

REVUE DE STATISTIQUE APPLIQUÉE

M. BENNANI DOSSE

Positionnement multidimensionnel d'un tableau à 3 voies

Revue de statistique appliquée, tome 43, n° 4 (1995), p. 63-75

[<http://www.numdam.org/item?id=RSA_1995__43_4_63_0>](http://www.numdam.org/item?id=RSA_1995__43_4_63_0)

© Société française de statistique, 1995, tous droits réservés.

L'accès aux archives de la revue « *Revue de statistique appliquée* » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

POSITIONNEMENT MULTIDIMENSIONNEL D'UN TABLEAU À 3 VOIES

M. Bennani Dosse

*Laboratoire d'Analyse des Données
Université RENNES 2 Haute Bretagne*

6 avenue Gaston Berger 35043 RENNES CEDEX

E-mail : Mohamed.Bennani@uhb.fr

RÉSUMÉ

Nous proposons une méthode nouvelle de positionnement multidimensionnel permettant de représenter graphiquement un tableau à 3 voies où les trois voies jouent des rôles symétriques et où la donnée est exprimée en termes de différence ou de ressemblance. Nous présentons un exemple de données réelles illustrant la méthode proposée.

Mots-clés : *dissimilarité, similarité, dépliage métrique à 3 voies.*

ABSTRACT

We propose a new Multidimensional scaling method which represents 3-way tables where all of the three ways are of equal consideration from the point of exploratory data analysis and whose elements are interpreted as measures of proximity. The treatment of real data is performed to illustrate the proposed method.

Keywords : *dissimilarity, similarity, 3-way metric unfolding.*

1. Introduction

Dans cet article, nous nous intéressons à des données mettant en relation trois ensembles d'objets, que nous noterons I , J et K , où à chaque triplet (i, j, k) de $I \times J \times K$ est associé un nombre mesurant leur «lien», proximité, ressemblance ou différence.

De nombreux tableaux à 3 voies (ou à 3 entrées) X_{IJK} peuvent être vus comme décrivant de cette façon la relation entre I , J et K . C'est notamment le cas des tableaux d'opinions (ou de notes), où chaque individu (ensemble K) exprime, par une note, sa préférence plus ou moins grande pour certains objets (ensemble J) suivant certaines caractéristiques (ensemble I).

C'est aussi le cas des tableaux de contingence à 3 voies où à chaque modalité i de la variable I , à chaque modalité j de la variable J et à chaque modalité k de la variable K on associe le nombre n_{ijk} d'individus possédant simultanément ces trois

modalités. Le nombre n_{ijk} mesure alors la ressemblance entre les trois modalités i, j et k .

Pour analyser de telles données, la démarche classique consiste à présenter le tableau à 3 voies comme un tableau rectangulaire (à 2 voies), et ce de trois manières différentes : $(I \times J) \times K$, $I \times (J \times K)$ ou $(I \times K) \times J$. Or, souvent, cette démarche présente deux inconvénients. D'une part, toutes les entrées ne sont pas traitées sur un même pied d'égalité ce qui entraîne une asymétrie dans l'importance accordée à chacun des trois ensembles, et donc une partie de l'information contenue dans les données risque d'être perdue. D'autre part, on n'a pas la possibilité d'avoir une représentation simultanée satisfaisante (c'est-à-dire gardant toute l'information) de ces trois ensembles, ce qui est parfois regrettable.

Pour effectuer une analyse faisant jouer le même rôle aux trois entrées, trois approches peuvent être envisagées :

- La première revient à faire l'analyse factorielle des correspondances (AFC) du tableau de Burt associé à X_{IJK} qui, rappelons le, s'écrit :

X_I 0	X_{IJ}	X_{IK}
X_{JI}	X_J 0	X_{JK}
X_{KI}	X_{KJ}	X_K 0

Tableau de Burt associé à X_{IJK}

où :

- Les tableaux non diagonaux X_{IJ} , X_{IK} etc. correspondent aux marges binaires.
- Les tableaux diagonaux n'ont des masses non nulles que sur leur diagonale, ces masses correspondant aux marges d'ordre un X_I , X_J , X_K , sur I , J et K respectivement.

