Distances à trois voies : concepts théoriques et applications

Mohammed Bennani Dosse

Laboratoire de Statistiques
Université Rennes 2 Haute Bretagne
Place du Recteur Henri Le Moal CS 24307
35043 Rennes Cedex, France
(e-mail : mohamed.bennani@uhb.fr)

Abstract. Distance models for three-way proximity data, which consist of numerical values assigned to triples of objects that indicate their joint (lack of) resemblance, require a generalization of the usual distance concept defined on pairs of objets. An axiomatic framework is given for characterinzing three-way dissimilarity, three-way similarity and three-way distance. The Minkowski-p or $\mathcal{M}_p$ model, which includes the perimeter model, is studied and an Euclidean representation is introduced. Finally, two monotonically convergent algorithms are described that find weighted least squares representations under the Euclidean $\mathcal{M}_1$ and $\mathcal{M}_2$ models.
Keywords: Three-way dissimilarity, Three-way distances, Multidimensional scaling.

1 Introduction

De nombreux domaines scientifiques utilisent le concept de distance dans des contextes très différents. Le présent travail puise ses origines dans l'analyse des données où les distances sont principalement utilisées pour modéliser des jugements subjectifs de différence de façon à découvrir des structures latentes de représentation.

Ce concept intervient aussi lorsqu'on cherche, comme en classification, à transformer un tableau de données X en un tableau de distance D. Cette transformation peut entrainer une perte d'information comme le montrent les deux exemples suivants.

L'exemple 1, tiré de [Daws, 1996], concerne une expérience de libre classement. On demande à chacun des N sujets de produire une partition de n objets reflétant leurs ressemblances perçues. Classiquement, on détermine à partir de l'ensemble des partitions, un tableau de similarité de la manière suivante : pour deux objets i et j, la similarité s_{ij} est définie comme étant le nombre de sujets qui ont classé i et j ensemble. La distance entre i et j est égale à $\delta_{ij} = N - s_{ij}$. Le tableau 1 donne, pour deux groupes différents de sujets, les résultats d'un libre classement (la notation $12 - 3 - 4$ signifie que le sujet a produit trois classes : $\{1, 2\}$, $\{3\}$ et $\{4\}$).

partition	**groupe 1**	**groupe 2**
1234	0	0
$123-4$	5	1
$124-3$	0	0
$134-2$	0	0
$1-234$	1	2
$12-34$	0	1
$13-24$	1	2
$14-23$	0	0
$12-3-4$	1	4
$13-2-4$	0	3
$1-23-4$	1	4
$14-2-3$	0	0
$1-24-3$	2	0
$1-2-34$	2	0
$1-2-3-4$	5	1
Total	18	18

Table 1.

Pour les deux groupes on obtient :

$$s_{12} = 6,\ s_{13} = 6,\ s_{23} = 7,\ s_{14} = 0,\ s_{24} = 4,\ s_{34} = 3$$

On voit donc que la similarité à deux voies s ne permet pas de distinguer les deux groupes de sujets. Si, pour trois objets i, j et k, on définit la similarité à trois voies s_{ijk} comme étant le nombre de sujets qui ont classé i, j et k ensemble, alors on obtient :

- pour le groupe 1 :

$$s_{123} = 5,\ s_{124} = 0,\ s_{134} = 0,\ s_{234} = 1$$

- pour le groupe 2 :

$$s_{123} = 1,\ s_{124} = 0,\ s_{134} = 0,\ s_{234} = 2$$

La similarité à trois voies fait apparaitre clairement que les deux groupes n'ont pas classé de la même manière les quatre objets.

Dans le second exemple, emprunté à [Cox *et al.*, 1991], quatre individus sont décrits par sept variables binaires de la manière décrite dans le tableau 2.

Calculons, à l'aide de l'indice de Jaccard les dissimilarités entre les quatre individus :

$$\delta_{ij} = \frac{q_{ij}}{n_{ij} + q_{ij}}$$

	v_1	v_2	v_3	v_4	v_5	v_6	v_7
1	0	1	0	1	1	0	0
2	0	0	0	0	1	1	1
3	1	0	1	0	1	0	0
4	0	1	1	0	0	0	1

Table 2.

où n_{ij} (resp. q_{ij}) est le nombre de concordances positives (resp. nombre de discordances) entre les individus i et j. On a :

$$\delta_{12} = \delta_{13} = \delta_{14} = \delta_{23} = \delta_{24} = \delta_{34} = \frac{4}{5}$$

L'indice de Jaccard indique que ces quatre individus sont équidistants. Or, il suffit de réordonner le tableau 2 selon la forme suivante pour voir que si les individus 1,2 et 3 jouent des rôles symétriques il n'en est pas de même pour l'individu 4 :

	v_2	v_7	v_3	v_5	v_4	v_6	v_1
1	1	0	0	1	1	0	0
2	0	1	0	1	0	1	0
3	0	0	1	1	0	0	1
4	1	1	1	0	0	0	0

Table 3.

