2379	.4457
18	MOIS	I	1.7377	-.6300
19	LEVU	J	1.6138	-.8121
20	CHAM	K	1.6989	-.8618

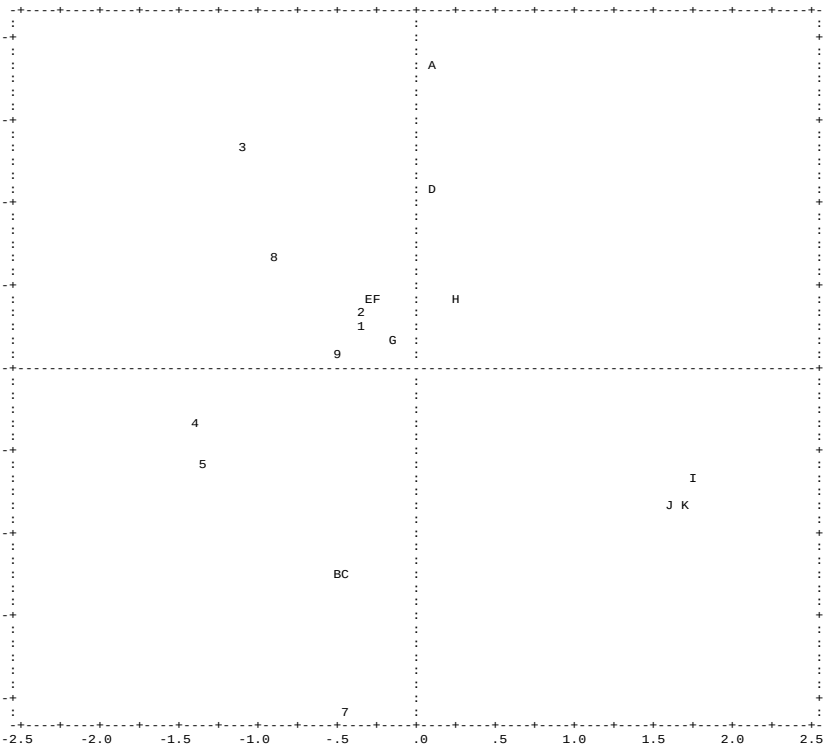

Figure 10: m.d.s. - exp. 3, conf.init. = Torg., directe

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	HOMM	1	.2525	.8816
2	SING	2	.0051	-.5155
3	CHIE	3	.6341	.4140
4	CHEV	4	.8653	-.4681
5	ANE.	5	.4933	.9055
6	PORC	6	.6171	.5257
7	LAPI	7	.4758	-.1224
8	KANG	8	.8234	.7183
9	CANA	9	.2481	-.8471
10	PIGE	A	.2487	-.7603
11	POUL	B	.3976	-.8130
12	MANC	C	.1331	.9101
13	TORT	D	1.0943	.0932
14	SERP	E	.7066	-.2307
15	THON	F	1.3566	.0073
16	MOUC	G	.2061	-.4789
17	PHAL	H	-.3712	.0220
18	MOIS	I	-2.3277	-1.4558
19	LEVU	J	-2.0264	.7640
20	CHAM	K	-3.8324	.4419

