MATHÉMATIQUES ET SCIENCES HUMAINES

GILDAS BROSSIER

Approximation des dissimilarités par des arbres additifs

Mathématiques et sciences humaines, tome 91 (1985), p. 5-21

<http://www.numdam.org/item?id=MSH_1985__91__5_0>

Math. Sci. hum. (23^{e} année, n°91, *1985*, p.5-21)

APPROXIMATION DES DISSIMILARITES PAR DES ARBRES ADDITIFS

GILDAS BROSSIER*

I - INTRODUCTION

Les arbres additifs, appelés aussi distances quadrangulaires ou distances arborées ont été étudiés depuis plusieurs années par un certain nombre d'auteurs (voir les références). Présentés souvent comme une généralisation des ultramétriques, leur utilisation en analyse des données a été restreinte du fait de la complexité des algorithmes à mettre en oeuvre, pour les construire à partir d'une matrice de dissimilarité.

Une méthode classique [13] pour approcher une matrice de dissimilarité par un arbre additif consiste à construire dans un premier temps la structure de l'arbre, puis dans un deuxième temps à calculer les quantités valuant les arêtes de l'arbre.

L'approche qui est présentée ici est fondée sur la décomposition de toute matrice de distance quadrangulaire en la somme d'une distance ultramétrique et d'une distance à centre.Pour réaliser l'approximation on utilisera les produits scalaires dans la métrique $\sqrt{D}$, métrique qui semble appropriée à l'étude des matrices de distances quadrangulaires.

En conséquence de ces deux choix, la procédure d'approximation qui est proposée permet de trouver tout d'abord la composante distance à centre, puis ensuite la composante ultramétrique à l'inverse des procédures classiques. On obtient ainsi des algorithmes beaucoup plus rapides,$O(n^2)$ au lieu de $O(n^4)$, et possédant certaines propriétés d'optimalité.

---

*U.E.R. des Sciences et Techniques - Université Rennes 2 Haute-Bretagne.

II - PROPRIETES DES DISTANCES QUADRANGULAIRES

On appelle E l'ensemble de cardinalité n des éléments sur lesquels les distances sont définies.

2.1.Définition

Soit Q une matrice de dissimilarité n x n, on dira qu'elle est quadrangulaire si et seulement si elle vérifie pour tout quadruplet l'inégalité du même nom. C'est à dire :

$\forall\ i, j, k, l \in E \quad Q_{ij} + Q_{kl} \leq \max (Q_{ik} + Q_{jl}, Q_{il} + Q_{jk})$

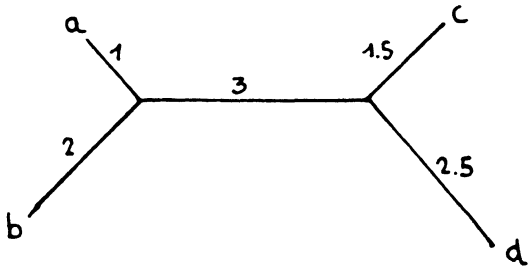
Il existe dans la littérature plusieurs démonstrations (voir [1], [6], [11]), de l'équivalence entre les distances quadrangulaires et les arbres additifs.

Plus précisément, un arbre additif sur un ensemble E est un graphe valué, connexe et sans cycle tel que :

- les sommets pendants (ou noeuds terminaux) de l'arbre sont des éléments de E. Notons que des éléments de E peuvent être des sommets intérieurs (ou noeuds intérieurs)

- la distance entre deux éléments de E est la longueur de l'unique chemin les reliant. La longueur d'un chemin étant, par définition, la somme des valeurs des arêtes composant ce chemin.

Exemple : E = {a, b, c, d}



Arbre additif

$Q =$

	a	b	c	d
a	0	3	5.5	6.5
b		0	6.5	7.5
c			0	4
d				0

Matrice de distance quadrangulaire Q

Figure 1 . Représentation des arbres additifs.