Cette conclusion est confirmée par le calcul de l'indice de Jaccard à trois voies (Bennani Dosse(1993)). En effet, cet indice donne :

$$\delta_{123} = \frac{6}{7},\ \delta_{124} = \delta_{134} = \delta_{234} = 1$$

Les deux exemples ci-dessus montrent l'intérêt de généraliser les concepts de similarité, dissimilarité et distances à trois (voire plusieurs) voies.

Dans la littérature, quelques auteurs se sont intéressés à ce problème. On peut citer les travaux de [Hayashi, 1972, Hayashi, 1989], [Gower, 1984], [Cox *et al.*, 1991], [Pan and Harris, 1991], [Daws, 1996]. Les premiers auteurs abordant les définitions axiomatiques et propriétés mathématiques sont [Joly and Le Calvé, 1989, Joly and Le Calvé, 1995], [Bennani Dosse, 1993] et [Heiser and Bennani Dosse, 1997]. Dans le paragraphe 2, nous allons faire quelques rappels de ces notions puis nous abordons quelques problèmes de représentations géométriques. Le dernier paragraphe traite un exemple réel à l'aide du modèle $\mathcal{M}_2$.

2 Définitions et propriétés

Soit **E** un ensemble fini non vide de cardinal n. On note ses éléments par $1, \ldots, i, j, k, \ldots, n$. Une dissimilarité à trois voies sur **E** est une mesure de dissemblance entre les éléments de **E** pris trois à trois. Plus la valeur de cette dissimilarité est grande, plus les éléments sont considérés comme différents.

Définir une dissimilarité à trois voies δ sur **E** consiste à associer à chaque triplet (i, j, k) de $\mathbf{E}^3$ un nombre réel positif ou nul, noté δ_{ijk}. Formellement :

Définition 1 *Une dissimilarité à 3 voies sur* **E** *est une application* δ *de* $\mathbf{E}^3$ *dans* $\mathbb{R}^+$ *telle que pour tout* $i, j, k \in \mathbf{E}$ *on a :*

$$\delta_{iii} = 0 \tag{1}$$
$$\delta_{ijk} = \delta_{ikj} = \delta_{jik} = \delta_{jki} = \delta_{kij} = \delta_{kji} \tag{2}$$
$$\delta_{iij} = \delta_{ijj} \tag{3}$$

Définition 2 *Soit* δ *une dissimilarité à 3 voies sur* **E**. *On appelle restriction de* δ *aux plans diagonaux l'application définie par :*

$$\rho_{ij} = \delta_{iij} (= \delta_{ijj})$$

Proposition 1 *L'application définie ci-dessus est une dissimilarité à 2 voies sur* **E**.

De façon duale, on définit le concept de similarité à 3 voies comme étant une mesure de ressemblance sur des triplets d'objets. Formellement :

Définition 3 *Une similarité à 3 voies sur* **E** *est une application* s *de* $\mathbf{E}^3$ *dans* $\mathbb{R}^+$ *telle que pour tout* $i, j, k \in \mathbf{E}$ *on a :*

$$s_{iii} = s_{jjj} = s_{kkk} \geqslant s_{ijk} \tag{4}$$
$$s_{ijk} = s_{ikj} = s_{jik} = s_{jki} = s_{kij} = s_{kji} \tag{5}$$
$$s_{iij} = s_{ijj} \tag{6}$$

Comme dans le cas "2 voies" les notions de dissimilarité et de similarités à 3 voies jouent des rôles opposés et on peut passer de l'une à l'autre par une fonction décroissante.

La généralisation de l'inégalité triangulaire qui a été proposé par Joly-Le Calvé(1989) est la suivante : pour tout $i, j, k, \ell \in \mathbf{E}^3$

$$\delta_{ijk} \leqslant \delta_{ik\ell} + \delta_{jk\ell} \tag{7}$$

Bennani Dosse(1993) propose l'inégalité suivante : pour tout $i, j, k, \ell \in \mathbf{E}^3$

$$2\,\delta_{ijk} \leqslant \delta_{ik\ell} + \delta_{jk\ell} + \delta_{ij\ell} \tag{8}$$

Proposition 2 *L'inégalité* (8) *implique l'inégalité* (7).