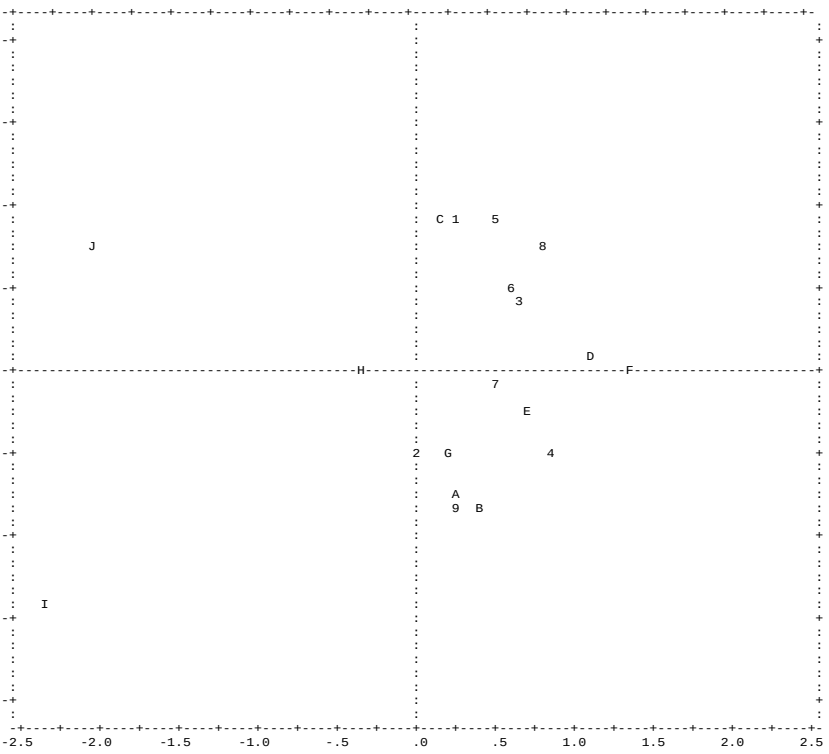

Figure 11: m.d.s. - exp. 3, conf.init. = Torg., transfor. Z

STIMULUS COORDINATES

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	2
1	HOMM	1	-.8127	.4163
2	SING	2	-.7535	.0849
3	CHIE	3	.0253	-.3236
4	CHEV	4	.0500	.3174
5	ANE.	5	-.5113	.2084
6	PORC	6	-.3255	.1982
7	LAPI	7	.1383	.2633
8	KANG	8	-.3572	.5191
9	CANA	9	.1385	-.2862
10	PIGE	A	.1449	-.2953
11	POUL	B	.0934	-.2962
12	MANC	C	-.3819	-.4098
13	TORT	D	-.0302	-.5910
14	SERP	E	-1.6581	.2795
15	THON	F	-.8082	-.7917
16	MOUC	G	-1.0020	-.5537
17	PHAL	H	-2.3671	-.7682
18	MOIS	I	2.6965	1.5360
19	LEVU	J	2.6053	-3.1279
20	CHAM	K	3.1315	3.6284

