MATHÉMATIQUES ET SCIENCES HUMAINES

A. BATBEDAT

L'algorithme proxel pour les dissimilarités

Mathématiques et sciences humaines, tome 102 (1988), p. 31-38

<http://www.numdam.org/item?id=MSH_1988__102__31_0>

Math. Inf. Sci. hum. (26e année, n°102, *1988*, p.31-38).

L'ALGORITHME PROXEL POUR LES DISSIMILARITES

A. BATBEDAT[1]

INTRODUCTION.

En 1976, Booth et Lueker [8] ont proposé un algorithme qui reconnaît si une famille de parties d'un ensemble fini X est représentable comme famille d'intervalles d'un ordre total C sur X. Une telle famille K sera appelée ici une prépyramide et C une orientation de K. L'algorithme permet aussi de déterminer les orientations de K. Il s'appuie sur la notion de "PQ-arbre" : on pourra voir [10] et les exposés 1, 4, de [11], où nous avons situé le passage d'un hypergraphe à son PQ-arbre. Cet éclairage mathématique donne aussi le lien naturel entre l'algorithme précédent et celui de Dubost et Oubina dans [10] (voir en outre, les exposés 1 et 3 de [11]), où les mêmes problèmes sont résolus au moyen de "formules". A ce propos, rappelons que la formule d'une hiérarchie est un produit dans une algèbre généralement non associative ; dans [10] on l'a étendue aux prépyramides en utilisant une algèbre avec deux opérations.

Un peu au-delà du cadre de notre article (qui est celui des prépyramides et des dissimilarités qui les accompagnent), signalons un résultat de Acharya et Las Vergnas [1] donnant un moyen simple, explicité par Leclerc [18], de reconnaître un hypergraphe arboré et dans ce cas de déterminer tous les arbres associés.

Dans le paragraphe 1 ci-après, nous présentons les liens réciproques HTS et DIS entre certains hypergraphes strictement indicés et toutes les dissimilarités. Ainsi, on peut envisager l'utilisation des algorithmes précédents pour reconnaître certains types de dissimilarités. Cependant un traitement direct est toujours utile ; en outre, il est plus facile de mettre en oeuvre un test sur une dissimilarité que sur un hypergraphe. D'où la démarche inverse : mise au point d'algorithmes directs pour les dissimilarités, puis éventuellement étude de leurs applications vers les hypergraphes en passant par HTS et DIS.

L'article actuel concerne les dissimilarités pyras (reliées aux prépyramides) et présente l'algorithme PROXEL, qui signifie "de proximité élevée". Nous avons testé un programme associé, fin 1985.

Comme autres travaux sur les prépyramides et les pyras, signalons Diday [12], [7], Fichet [15], [14], Bertrand [6], [7] et Durand [13], [14]. Nous avons d'autres résultats. Retenons aussi l'article [16] d'Hubert dont nous extrayons deux exemples au paragraphe 5.

Nous nous proposons de détailler l'utilisation de PROXEL pour les hypergraphes, dans un second article en prolongement direct de celui-ci.

---

[1] Mathématiques, Université des Sciences, Montpellier.

1. MISE EN PLACE.
Nous rappelons quelques définitions et résultats de [2], [3], [4] et [9].

1.1. On s'est donné un ensemble X avec n éléments (encore appelés singletons), n > 2. Les graphes, les hypergraphes et les dissimilarités sont tous sur X.

1.2. Un hypergraphe K est ici (dans un sens un peu restrictif) une famille de parties de X (les paliers de K) ne contenant pas la partie vide, contenant les singletons et la partie pleine.

1.3. Pour K, un indice est une application croissante définie sur les paliers intérieurs (non singletons) et à valeurs réelles strictement positives. On dit aussi que K est indicé. D'où le cas particulier strictement indicé, lorsque l'indice est strictement croissant.

1.4. Une chaîne C sur X est la suite des singletons définie par un ordre total sur X.

1.5. Une prépyramide P est un hypergraphe pour lequel il existe une chaîne C telle que tout palier de P soit C-connexe. Alors C est appelée une orientation de P.

1.6. Pour les parties de X nous adoptons la notation en mot. Ainsi la paire en les singletons x et y (on sous-entend qu'ils sont distincts) est écrite xy ou yx.

1.7. Une dissimilarité § est une application réelle strictement positive, définie sur les paires.

1.8. Pour un hypergraphe K indicé par t, l'application DIS donne la dissimilarité & où &(xy) est le plus petit des t(a) pour a dans K et contenant xy.