On peut facilement vérifier que l'inégalité (8) n'est pas suffisante pour que ρ vérifie l'inégalité triangulaire. Par contre on montre (Heiser et Bennani Dosse(1997)) que l'on a :

$$\rho_{ij} \leqslant \frac{5}{4}(\rho_{ik} + \rho_{jk})$$

Pour que ρ vérifie l'inégalité triangulaire, Joly & Le Calvé(1989) introduisent la contrainte suivante :

$$\delta_{iij} \leqslant \delta_{ijk} \tag{9}$$

Définition 4 *Une application qui vérifie les axiomes* (1), (2), (3), (7) *et* (9) *est appelée distance à 3 voies.*

Définition 5 *Une application qui vérifie les axiomes* (1), (2), (3), (8) *est appelée distance triadique.*

Proposition 3 *Les indices de Daws et de Jaccard défnis dans le premier paragraphe sont des dissimilarités à 3 voies qui vérifient les inégalités* (8) *et* (9).

Définition 6 *une application δ de $\mathbf{E}^3$ dans $\mathbb{R}^+$ est dite distance à centre à 3 voies s'il existe un vecteur $u \in \mathbb{R}_+^n$ tel que :*

$$\begin{aligned}\delta_{iii} &= 0\\ \delta_{iij} &= u_i + u_j\\ \delta_{ijk} &= u_i + u_j + u_k\end{aligned}$$

3 Modèles de Minkowski d'ordre p

Étant donnée une dissimilarité à 2 voies d sur $\mathbf{E}$, on peut construire de nombreux modèles de dissimilarités à 3 voies. Les modèles de Minkowski d'ordre p, $p \geqslant 1$, sont définis par :

$$\delta_{ijk}^p = d_{ij}^p + d_{ik}^p + d_{jk}^p \tag{10}$$

Proposition 4 *si d est une distance à 2 voies alors δ est une distance triadique. De plus, la restriction de δ aux plans diagonaux est une distance à 2 voies.*

Remarque 1 *Si $p = 1$ on obtient le modèle périmètre; si $p = 2$ on obtient le modèle $\mathcal{M}_2$ et si $p = \infty$ on obtient le modèle max.*

4 Représentations géométriques

Considérons le problème suivant : étant donnée une dissimilarité à trois voies δ sur $\mathbf{E}$, on cherche à représenter les éléments de $\mathbf{E}$ par des points dans un espace de dimension finie de manière que les distances à 3 voies dans cet espace approchent le plus possible les données initiales. Ce problème est une extension du multidimensional scaling (voir [Borg and Groenen, 1998]).

4.1 Approximation par une distance périmètre

Le problème posé est de minimiser la fonction :

$$\sigma_1 = \sum_i \sum_j \sum_k w_{ijk} \left(\delta_{ijk} - d_{ij} - d_{ik} - d_{jk}\right)^2$$

où d est une distance euclidienne dans un espace de dimension donnée $\mathbf{p}$ et les w_{ijk} sont des poids positifs ou nuls donnés.

4.2 Approximation par une distance $\mathcal{M}_2$

Le problème posé est de minimiser la fonction :

$$\sigma_2 = \sum_i \sum_j \sum_k w_{ijk} \left(\delta_{ijk} - \sqrt{d_{ij}^2 + d_{ik}^2 + d_{jk}^2}\right)^2$$

5 Application

Hayashi(1972) a collecté directement des données de dissimilarité à 3 voies portant sur l'improductivité d'équipes formées de trois individus. Vu la taille restreinte des données ($n = 6$), cet exemple présente surtout un caractère pédagogique. Les données sont présentées dans le tableau 4 :

Hayashi propose, pour faire une représentation euclidienne de la dissimilarité à 3 voies δ, d'utiliser le carré de la surface du triangle. La figure 1 présente le positionnement des 6 individus.

Nous avons analysé ces données à l'aide du modèle $\mathcal{M}_2$. La figure 2 montre que ces données mettent en evidence deux groupes d'individus $\{1, 2, 3\}$ et $\{4, 5, 6\}$.

$\delta_{123} = 1$	$\delta_{124} = 7$	$\delta_{125} = 6$	$\delta_{126} = 9$
$\delta_{134} = 7$	$\delta_{234} = 8$	$\delta_{135} = 6$	$\delta_{235} = 7$
$\delta_{136} = 9$	$\delta_{236} = 9$	$\delta_{145} = 4$	$\delta_{245} = 6$
$\delta_{345} = 3$	$\delta_{146} = 9$	$\delta_{246} = 8$	$\delta_{346} = 5$
$\delta_{156} = 6$	$\delta_{256} = 7$	$\delta_{356} = 3$	$\delta_{456} = 1$

Table 4.