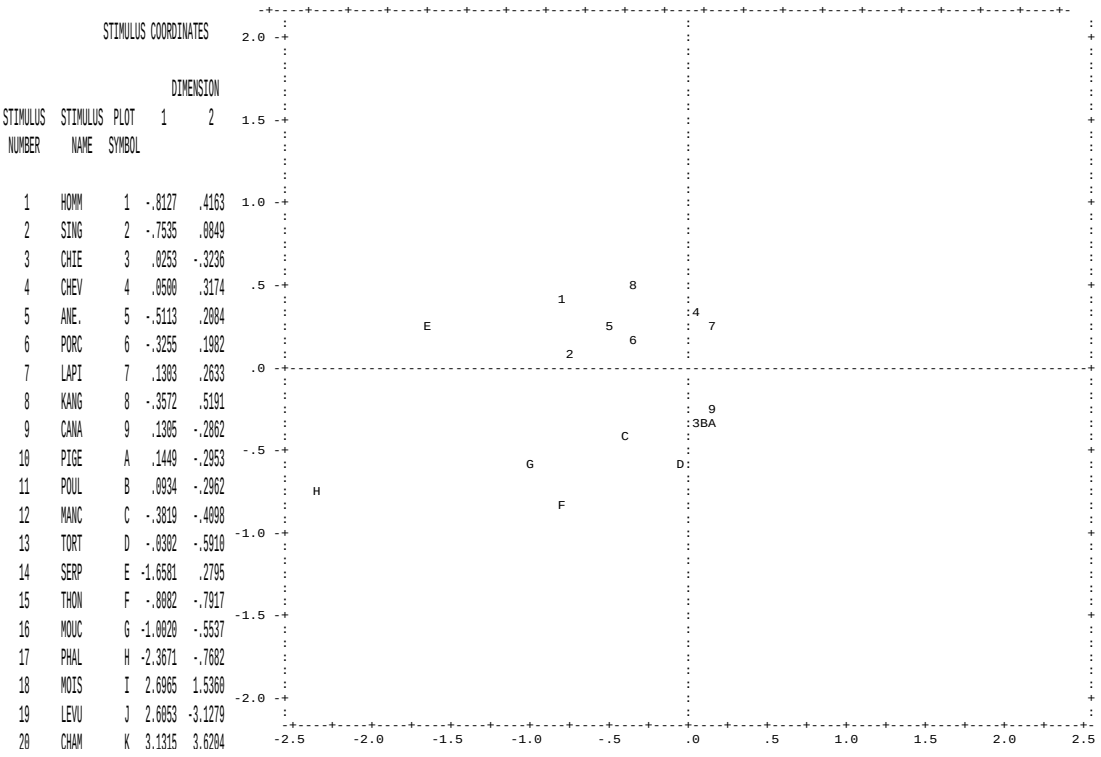

Figure 12: m.d.s. - exp. 3, conf.init. = Torg. transfor. AVL

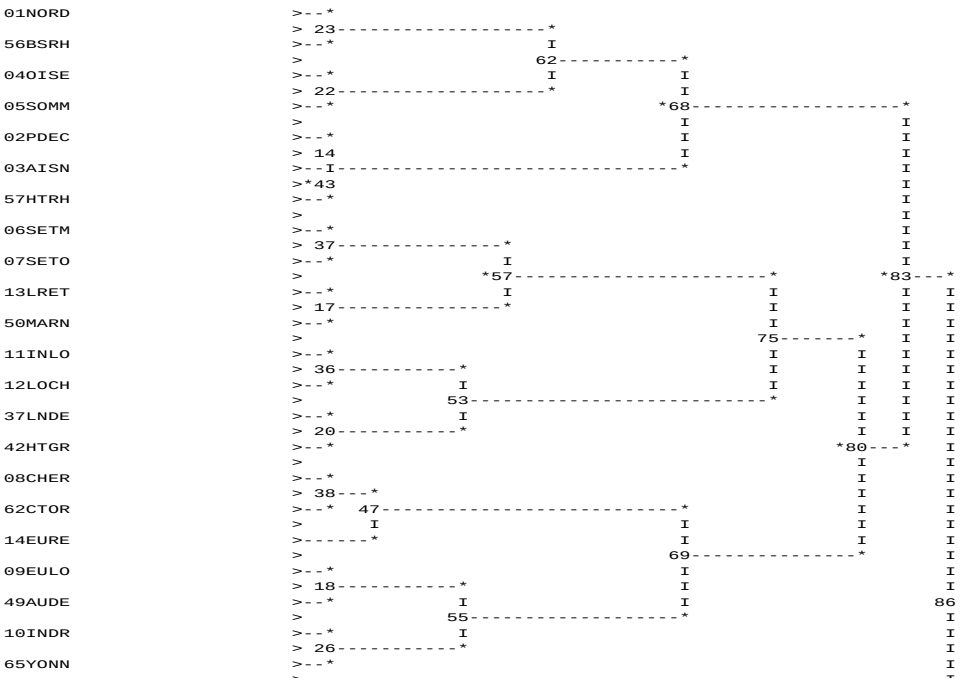

Figure 13: classification - expérience 2, CHAVL

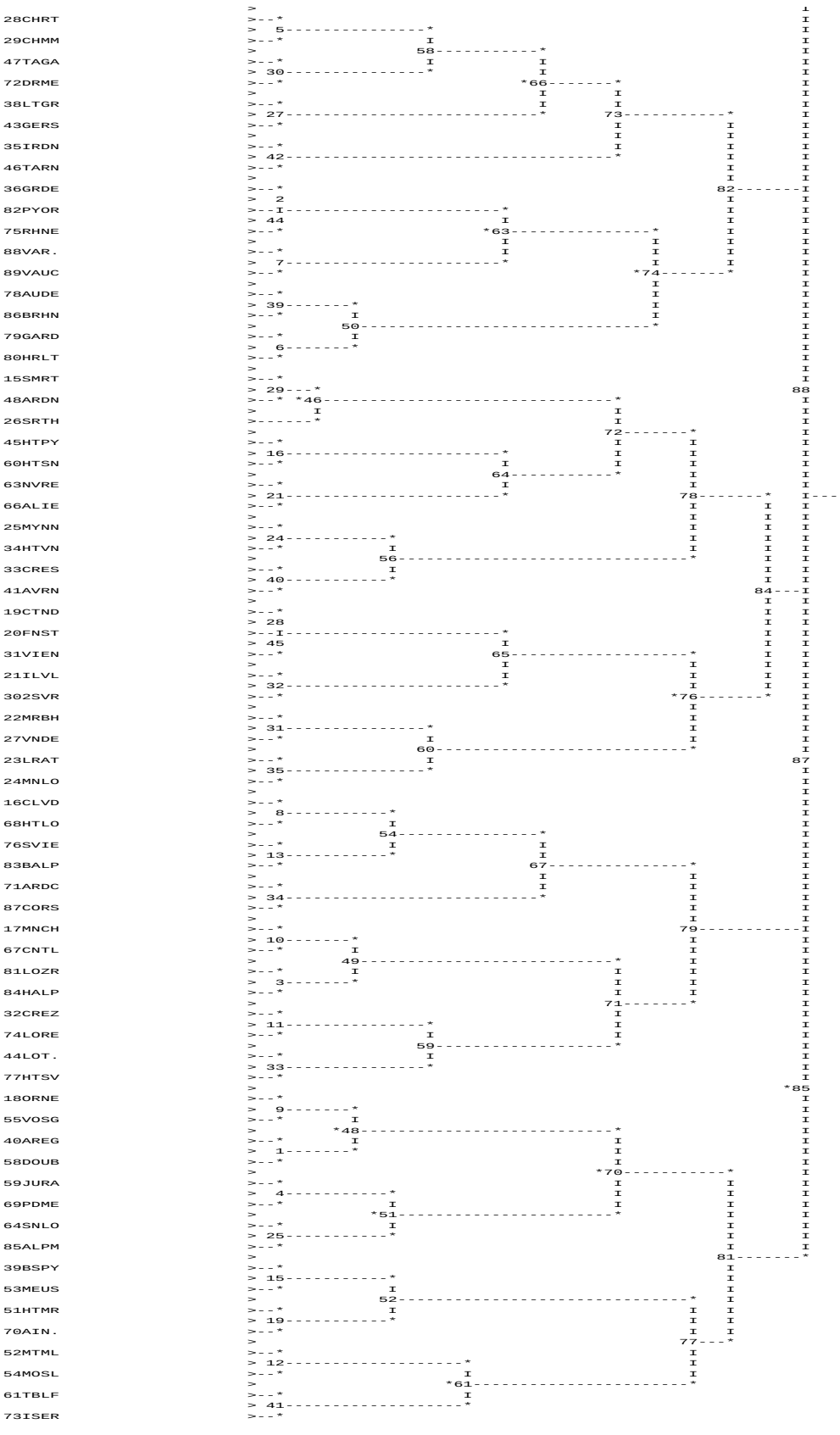

Figure 14: classification - expérience 2,CHAVL

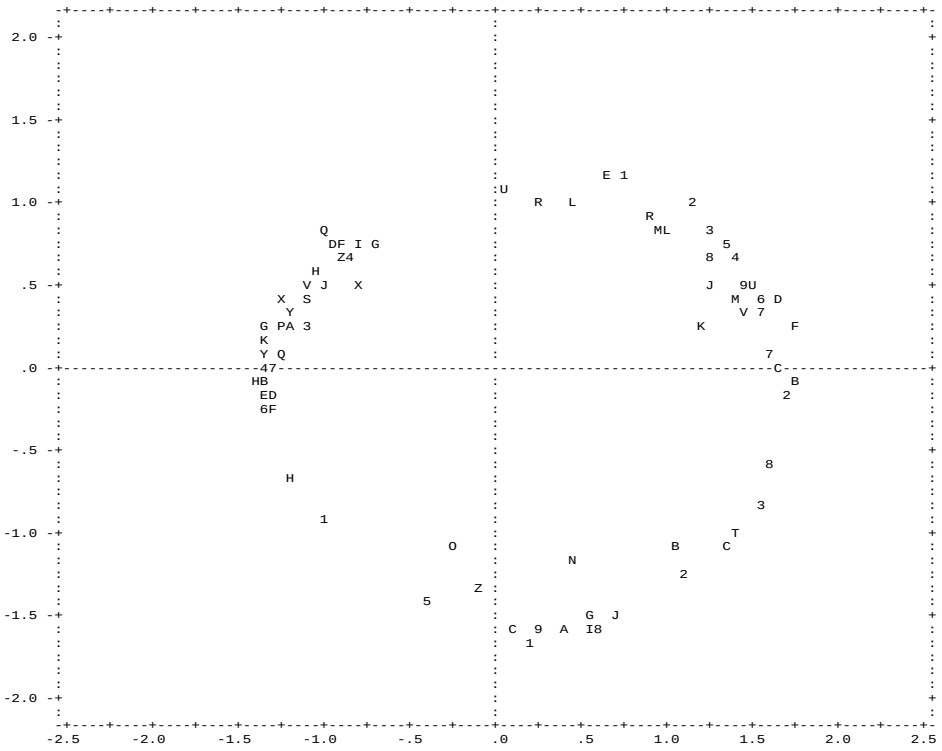

Figure 15: m.d.s. - exp. 2, indice = diss. prob.

STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	DIMENSION 2	STIMULUS NUMBER	STIMULUS NAME	PLOT SYMBOL	DIMENSION 1	DIMENSION 2
1	R01NORD	1	.7493	1.1429	45	R45HTPY	A	-1.2206	.2144
2	R02PDEC	2	1.1630	.9725	46	R46TARN	B	1.0359	-1.0951
3	R03AISN	3	1.2463	.8462	47	R47TAGA	C	1.3581	-1.0478
4	R04OISE	4	1.4008	.7008	48	R48ARDN	D	-.9381	.7755
5	R05SOMM	5	1.3445	.7553	49	R49AUBE	E	1.5208	.4703
6	R06SETM	6	1.5556	.4310	50	R50MARN	F	1.7385	.2452
7	R07SETO	7	1.5461	.3160	51	R51HTMR	G	-.7244	.7566
8	R08CHER	8	1.2376	.7077	52	R52MTML	H	-1.0547	.5806
9	R09EULO	9	1.4328	.4623	53	R53MEUS	I	-.8212	.7384