Cette première approche fait jouer un rôle symétrique aux trois ensembles mais ne tient compte que des interactions d'ordre un et deux et non de l'interaction d'ordre trois.

– La deuxième approche est basée sur des modèles qui généralisent la propriété de la décomposition en valeurs singulières. Parmi ces modèles, on peut citer **TUCKALS** de Tucker [15], **PARAFAC** de Harshman [8], **CANDECOMP** de Carroll et Chang [4] ou encore le modèle **ROCKET FORM** de Denis et Dhome [7].

Outre leur complexité, dûe entre autres à l'utilisation de calcul tensoriel d'ordre 3, aucun de ces modèles ne permet une représentation simultanée des trois ensembles.

– La troisième approche, qui fait l'objet de ce papier, est basée sur la notion de **distance à 3 voies** introduite par Joly et Le Calvé [12] et étudiée dans [2]. Cette notion

constitue une extension «à 3 voies» de celle de distance. Dans le paragraphe suivant nous rappelons brièvement la notion de distances à 3 voies et dans le paragraphe 3 nous donnons une généralisation du modèle de dépliage. Dans le dernier paragraphe nous donnons un exemple d'application où nous comparons notre méthode avec la première.

2. Distances à 3 voies

2.1. Définitions

On considère un ensemble E fini, non vide, de cardinal n . C'est l'ensemble des éléments que l'on désire représenter, classer, interpréter. On notera ses éléments par $\{1, 2, \dots, i, j, k, \dots, n\}$.

On définit sur E une relation de dissemblance à 3 voies T de la façon suivante : à chaque triplet (i, j, k) de E^3 on associe un nombre positif, noté T_{ijk} , mesurant leur «différence». Plus la valeur de T_{ijk} est grande plus les trois éléments i, j et k sont considérés comme différents, éloignés ou écartés.

Remarque 1 Cette relation de dissemblance est définie sur des triplets d'éléments d'un même ensemble E . Le cas où les éléments appartiennent à trois ensembles distincts sera développé dans la suite.

Définition 1 (Joly-Le Calvé [12])

Une application T de E^3 dans $\mathbb{R}^+$ est une **distance à 3 voies** sur E si et seulement si, pour tout i, j, k, l de E elle vérifie les 5 axiomes :

1. $T_{ijk} = 0 \Leftrightarrow i = j = k$
2. $T_{ijk} = T_{\sigma(i)\sigma(j)\sigma(k)}$ pour toute permutation σ sur $\{i, j, k\}$
3. $T_{iij} = T_{ijj}$
4. $T_{iij} \leq T_{ijk}$
5. $\max(T_{ijk}, T_{ijl}) \leq T_{ilk} + T_{jlk}$

1. est l'axiome de propreté, 2. indique la symétrie : la distance entre 3 éléments ne dépend pas de leur ordre de présentation. 3. annonce un lien entre certaines distances à 3 voies et à 2 voies (on montre que T_{iij} est une distance à 2 voies). Dans le quatrième axiome on impose que la distance entre trois éléments soit supérieure ou égale à celle entre deux de ces trois éléments. Enfin le dernier axiome est une généralisation de l'inégalité triangulaire à un tétraèdre.

Comme dans le cas classique à 2 voies, si on relâche certains axiomes (1, 4 et 5) nous parlons alors de **dissimilarité à 3 voies**.