1.9. Nous avons prolongé à toutes les dissimilarités les trois bijections de Benzécri/Johnson, Diday et Fichet (voir la suite). Ces deux dernières qui sont divergentes, ont donné lieu à deux prolongements distincts. Nous retenons ici la bijection HTS qui construit de façon ascendante classique l'hypergraphe strictement indicé aux seuils pour chaque valeur de la dissimilarité. Sur l'image de HTS, la restriction de DIS est la réciproque de HTS (bijections réciproques).

1.10. Ces bijections mettent en valeur certains types de dissimilarités. Nous sommes concernés ici par les pyras : une dissimilarité § est pyra ssi HTS(§) est une prépyramide strictement indicée (P,s) ; alors les orientations de P sont appelées les orientations de §.
Complément : la bijection de Benzécri/Johnson relie les ultramétriques aux hiérarchies strictement indicées. Celle de Diday n'atteint pas toutes les prépyramides et celle de Fichet n'atteint pas toutes les pyras.

1.11. Donnons-nous une chaîne C : un C-demi-tableau T présente ses valeurs (réelles, strictement positives) T(i,j) pour i de 1 à (n-1), j de (i+1) à n et i < j. Les C-demi-tableaux correspondent biunivoquement aux dissimilarités.

1.12. T de 1.11 est Robinson s'il est croissant en ligne et décroissant en colonne.

1.13. Enfin, après ces définitions, nous avons démontré qu'une dissimilarité est pyra ssi elle possède un demi-tableau Robinson : ensuite ses orientations sont les chaînes de ses demi-tableaux Robinson.

2. EXEMPLES ET CONTRE-EXEMPLES.

2.1. Donnons-nous une chaîne C.
2.1.1. Parmi les prépyramides d'orientation C, celle qui a le plus de paliers est l'hypergraphe de

toutes les parties C-connexes ; ses orientations sont C et C^{op} (l'opposée). Celle qui a le moins de paliers a pour seul palier intérieur la partie pleine ; toute chaîne en est une orientation.
2.1.2. Construisons arbitrairement un C-demi-tableau Robinson R : nous obtenons une pyra d'orientation C. Toute pyra peut être construite ainsi... mais il faut connaître une orientation.

2.2. Pour R de 2.1.2., la plus grande valeur est au sommet (en R(1,n)).
2.2.1. Ainsi, pour n = 3 toute dissimilarité est pyra (car une paire à plus grande valeur sera extrême pour une orientation).
2.2.2. Pour la même raison, la dissimilarité suivante n'est pas pyra :

	y	z	u
x	1	1	3
y		2	1
z			1

3. L'ALGORITHME PROXEL.

3.1. On se donne une dissimilarité quelconque §.
3.1.1. A chaque chaîne C est associé le C-demi-tableau T de §.
3.1.2. On pourra munir C de certaines propriétés : ce sera, bien entendu, relativement à §.
3.1.3. On notera q un indice entier vérifiant toujours $1 \leqslant q < n$. Il donnera la section commençante Cq de C (du rang 1 au rang q inclus). On ajoute par convention : C0 = vide.

3.2. On dit qu'un singleton xk est proche de Cq si xk est hors de Cq (k >q) et pour i de 1 à q, pour j de (q+1) à n : $T(i,k) \leqslant T(i,j)$.
3.2.1. Pour q = 1, il existe toujours un proche. Mais pour q > 1 l'existence d'un proche à Cq n'est pas assurée.

3.3. On dit qu'une chaîne C est de proximité (selon 3.1.2.) si pour tout q (voir 3.1.3.), x(q+1) est proche de Cq.
3.3.1. Pour la dissimilarité de 2.2.2., aucune chaîne n'est de proximité.
3.3.2. Il existe des non-pyras qui possèdent une chaîne de proximité (comparer 1.13 à la caractérisation suivante).
3.3.3. PROPOSITION : La chaîne C est de proximité ssi son demi-tableau T est croissant par ligne.

PREUVE : Lorsque C est de proximité, on considère une ligne $i < (n-1)$ puis $i < k < j \leqslant n$: posant $(q+1) = k$, il vient : $T(i,q+1) \leqslant T(i,j)$ soit $T(i,k) \leqslant T(i,j)$. Réciproque facile.