10	R10INDR	A	1.4924	.5324	54	R54MOSL	J	-1.0218	.5349
11	R11INLO	B	1.7563	-.1003	55	R55VOSG	K	-1.3446	.1554
12	R12LOCH	C	1.6402	-.0175	56	R56BSRH	L	1.0134	.8418
13	R13LRET	D	1.6741	.4325	57	R57HTRH	M	1.3839	.4514
14	R14EURE	E	.6701	1.1354	58	R58DOUB	N	-1.3412	-.0220
15	R15SMRT	F	-.9113	.7911	59	R59JURA	O	-1.3244	.0170
16	R16CLVD	G	-1.3415	.2649	60	R60HTSN	P	-1.2476	.2848
17	R17MNCH	H	-1.3872	-.0658	61	R61TBLF	Q	-1.2614	.0537
18	R18ORNE	I	-1.3599	.1117	62	R62CTOR	R	.9240	.8975
19	R19CTND	J	1.2634	.5009	63	R63NVRE	S	-1.1085	.4304
20	R20FNST	K	1.1820	.2537	64	R64SNLO	T	-1.3104	-.0010
21	R21ILVL	L	.4728	1.0268	65	R65YONN	U	1.4896	.4679
22	R22MRBH	M	.9458	.8745	66	R66ALIE	V	-1.0814	.4858
23	R23LRAT	N	.4375	-1.1365	67	R67CNTL	W	-1.3539	-.0169
24	R24MNLO	O	-.2422	-1.0916	68	R68HTLO	X	-1.2685	.4078
25	R25MYNN	P	-1.1937	.2910	69	R69PDME	Y	-1.3338	.0660
26	R26SRTH	Q	-.9885	.8425	70	R70AIN.	Z	-.9244	.6271
27	R27VNDE	R	.2447	1.0306	71	R71ARDC	1	-1.0192	-.9009