Définition 2 (Bennani [2])

Une application T de E^3 dans $\mathbb{R}^+$ est une **dissimilarité à 3 voies** sur E si et seulement si, pour tout i, j, k, l de E , on a :

1. $T_{iii} = 0$
2. $T_{ijk} = T_{\sigma(i)\sigma(j)\sigma(k)}$ pour toute permutation σ sur $\{i, j, k\}$
3. $T_{ijj} = T_{ijj}$

2.2. Représentation euclidienne

Un des problèmes qui se pose dans l'étude des distances à 3 voies est celui de la représentation euclidienne. Plus précisément, étant donnée une distance (ou plus généralement une dissimilarité) à 3 voies T sur E , nous cherchons à représenter les éléments de E par n points $M_1, M_2, \dots, M_n$ dans un espace euclidien de faible dimension de façon que la distance entre M_i, M_j et M_k , notée D_{ijk} (distance qui sera définie dans la suite), approche «au mieux» la quantité T_{ijk} .

Il est clair que suivant le sens que l'on donne d'une part à la proximité entre D et T et d'autre part à la distance à 3 voies D on ouvre la voie à différentes techniques de résolution.

Donner un sens à la proximité entre T et D revient à choisir une mesure de proximité entre tableaux de dissimilarité. Nous utilisons la fonction perte (loss function) :

$$\sigma = \frac{1}{6} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \omega_{ijk} (T_{ijk} - D_{ijk})^2$$

où ω_{ijk} désigne la pondération associée à la quantité T_{ijk} . L'introduction de ω_{ijk} permet de traiter le problème dans le cas, par exemple, où certaines dissimilarités sont manquantes ou on n'est pas sûr de leur mesure. On supposera que $\omega_{ijk} = \omega_{\sigma(i)\sigma(j)\sigma(k)}$ pour toute permutation σ de $\{i, j, k\}$.

Pour mesurer la distance entre trois points, plusieurs choix sont possibles [2]. Nous présentons ici le modèle **périmètre euclidien** : la distance entre M_i, M_j et M_k est égale au périmètre du triangle $M_i M_j M_k$. Cette distance s'écrit :

$$D_{ijk} = d_{ij} + d_{ik} + d_{jk}$$

où d_{ij} désigne la distance euclidienne entre M_i et M_j . On démontre que D vérifie tous les axiomes d'une distance à 3 voies.

Nous considérons donc le problème d'approximation (P) :

Etant donnée une dissimilarité à 3 voies T sur E , trouver une configuration $X = (x_{il})$ de n points dans un espace euclidien de dimension p , p fixé, minimisant la somme :

$$\sigma(X) = \frac{1}{6} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \omega_{ijk} (T_{ijk} - d_{ij} - d_{ik} - d_{jk})^2$$

où :

$$d_{ij} = \sqrt{\sum_{l=1}^p (x_{il} - x_{jl})^2}$$

Nous supposons (sans perte de généralité) que la matrice X est centrée.

Pour approcher une solution au problème (P), nous proposons une méthode itérative basée sur l'approche de «Majorization» [6]. Cette méthode est une généralisation de SMACOF [11].

2.3. Algorithme

En développant l'expression de $\sigma(X)$ on obtient :

$$\sigma(X) = K + \frac{1}{2}\alpha(X) - \beta(X) + \gamma(X)$$

où :

$$\begin{aligned} K &= \frac{1}{6} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^n \omega_{ijk} T_{ijk}^2 \\ \alpha(X) &= \sum_i \sum_j (\omega_{ij.} + 2\omega_{iij}) d_{ij}^2(X) \\ \beta(X) &= \sum_i \sum_j u_{ij} d_{ij}(X) \\ \gamma(X) &= \sum_i \sum_j \sum_{k \neq i,j} \omega_{ijk} d_{ij}(X) d_{ik}(X) \end{aligned}$$

avec :

$$\begin{aligned} \omega_{ij.} &= \sum_k \omega_{ijk} \\ u_{ij} &= \sum_k \omega_{ijk} T_{ijk} \end{aligned}$$

Pour décrire l'algorithme, il est commode d'écrire les fonctions $\alpha(X)$, $\beta(X)$ et $\gamma(X)$ sous une forme matricielle. Pour une démonstration des différents résultats cités le lecteur peut consulter [2].