3.4. Il résulte de cette proposition que toute orientation d'une pyra est une chaîne de proximité, donc nous retiendrons ce critère pour un futur algorithme de reconnaissance pyra. Mais il ne suffit pas. Alors une autre voie prometteuse est celle des valeurs maximales (en raison de 2.2.).
3.4.1. Un singleton xk est dit haut après q si xk est hors de Cq et si, pour au moins un p après q, on a : $§(k,p) = \mathrm{Max}(§(i,j) / i > q , j > q)$.
3.4.2. La chaîne C est haute si chaque xq est haut après (q-1).
3.4.3. Pour une pyra, toute orientation est une chaîne de proximité haute (voir 3.4.).
3.4.4. Voici une non-pyra présentée par une chaîne de proximité haute :

1	1	3
	2	2
		1

3.5. La notion suivante va se révéler décisive.
3.5.1. Un singleton xk est dit élevé après q si xk est hors de Cq et si pour tous indices j et h strictement après q : T(j,h) ≤ Max (T(k,j), T(k,h)).
3.5.2. Après q, tout élevé est haut.
3.5.3. La chaîne C est élevée si chaque xq est élevé après (q-1).

3.6. THEOREME : Soit § une dissimilarité :
i) Si § possède une chaîne de proximité élevée, alors § est pyra.
ii) Lorsque § est pyra, ses orientations sont ses chaînes de proximité élevée.

PREUVE : Soit C une chaîne de proximité élevée pour la dissimilarité § : il résulte de 3.3.3. que le C-demi-tableau T de § est croissant en ligne : alors, pour i < h < j, T(i,h) ≤ T(i,j), ce qui donne Max (T(i,h), T(i,j)) = T(i,j). Ainsi, l'élévement de xi impose T(h,j) ≤ T (i,j). D'où la décroissance en colonne pour T. Au bout du compte, T est Robinson (1.12) donc § est pyra (1.13.). Ceci prouve le i) et montre pour le ii) que toute chaîne de proximité élevée pour § est une orientation. Enfin, si un C-demi-tableau T de § est Robinson, on complète la propriété 3.3.3. en montrant que xq est élevé après (q-1) : pour q < h < j, la décroissance sur la colonne j donne : T(h,j) ≤ T(q,j).

3.7. *L'algorithme PROXEL pour une dissimilarité.*
Entrer une dissimilarité § (a priori quelconque) : on veut savoir si § est pyra et dans ce cas, on veut une orientation.
Etape 1 : préciser l'ensemble E1 des singletons élevés (après zéro). Lorsque E1 est vide, FIN. Sinon, choisir un élevé et le placer au rang 1.
Etape (q+1) de l'itération :
nous avons Cq de x1 jusqu'à xq. Préciser l'ensemble E(q+1) des singletons qui sont élevés après q et proches de Cq.
Si E(q+1) est vide, revenir à l'étape q avec (Eq-xq).
Si E(q+1) n'est pas vide, choisir x(q+1) dans E(q+1).
3.7.1. Cet algorithme PROXEL nous dit si § est pyra et dans ce cas il sort une orientation.
3.7.2. On peut ensuite obtenir toutes les orientations : quand on veut la branche de C(q-1) non suivie par xq, on termine l'étape q par (Eq-xq).
3.7.3. Voici des branches mortes courtes pour PROXEL sur l'orientation d'une pyra :
i) En proximité :
On entre la pyra :

	y	z	t
x	8	8	9
y		8	8
z			1

On peut commencer par x puis z, mais on n'a plus de proche.
(ii) En élévement :
On entre la pyra :

	y	z	t
x	2	2	3
y		1	3
z			1

On peut commencer par y, mais son unique proche z n'est pas élevé après 1.

3.8. *L'algorithme PROXEL pour plusieurs pyras.*
Nous sommes ici dans le contexte très important du consensus : le problème est de trouver si possible une orientation commune à plusieurs pyras. Nous le résolvons pour deux (analogue pour plus de deux).
Entrer les pyras § et & (sur le même X).
Etape 1 : Préciser l'ensemble E1 des singletons qui sont §-élevés et &-élevés.
Lorsque E1 est vide, FIN.
Si E1 n'est pas vide, choisir x1 dans E1.
Etape (q+1) de l'itération :
Nous avons Cq de x1 jusqu'à xq.
Préciser l'ensemble E(q+1) des singletons qui sont §-élevés et &-élevés après q et qui sont §-proches et &-proches de Cq. Finir comme en 3.7.
3.8.1 Comme en 3.7.2., on peut obtenir toutes les orientations communes à plusieurs pyras sur le même ensemble.
3.9. Nous allons ci-après expliciter d'autres critères de reconnaissance pyra.