28	R28CHRT	S	1.4056	-.9667	72	R72DRME	2	1.1044	-1.2789
29	R29CHMM	T	1.3875	-.9837	73	R73ISER	3	-1.0925	.2866
30	R302SVR	U	.0694	1.1036	74	R74LORE	4	-1.3428	.0271
31	R31VIEN	V	1.4567	.3716	75	R75RHNE	5	-.3959	-1.4060
32	R32CREZ	W	-1.3733	.0868	76	R76SVIE	6	-1.3313	-.2311
33	R33CRES	X	-.7863	.5164	77	R77HTSV	7	-1.3221	-.0242
34	R34HTVN	Y	-1.2115	.3422	78	R78AUDE	8	.6088	-1.5723
35	R35IRDN	Z	-.0809	-1.3534	79	R79GARD	9	.2662	-1.6161
36	R36GRDE	1	.1763	-1.6474	80	R80HRTL	A	.4118	-1.6128
37	R37LNDE	2	1.6950	-.1706	81	R81LOZR	B	-1.3359	-.0670
38	R38LTGR	3	1.5337	-.8452	82	R82PYOR	C	.1199	-1.6069
39	R39BSPY	4	-.8452	.6606	83	R83BALP	D	-1.3079	-.1751
40	R40AREG	5	-1.3505	-.0041	84	R84HALP	E	-1.3412	-.1728
41	R41AVRN	6	-1.1047	.2160	85	R85ALPM	F	-1.3124	-.2572
42	R42HTGR	7	1.5999	.0655	86	R86BRHN	G	.5483	-1.4926
43	R43GERS	8	1.5905	-.5973	87	R87CORS	H	-1.2180	-.6419
44	R44LOT.	9	-1.2844	-.2803	88	R88VAR	I	.5506	-1.5801
45	R45HTPY	A	-1.2206	.2144	89	R89VAUC	J	.7173	-1.5221