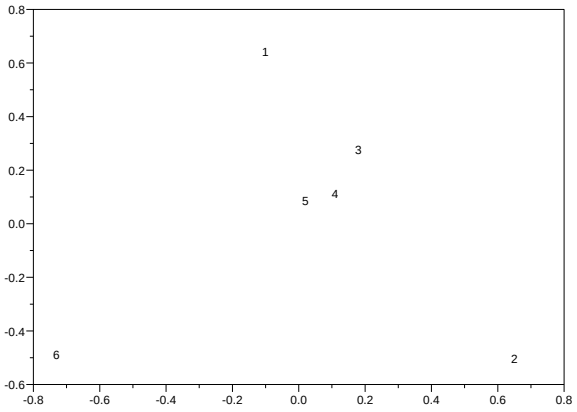

Fig. 1. données de Hayashi : modèle surface du triangle.

6 Conclusion

Ce travail montre qu'il est possible, grâce à quelques outils mathématiques élémentaires, d'étendre à plusieurs voies les notions de dissimilarité, similarité et distance. Nous avons choisi de mettre l'accent sur les représentations Euclidiennes mais d'autres sont possibles (comme les représentations hiérarchiques).

Un champ, particulièrement intéressant dans les applications, est celui où l'on dispose d'un tableau à trois voies où la donnée exprime une dissimilarité entre les éléments de trois ensembles disjoints. Cette approche est une généralisation du dépliage métrique (metric unfolding). Le lecteur intéréssé peut consulter Bennani Dosse(1995)[Bennani Dosse, 1995].

References

[Bennani Dosse, 1993]M. Bennani Dosse. *Analyses métriques à 3 voies.* Thèse de Doctorat-Université de Rennes 2 Haute Bretagne, Rennes, 1993.

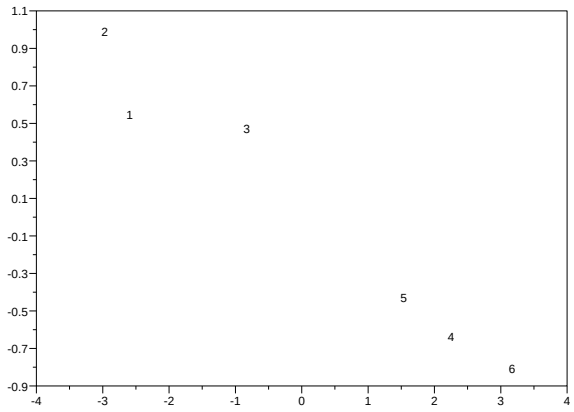

Fig. 2. données de Hayashi : modèle $\mathcal{M}_2$.

[Bennani Dosse, 1995]M. Bennani Dosse. Positionnement multidimensionnel d'un tableau à 3 voies. *Revue de Statistque Appliquée*, pages 63–75, 1995.

[Borg and Groenen, 1998]I. Borg and P. Groenen. *Modern Multidimensional Scaling : Theory and Applications*. Springer Series in Statistics, Rennes, 1998.

[Cox *et al.*, 1991]T.F. Cox, M.A.A. Cox, and J.A. Branco. Multidimensional scaling for n-tuples. *British Journal of Mathematical and Statistical Psychology*, pages 195–206, 1991.

[Daws, 1996]J.T. Daws. The analysis of free-sorting data : Beyond pairwise coocurrences. *Journal of classification*, pages 57–80, 1996.

[Gower, 1984]J.C. Gower. Multidimensional scaling displays. In H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. MacDonald, editors, *Research methods for multimode data analysis*, 1984.

[Hayashi, 1972]C. Hayashi. Two dimensional quantification based on the measure of dissimilarity among three elements. *Annals of the Institute of Statistical Mathematics*, pages 251–257, 1972.

[Hayashi, 1989]C. Hayashi. Multiway data matrices and methods of quantification of qualitative data as strategy of data analysis. In R. Coppi and S. Bolasco, editors, *Multiway data analysis*, 1989.

[Heiser and Bennani Dosse, 1997]W. J. Heiser and M. Bennani Dosse. Triadic distance models : Axiomatization and least squares representation. *Journal of Mathematical Psychology*, pages 189–206, 1997.

[Joly and Le Calvé, 1989]S. Joly and G. Le Calvé. Three-way distances. *Rapport de recherche*, 1989.

[Joly and Le Calvé, 1995]S. Joly and G. Le Calvé. Three-way distances. *Journal of Classification*, pages 191–205, 1995.

[Pan and Harris, 1991]G. Pan and D.P. Harris. A new multidimensional scaling technique based upon association of triple objects p_{ijk} and its application to the analysis of geochemical data. *Mathematical Geology*, pages 861–886, 1991.