Proposition 1 Soit V la matrice carrée symétrique d'ordre n définie par :

$$\begin{cases} v_{ij} = -2u_{ij} & \text{pour } i \neq j \\ v_{ii} = -\sum_{k \neq i} v_{ik} \end{cases}$$

Alors : $\alpha(X) = \text{Trace}(X^t V X)$

Proposition 2 Soit $B(X)$ la matrice carrée symétrique d'ordre n définie par :

$$\begin{cases} b_{ij}(X) = 0 & \text{si } d_{ij} = 0 \\ b_{ij}(X) = -\frac{2u_{ij}}{d_{ij}} & \text{si } (i \neq j) \text{ et } (d_{ij} \neq 0) \\ b_{ii}(X) = -\sum_{k \neq i} b_{ik}(X) \end{cases}$$

Alors : $\beta(X) = \text{Trace}(X^t B(X) X)$

Proposition 3 Soit $\xi(X)$ la matrice carrée symétrique d'ordre n définie par :

$$\begin{cases} \xi_{ij}(X) = -(c_{ij}(X) + c_{ji}(X)) & \text{si } i \neq j \\ \xi_{ii}(X) = -\sum_{k \neq i} \xi_{ik}(X) \end{cases}$$

avec :

$$c_{ij}(X) = \begin{cases} 0 & d_{ij} = 0 \\ \sum_{k \neq i, j} \frac{w_{ijk} d_{ik}}{d_{ij}} & \text{sinon} \end{cases}$$

Alors : $\gamma(X) = \text{Trace}(X^t \xi(X) X)$

Finalement la fonction perte s'écrit donc :

$$\sigma(X) = K + \frac{1}{2} \text{Trace}(X^t A(X) X) - \text{Trace}(X^t B(X) X)$$

avec :

$$A(X) = V + 2\xi(X)$$

Il est facile de voir (cf [2]) que la matrice A est semi-définie positive de rang $n - 1$ et que le vecteur e_n , vecteur $n \times 1$ dont toutes les composantes valent 1, est vecteur propre de A associé à la valeur propre zéro.

Proposition 4 L'inverse généralisée de Moore-Penrose de $A(X)$ est donnée par :

$$A^+(X) = (A(X) + e_n \cdot e_n^t)^{-1} - e_n \cdot e_n^t$$

Grâce à ces résultats préliminaires, l'algorithme s'écrit :

Partant d'une configuration initiale X^0 , nous construisons la suite $X^1, X^2, \dots, X^k, \dots$ par :

$$X^{k+1} = A^+ (X^k) B (X^k) X^k$$

On démontre que la suite $(\sigma(X^k))$ est convergente.

Remarque 2

1. Pour faire démarrer l'algorithme, on a besoin d'une configuration initiale qui peut être soit fixée par l'utilisateur soit générée d'une façon aléatoire.
2. Comme dans le cas de la quasi-totalité des algorithmes utilisés en «**Multi-dimensional Scaling**», on n'est pas assuré d'obtenir un minimum global. On peut cependant, en tirant plusieurs configurations initiales, retenir la solution qui fournit la fonction perte minimale.

3. Modèle de dépliage métrique à 3 voies

Dans cette partie nous proposons une extension à 3 voies du modèle de dépliage métrique à 2 voies en utilisant les notions rappelées dans les paragraphes précédents.

Soit Y un tableau à 3 voies défini sur $I \times J \times K$. Nous supposons que Y_{ijk} exprime une dissimilarité entre trois éléments i, j et k appartenant à trois ensembles disjoints. Si la donnée est exprimée en terme de ressemblance ou de similarité on se ramènera au cas précédent en considérant une transformation décroissante simple, par exemple du type $C - Y_{ijk}$, C étant une constante adéquate.

La méthode que nous proposons s'inspire de l'approche dite du «unfolding metric problem» (problème de la métrique dépliée) pour des distances à 2 voies introduite initialement par Coombs [5] dans le cadre euclidien unidimensionnel, généralisée au cas multidimensionnel par Hays et Bennet [9]. Cette approche a donné lieu à une littérature abondante (voir par exemple [2], [10], [14]).