Table 5: STIMULUS COORDINATES

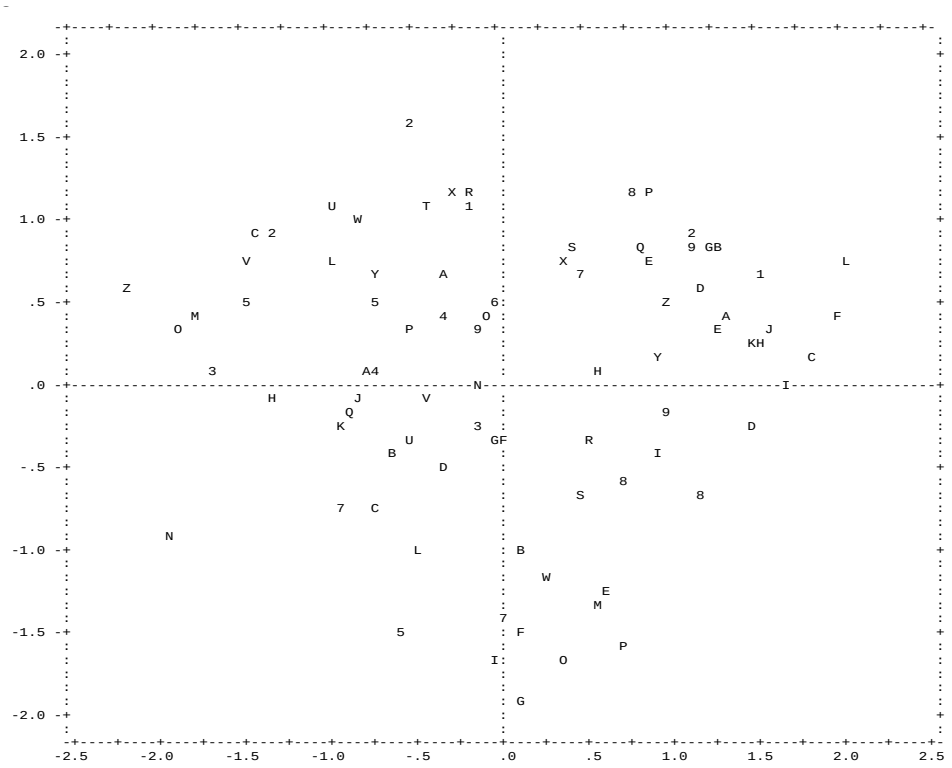

Figure 16: m.d.s. - exp. 2 ,indice = distance euclidienne.