4. D'AUTRES CARACTERISATIONS.

4.1. On reprend la situation générale de 3.1. à 3.1.3.

4.2. On commence par une forme affaiblie de la proximité.
4.2.1. On dit qu'un singleton xk est proche de xq après q si $q < k$ et pour tout j de (q+1) à n : $T(q,k) \leqslant T(q,j)$. Il existe toujours un proche de xq après q.
4.2.2. On dit qu'une chaîne C est de proche en proche si pour chaque q, x(q+1) est proche de xq après q.
4.2.3. Toute chaîne de proximité est de proche en proche.
4.2.4. Voici une non-pyra présentée par une chaîne de proche en proche élevée :

1	6	5
	2	4
		3

4.2.5. Ainsi pour la reconnaissance pyra d'une dissimilarité, nous avons vu que la proximité élevée permet de répondre. Par contre, ne suffisent pas :
*) Proximité haute
**) De proche en proche élevé.

4.3. On prend une chaîne C a priori quelconque et $q > 1$: on dit qu'un singleton y est en harmonie après Cq s'il est hors de Cq et si : $\S(x1,y) \geqslant \ldots \geqslant \S(xq,y)$.
4.3.1. Une chaîne C est harmonieuse si pour tout $q > 1$, x(q+1) est en harmonie après Cq.
4.3.2. C est harmonieuses ssi T est décroissant en colonne.
D'où le :

CRITERE :
(i) Une dissimilarité est pyra ssi elle possède une chaîne de proximité harmonieuse.
(ii) Les orientations d'une pyra sont ses chaînes de proximité harmonieuses.

4.4. C est à nouveau une chaîne a priori quelconque et $q > 1$: on dit qu'un singleton xk est en étage après q si xk est hors de Cq et pour i de 1 à (q-1) : $T(i,q) \leqslant T(i,k)$.
4.4.1. De façon imagée, on peut dire que ce xk est pour chaque xi, aussi près ou plus loin que xq, alors que dans la proximité le singleton considéré est un des plus près parmi ceux qui sont après q.
4.4.2. On dit que la chaîne C est étagée si pour tout $q > 1$, x(q+1) est en étage après q.

4.4.3. C est étagée ssi T est croissant en ligne.
4.4.4. Par conséquent, C est étagée ssi C est de proximité. Par suite, toute chaîne étagée est de proche en proche.
On en déduit les deux critères suivants :

4.4.5. CRITERE :
i)Une dissimilarité est pyra ssi elle possède une chaîne étagée élevée.
ii) Pour une pyra, les orientations sont les chaînes étagées élevées.

4.4.6. CRITERE :
i) Une dissimilarité est pyra ssi elle possède une chaîne étagée harmonieuse.
ii) Pour une pyra, les orientations sont les chaînes étagées harmonieuses.

4.5. Chacun des critères 4.3.2, 4.4.5 ou 4.4.6, peut conduire à un algorithme de reconnaissance pyra et de détermination des orientations, mais l'harmonie ou l'étagement développent beaucoup de branches mortes.

4.6. On dit qu'une dissimilarité § est à bordures, lorsque la valeur maximale est prise sur une seule paire, disons ici xy. Alors, la bordure de x pour § est la famille des valeurs §(xz). L'autre bordure est celle de y.
4.6.1. On dit que § est à bordure sans ex aequo lorsque § possède une bordure (4.6) sans ex aequo.
Ce cas est plutôt fréquent dans les situations réelles.

4.6.2. CRITERE (avec 2.2.) :
Soit § une dissimilarité à bordure de x sans ex aequo.
*) Ordonner en croissant les valeurs §(xz).
*) Ceci détermine la chaîne C commençant par x.
*) D'où le C-demi-tableau T de §.
*) TEST ROBINSON : § est pyra ssi T est Robinson. Dans ce cas, C est une orientation.

4.6.3. PROPRIETE : Une pyra à bordure sans ex aequo possède exactement deux orientations (opposées).

5. DEUX EXEMPLES DE [16].

5.1. Dans l'article [16] de 1974, Hubert cherche des présentations Robinson associées à un demi-tableau T donné. Nous reprenons ici deux exemples qui sont très intéressants pour nous.
5.1.1. Précisons que pour appliquer correctement nos définitions, il faut munir ici la chaîne des réels de l'ordre opposé à l'ordre classique et oublier la diagonale.

5.2. Premier exemple (on a multiplié les valeurs par 100) :

	2	3	4	5	6
1	564	429	577	742	472
2		389	476	621	394
3			548	411	639
4				503	688
5					461

5.2.1. Dans ce cas, Hubert construit la chaîne de présentation Robinson :
C : 2 < 5 < 1 < 4 < 6 < 3.
Ainsi, dans notre langage et sous 5.1.1., la dissimilarité § de ce T est pyra et C une orientation.