Le principe de la méthode est simple : nous considérons le tableau Y comme un sous-tableau extrait d'un cube T défini sur E^3 où $E = I \cup J \cup K$. A ce cube T nous associons une dissimilarité à 3 voies, notée aussi T , dans laquelle les relations entre les éléments d'un ensemble (I, J ou K) ou entre deux ensembles sont inconnues. Notons que dans ce cas la situation est beaucoup plus complexe que dans le cas à 2 voies où on connaît les relations entre deux ensembles I et J mais où les relations entre les éléments de I ou entre les éléments de J sont inconnues.

Notre problème s'exprime donc : trouver des points $N_i, i \in I, P_j, j \in J$ et $Q_k, k \in K$ tels que : $D(N_i, P_j, Q_k)$ (périmètre du triangle $N_i P_j Q_k$) approche au sens des moindres carrés la quantité T_{ijk} .

Pour faire un tel positionnement multidimensionnel de Y nous utilisons l'algorithme d'approximation d'une dissimilarité à 3 voies par le modèle périmètre euclidien où on pose $\omega_{ijk} = 0$ si la quantité T_{ijk} est inconnue :

$$\omega_{ijk} = \begin{cases} 1 & \text{si } (i, j, k) \in I \times J \times K \\ 0 & \text{sinon} \end{cases}$$

Les dissimilarités inconnues sont choisies arbitrairement puisque leur poids est nul.

4. Application

Pour illustrer la méthode proposée, nous considérons l'exemple suivant tiré de [13]. Il s'agit d'une expérience psycho-sensorielle où un jury composé de cinq professionnels a évalué dix variétés de foies gras selon quatre critères :

- La présentation générale ou aspect (notée sur 2)
- La présentation de la coupe (notée sur 4)
- L'odeur (notée sur 4)
- Le goût (noté sur 15)

L'exemple a un caractère pédagogique étant donné le nombre restreint de dégustateurs.

Nous considérons qu'une note reflète une mesure de ressemblance ou **similarité** S sur le triplet constitué par (foie gras, critère, dégustateur). Pour déduire une dissimilarité T d'une similarité S donnée (et réciproquement), de nombreuses transformations sont proposées dans la littérature. Par exemple :

$$T_{ijk} = C - S_{ijk}$$

où C est une constante adéquate.

Afin de prendre en compte les différences de niveau entre notes, nous avons opté pour la transformation :

$$T_{ijk} = 1 - \frac{S_{ijk}}{C_i}$$

où C_i est la note maximale pour le critère i .

Sur la figure 1, nous avons représenté la configuration obtenue en dimension deux avec une fonction perte de 0,1 (sur des données normées). Les critères sont notés { Asp, Cop, Odr, Goû } pour {Présentation générale, Présentation de la coupe, Odeur, Goût}; les foies gras sont désignés par un nombre entre 1 et 10; enfin les dégustateurs sont notés {J1, J2, J3, J4, J5 } pour {Dégu1,Dégu2,...,Dégu5}.

Sur la figure 1 on a une représentation simultanée (au sens du dépliage métrique) où il est possible de lire aussi bien la proximité entre les foies gras, les dégustateurs et les critères que les proximités entre les foies gras ou entre les dégustateurs ou entre les dégustateurs et les foies gras etc.

Sur le graphique ci-dessous on constate que les foies gras 4, 6 et 8 ont été appréciés par l'ensemble des dégustateurs (en particulier J1). Signalons que, dans l'article de Lavialle, Qannari et Vidal [13], ces 3 produits ont été classés en tête par toutes les méthodes de classement exposées dans leur papier. A l'opposé les foies gras 5, 2 et 9 ont été globalement mal appréciés (Ces produits ont été les plus mal classés dans l'article cité précédemment).