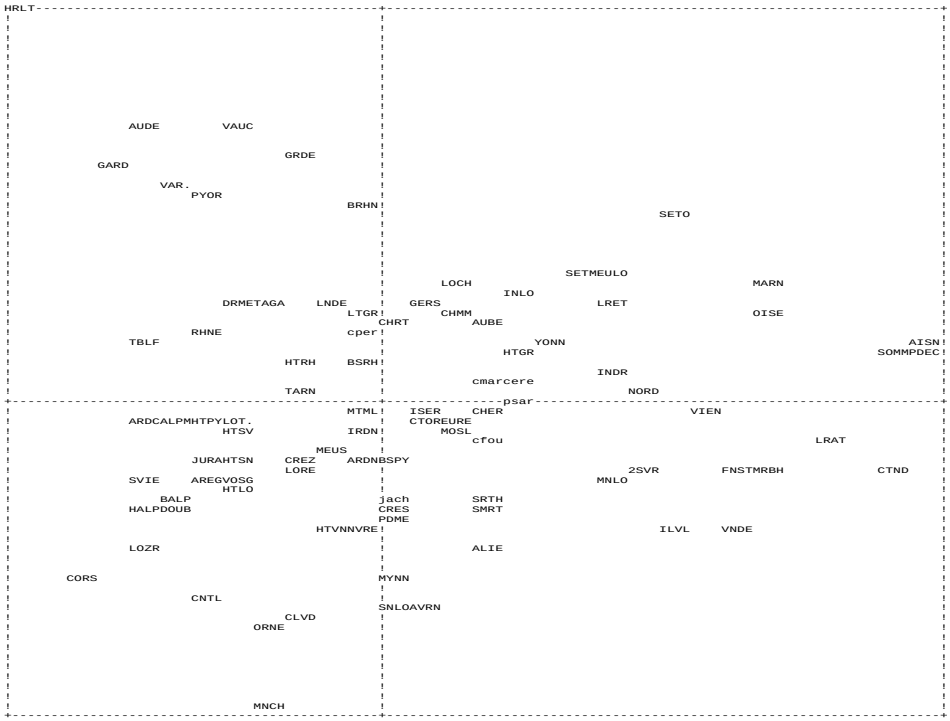

Figure 17: a.c.p. - a.c.p. - exp. 2, ancomp (ADDAD)

5 Conclusion et Perspectives

Les expérimentations numériques - premier objectif de ce travail - nous ont permis d'une part, de mettre en évidence le bon comportement de la famille d'indices probabilistes dans l'analyse de tableaux de données par les méthodes de type exploratoire de la MDS.

Nous ne pouvions étendre notre étude aux méthodes probabilistes de type confirmatoire tels que ceux de Ramsay qui supposent que les dissimilarités entre les objets sont liées aux distances - dans l'espace des stimuli - par un modèle normal ou log normal, compte tenu de leur mise en oeuvre numérique difficile et de leurs fondements formels et statistiques.

D'autre part et surtout, nous nous sommes rendu compte de la difficulté totale d'une mise en oeuvre d'un algorithme de la MDS. En effet, outre le caractère fermé des programmes disponibles,

1. Les algorithmes - disponibles actuellement à notre connaissance - de sous gradient ou de gradient fournissent toujours une solution dont la qualité et la pertinence [cf. figure 17] sont parfois difficiles à évaluer bien que Kruskal [10, 1964] ait proposé un barême d'évaluation de la configuration obtenue.

2. La lecture des graphiques résultats en termes de distances interpoints - et non en terme de classes comme nous esperons le mettre au point avec nos indices - est d'autant plus subjective que l'on dispose de peu d'informations a priori sur le domaine étudié, ou de théorie sur le processus de génèse des données (e.g. les séquences génétiques). Nous estimons dans le cas de l' Expérience II, que notre indice a apporté un plus dans ce sens.

Dans le futur, nous esperons pouvoir étendre notre étude aux tableaux de données plus généraux où les variables de description sont de nature qualitative ou logique. Nous chercherons à affiner nos idées sur les transformations les mieux adapées à nos indices. Nous allons également nous intéresser aux problèmes que posent l'organisation typologique des séquences génétiques. Dans ce contexte, nous essayerons d'améliorer les résultats de la MDS en concevant des configurations initiales telles que celles fournies par les meťhodes des pôles d'attraction [12, 1979] .

Une préoccupation majeure concerne la définition d'un "bon " critère d'adéquation conforme à nos indices. Mieux encore, nous proposerons de nouvelles idées algorithmiques en rélation avec la cellule d'Analyse des données du CNET de Lannion qui auront le mérite d'alleger - supprimer si possible - les dificultés techniques des méthodes MDS.