5.2.2. On voit que § est sans ex aequo. Par conséquent, si le problème pyra n'était pas résolu on pourrait appliquer 4.6.2. Maintenant, 4.6.3. nous dit que les seules orientations de § sont C et C^{op}.

5.3. Deuxième exemple :

	2	3	4	5	6	7	8	9	0
1	125	31	13	24	23	41	74	114	105
2		118	46	15	28	49	50	55	37
3			75	53	67	34	33	17	19
4				111	68	47	26	15	14
5					59	36	19	10	10
6						130	44	16	36
7							130	62	30
8								117	78
9									159

5.3.1. Passant en revue diverses méthodes... Hubert n'obtient pas d'orientation ("the resulting matrix transformations are inadequate with respect to the Robinson form").
5.3.2. La dissimilarité § de ce 5.3. n'est pas à bordures car la valeur maximale 10 (voir 5.1.1.) est répétée.
5.3.3. C'est pourquoi nous appliquons PROXEL.

La première étape doit préciser l'ensemble E1 des singletons élevés.

D'après 3.5.2., les seuls cas à examiner sont 5, 9 et 0 (les singletons hauts).

On commence par 5 en le couplant avec 4 qui est à côté dans T : aussitôt 1 est retenu car T(1,4) = 13 > T(1,5) = 24 et 13 > T(4,5) = 111 (toujours 5.1.1.). Ainsi 1 n'est pas élevé.

On couple 9 et 0 ensemble car T(9,0) = 159 est loin de la valeur maximale ; de plus, ils sont à côté. Alors 1 montre que 9 n'est pas élevé et 3 que 0 n'est pas élevé. FIN de PROXEL : § n'est pas pyra.

BIBLIOGRAPHIE

[1] ACHARYA B.D., LAS VERGNAS M., "Hypergraphs with cyclomatic number zero, triangulated graphs and an inequality", *J. Combinatorial Theory B*, 33 (1982), 52-56.

[2] BATDEBAT A., "Deux prolongements de la bijection de Benzécri/Johnson pour toutes les dissimilarités", Séminaire "Mathématiques discrètes et Sciences sociales", au Centre d'Analyse et de Mathématique Sociales, Paris, 16 mars 1987 (un résumé est disponible).

[3] BATBEDAT A., "Les isomorphismes HTS et HTE", Soumis.

[4] BATDEBAT A., "Applications des isomorphismes HTS et HTE", Soumis.

[5] BENZECRI J.P., *L'analyse des Données.1. La Taxinomie*, Paris, Dunod, 1973.

[6] BERTRAND P., *Etude de la représentation pyramidale,* thèse de 3ème cycle, Université de Paris-Dauphine et INRIA-Rocquencourt, 1986.

[7] BERTRAND P., DIDAY E., "A visual representation of the compatibility between an order and a dissimilarity index : the pyramids", *Computational Statistics Quartely 2*, n°1 (1985), 31-42.

[8] BOOTH K., LUEKER G., "Testing for the consecutive ones property, interval graphs and graph planarity using PQ-tree algorithms", *J. Computer Syst. Sci.* 13, n°3 (1976), 335-379.

[9] "Cahier N-Q", *Cahiers de DEA A.Batbedat*, Université des Sciences, Montpellier, 1985.

[10] "Cahier S", *Cahiers de DEA A. Batbedat*, Université des Sciences, Montepellier, 1985.

[11] *Comptes rendus du séminaire argentin en graphes 1985,* Université des Sciences, La Plata, Argentine.
[12] DIDAY E., "Une représentation visuelle des classes empiétantes : les pyramides", *Rapport de recherches INRIA*, n°291 (1984).
[13] DURAND C., *Sur la représentation pyramidale en Analyse des Données*, Mémoire de DEA, Université de Provence (1986).
[14] DURAND C., FICHET B., "One-to-one correspondances in pyramidal representation : a unified approach", in H.H. Bock ed., *Classification and related methods of data analysis*, Amsterdam, North-Holland (1987), 85-90.
[15] FICHET B., "Une extension de la notion de hiérarchie et son équivalence avec certaines matrices de Robinson", *Journées de Statistique*, La Grande-Motte (1984).
[16] HUBERT, L., "Some applications of graph theory and related non metric techniques to problems of approximate seriation : the case of symmetric proximity mesures", *The British J. of mathematical and statistical psychology,* 27 (1974), 133-153.
[17] JOHNSON S.C., "Hierarchical clustering schemes", *Psychometrica*, 39 (1974), 283-309.
[18] LECLERC B., "Arbres minimaux communs et compatibilité de données de types variés", *Math. Sci. hum.*, 98 (1987), 41-67.