Sur la figure 1, on remarque aussi que J3 se distingue des autres dégustateurs (J3 a plutôt bien apprécié les foies gras 5, 3 et 1 ce qui n'est pas le cas pour les autres

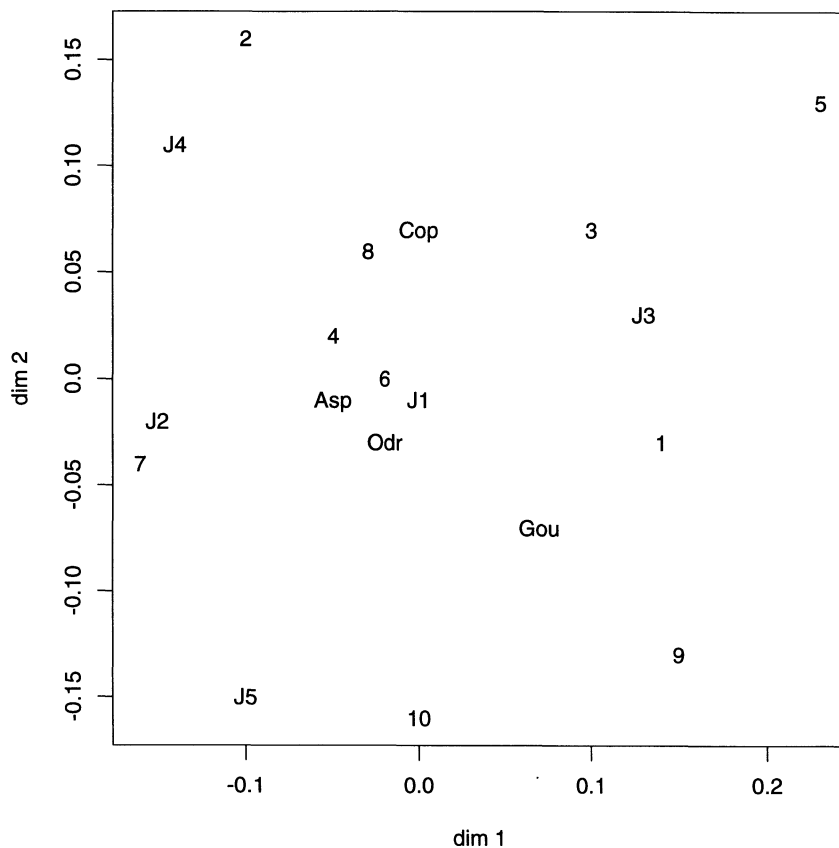


FIGURE 1 :

Plan 1-2 du modèle de dépliage métrique à 3 voies

dégustateurs). Dans un travail publié par ADAMES (cf [1]) et utilisant l'analyse des correspondances avec différents codages, la singularité du dégustateur J3 a également été mise en évidence.

Ce type de remarques est intéressant à relever car il permet de segmenter la population des consommateurs en fonction de leurs préférences pour, éventuellement, proposer différents produits adaptés aux différents segments.

Pour mesurer la qualité de l'approximation, on représente le diagramme de Shepard (figure 2) dans lequel on visualise la relation existant entre les dissimilarités de départ et les distances à 3 voies reconstituées.