Remerciements

Je remercie I.C. Lerman pour son aide et ses conseils lors de la rédaction de ce texte.

References

[1] Basavanneppa Tallur. *Contribution á l'analyse exploratoire de tableaux de contingence par la classification.* PhD thesis, Universite de RENNES I.

[2] Farid Beninel. Problèmes de représentations sphériques des tableaux de dissimilarité. October 1987.

[3] M. W. Browne. The young-householder algorithm and the least squares multidimensional scaling of squared distances. *Journal of Classification*, 4:175–190, 1987.

[4] Francis Cailliez. The analytic solution of the additive constant problem. *Psychometrika*, 48(2):4, July 1983.

[5] G. D. d' Aubigny. *l'Analyse Multidimensionnelle Des Donnees De Dissimilarite.* PhD thesis, Universite Joseph Fourier- Grenoble I, January 1989.

[6] Jan de Leeuw and J. Meulman. A special jacknife for multidimentional scaling. *Journal of Classification*, 3:97–112, 1986.

[7] J. A. Hartigan. *Clustering Algorithms.* John Wiley and Sons, April 1974.

[8] W. J. Heiser and J. de Leeuw. Smacof 1. *Departement of Data Theory - University of Leiden*, UG-86-02, 1986.

[9] G. de Soete J. D. Caroll and S. Pruzansky. An evaluation of five algorithms for generating an initial configuration for sindscal. *Journal of Classification*, (6):105, June 1989.

[10] J. B. Kruskal. Nonmetric multidimensional scaling. *Bell Telephone System*, 4821, June 1964.

[11] J. B. Kruskal and M. Wish. *Multidimensional Scaling.* Sage Publications, 1984.

[12] Henri Leredde. *La Methode des Poles d'attraction; la methode des poles d'aggregation: deux nouvelles familles d'algorithmes en Classification automatique et seriation.* PhD thesis, Universite de Paris VI, October 1979.

[13] I. C. Lerman. *Classification et Analyse Ordinale des Données.* Dunod-Paris, 1981.

[14] I. C. Lerman and Ph. Peter. Elaboration d'un indice de similarite entre objets d'un type quelconque. application au probleme de consensus en classification. *Publication interne IRISA*, (232):114, August 1985.

[15] I. C. Lerman, Ph. Peter, and H. Leredde. Principes et calculs de la méthode implantée dans chavl (classification hiérarchique par analyse de la vraissemblance des liens). *à paraître dans la revue Modulad.*

[16] M. Machmouchi. *Contributions à la mise en oeuvre des méthodes d'Analyse des Données de Dissimilarité.* PhD thesis, Université Pierre Mendes-France Grenoble II, October 1992.

[17] A. Mead. Review of the development of multidimensional scaling methods. *The Statistician*, 41:27–39, 1992.

[18] Roger Ngouenet. *Biplot non Lineaire.* Master's thesis, Universite Paul Sabatier - Toulouse III, June 1992.

[19] S. S. Schiffman, M. L. Reynolds, and F. W. Young. *Introduction to Multidimensional Scaling. Theory, Methods, and Applications.* Academic Press, 1981.

[20] Warren S. Torgerson. *Theory and Methods of Scaling.* New York . John Wiley and Sons, inc, 1958.

[21] A. K. Manrai Wayne s. Desarbo, Michael D. Johnson and al. Tscale: a new multidimensional scaling procedure based on tverky's contrast model. *Psychometrika*, 57(1):43–69, March 1992.

Unité de recherche INRIA Lorraine, Technôpole de Nancy-Brabois, Campus scientifique,
615 rue de Jardin Botanique, BP 101, 54600 VILLERS LÈS NANCY
Unité de recherche INRIA Rennes, IRISA, Campus universitaire de Beaulieu, 35042 RENNES Cedex
Unité de recherche INRIA Rhône-Alpes, 46 avenue Félix Viallet, 38031 GRENOBLE Cedex 1
Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105,
78153 LE CHESNAY Cedex
Unité de recherche INRIA Sophia-Antipolis, 2004 route des Lucioles, BP 93, 06902 SOPHIA-ANTIPOLIS
Cedex

Éditeur
INRIA, Domaine de Voluceau, Rocquencourt, BP 105, 78153 LE CHESNAY Cedex
(France)
ISSN 0249-6399