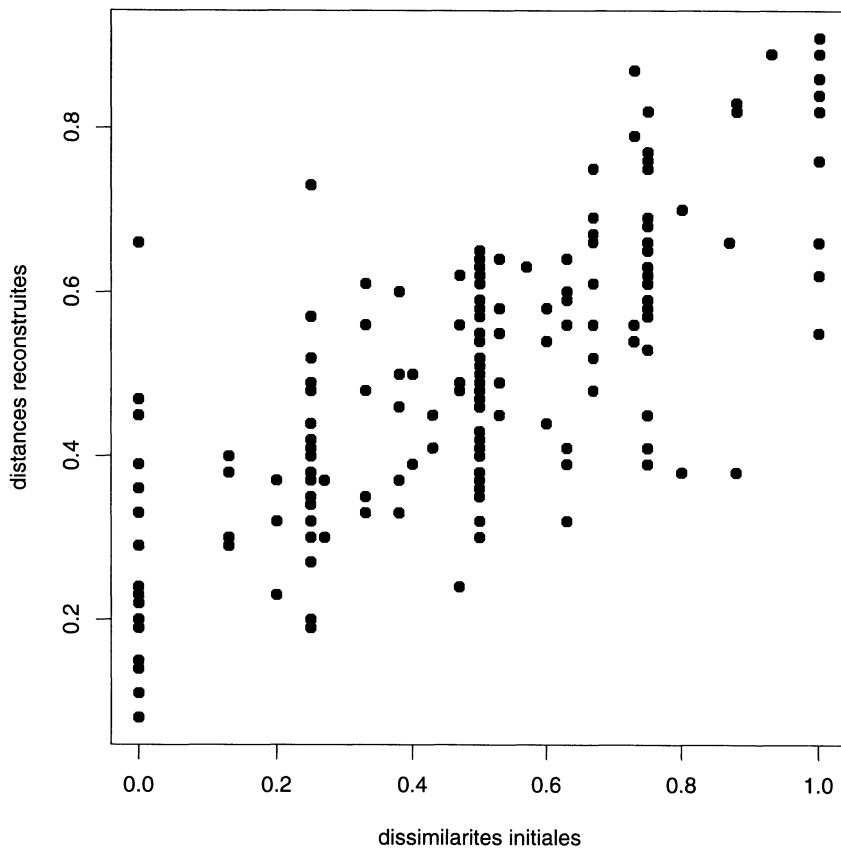


FIGURE 2 :
Diagramme de Shepard

Comparaison avec l'AFC du tableau de Burt

La figure 3 donne le premier plan factoriel (18.16 % de l'inertie dont 9.31 % pour le premier axe) issu de l'analyse des correspondances du tableau de Burt construit à partir des similarités :

$$X_{ijk} = \frac{S_{ijk}}{C_i} = 1 - T_{ijk}$$

On constate que cette figure est très voisine de la figure 1, ce qui sous-entend que sur cet exemple l'interaction d'ordre 3 est relativement faible.

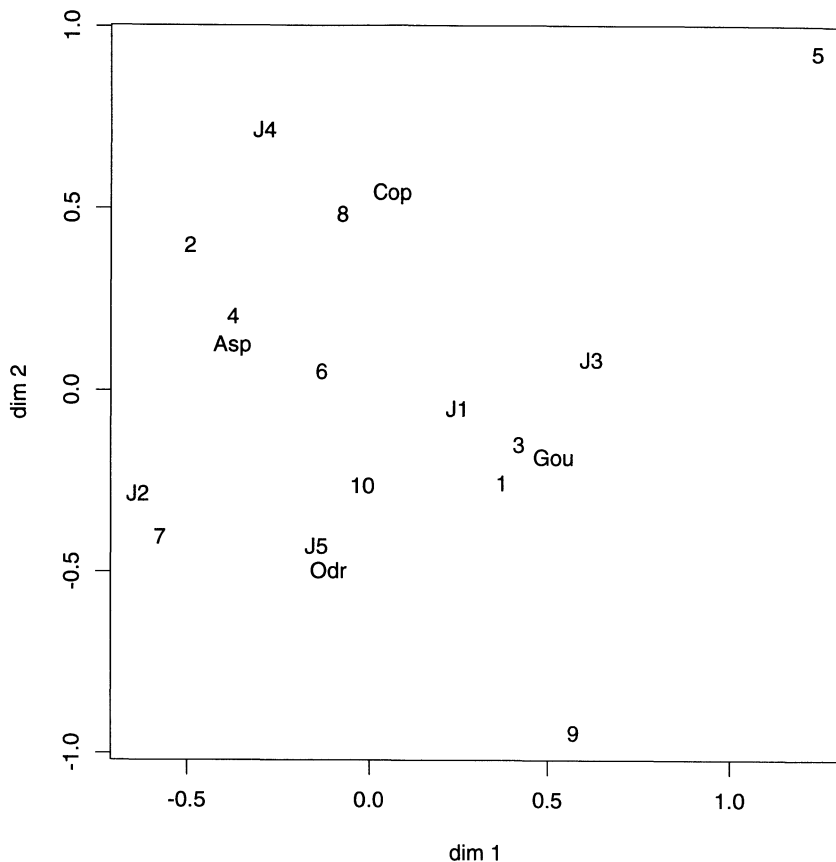


FIGURE 3 :

Plan 1-2 issu de l'AFC du tableau de BURT

5. Conclusion

Tous les tableaux à 3 voies pouvant s'interpréter en termes de dissemblance ou de ressemblance peuvent être représentés par notre approche. On peut penser aux tableaux exprimant directement une dissimilarité entre trois ensembles.

Mais on trouve beaucoup de tableaux exprimés en terme de similarité tels les tableaux de préférence où la donnée S_{ijk} est une appréciation entre un consommateur i , un produit j et un critère k . D'une façon générale rentrent dans cette catégorie tous les tableaux d'opinions (ensembles d'objets, ensembles de critères, ensembles de juges et opinion).

Une catégorie particulière est formée par les tableaux de contingences à 3 voies S où S_{ijk} peut être considérée comme une mesure de similarité entre les trois modalités i, j et k .

Quand le tableau de données peut s'interpréter en ces termes alors l'utilisation de notre approche fournit de nombreux avantages :

- on obtient une représentation (euclidienne) simultanée des trois ensembles
- on peut extraire les représentations séparées de chaque ensemble
- on accorde la même importance aux trois ensembles : on ne privilégie pas une entrée sur une autre
- la comparaison entre les données initiales et celles représentées est facile.

Remerciements

Je remercie les rapporteurs de cet article et P. Cazes pour leurs nombreuses remarques qui ont permis d'améliorer ce travail.

Références

- [1] ADAMES G. – Foie gras et tradition : analyse des résultats d'un concours. *Les cahiers de l'Analyse des Données*, vol. 18, n°4, 1993.
- [2] BENNANI DOSSE G. – *Analyse métriques à 3 voies*. – Thèse de Doctorat - Université de Rennes 2 Haute Bretagne, 1993.
- [3] BROSSIER G. – Etude des matrices de proximité rectangulaire en vue de la classification. *Revue de Statistique Appliquée*, vol. 35, n°4, 1986.
- [4] CARROLL J.D. et CHANG J.J. – Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 1970.
- [5] COOMBS C.H. – *A Theory of Data*. – New York, Wiley, 1960.
- [6] DE LEEUW J. – Convergence of the majorization method for multidimensional scaling. *Journal of classification*, vol. 5, 1988.
- [7] DENIS J.B. et DHORNE T. – Orthogonal tensor decomposition of 3-way tables. *Multiway Data Analysis*, 1989.
- [8] HARSHMANN R.A. – Foundations of the parafac procedure : models and conditions for an explanatory multimode factor analysis. *UCLA working papers on phonetics*, 1989.
- [9] HAYS W.L. et BENNET J.F. – Multidimensionnal unfolding : determining configuration from complete rank order preference data. *Psychometrika*, vol. 26, n°2, 1961.
- [10] HEISER W. – *Unfolding Analysis of Proximity Data*. – PHD thesis – Leiden University, 1981.
- [11] HEISER W. et DE LEEUW J. – How to use SMACOF I. *Rapport technique – Leiden university*, 1977.
- [12] JOLY S. et LE CALVÉ G. – Three-way distances. *manuscrit*, 1989.

- [13] LAVIALLE O., QANNARI E.M. et VIDAL C. – Agrégation d'ordre sous contraintes : classement de produits à partir de données sensorielles. *Revue Statistique Appliquée*, vol. 38, n°4, 1990.
- [14] SCHONEMANN P.H. – On metric multidimensional unfolding. *Psychometrika*, vol. 35, n°3, 1970.
- [15] TUCKER L.R. – Some mathematical notes on three-mode factor analysis. *Psychometrika*, vol. 31, 1966.