On vérifie aisément que l'ensemble des matrices de distances quadrangulaires

- contient celui des distances ultramétriques, c'est-à-dire les distances U, telles que $\forall i,k,j \in E \quad U_{ij} \leq \max (U_{ik}, U_{kj})$

- contient celui des distances à centres, c'est-à-dire les distances C, telles qu'il existe un vecteur X tel que :

$$\forall\ i \neq j\ C_{ij} = X_i + X_j \text{ et } C_{ii} = 0$$

- est contenu dans l'ensemble des distances, c'est-à-dire les dissimilarités T telles que $\forall\ i, j, k \in E\ \ T_{ij} \leq T_{ik} + T_{kj}$

2.2. Lien avec les distances Euclidiennes

2.2.1. Rappels sur les distances Euclidiennes

Soit F un espace affine euclidien de dimension finie et D une dissimilarité sur E fini. On appelle image Euclidienne dans F de (E,D), ou encore de D, une famille de points de F notés (M_i), $i \in E$ et vérifiant :

$$\text{pour tout } i, j \in E \qquad |\overrightarrow{M_i M_j}| = D_{ij}$$

On appelle distance Euclidienne, toute distance admettant une image Euclidienne.

Pour caractériser les distances Euclidiennes on introduit la forme $W_k(D)$ définie par :

$$\text{pour tout } i, j, k \in E \quad W_k(D)_{ij} = \tfrac{1}{2}(D_{ki} + D_{kj} - D_{ij})$$

Un résultat classique (voir par exemple [2]) montre que D est une distance Euclidienne si et seulement s'il existe k appartenant à E tel que $W_k(D^2)$ soit semi définie positive (s.d.p.), de plus si la forme $W_k(D^2)$ est s.d.p. pour k, elle l'est aussi en tout point M appartenant à l'espace euclidien, en particulier le centre de gravité des points M_i.

Alors la dimension de l'espace (et par abus la dimension de D) est égale au rang de la matrice $(W_k(D^2))_{ij}$. Cette matrice s'interprète alors comme la matrice des produits scalaires des vecteurs d'origine M_k, soit $W_k(D^2)_{ij} = \overrightarrow{M_k M_i} . \overrightarrow{M_k M_j}$

2.2.2. Propriétés Euclidiennes des distances quadrangulaires

Un résultat classique dû à Holman [9] montre que les distances ultramétriques sont Euclidiennes de dimension (n - 1). On vérifie aisément que les distances à centre ne sont pas en général Euclidiennes, il en est donc de même pour les distances quadrangulaires. Cependant on peut montrer le résultat suivant :

PROPRIETE 1.

Soit Q une distance quadrangulaire sur E, alors Q est le carré d'une distance Euclidienne de dimension (n - 1) (Le Calvé [10]).

Autrement dit pour toute distance quadrangulaire Q il existe une distance Euclidienne D telle que $Q_{ij} = D^2_{ij}$ pour tout i et j, ou, ce qui revient au même, la matrice $(\sqrt{Q})$ définie par $(\sqrt{Q})_{ij} = \sqrt{Q_{ij}}$ est Euclidienne de dimension (n - 1).

Si on note par Φ l'ensemble des distances Euclidiennes, on notera par $\Phi^{\frac{1}{2}}$ l'ensemble des distances dont la racine carrée est Euclidienne. L'intérêt de $\Phi^{\frac{1}{2}}$ est qu'il contient non seulement les quadrangulaires, et donc les ultramétriques et les distances à centre, mais aussi la plupart des dissimilarités définies sur des attributs binaires par des indices tels que ceux de Jaccard, Sokal et Sneath, etc qui ne sont pas en général Euclidiennes [8].

2.2.3.Métrique $\sqrt{D}$

Si on considère une matrice de dissimilarité D, on dit que l'on "travaille en métrique $\sqrt{D}$", si on mesure les distances entre i et j par $\delta_{ij} = \sqrt{D_{ij}}$, et non par D_{ij}.

Dans ce contexte dire que D appartient à $\Phi^{\frac{1}{2}}$ c'est dire que D est une distance Euclidienne dans la métrique $\sqrt{D}$.

L'intérêt de ce point de vue est de pouvoir parler des produits scalaires induits par D, (quand D appartient à $\Phi^{\frac{1}{2}}$), ce sont ceux définis dans la métrique $\sqrt{D}$. Soit :

$$W_k(D)_{ij} = W_k(\delta^2)_{ij} = \tfrac{1}{2}(D_{ik} + D_{kj} - D_{ij}) = \overrightarrow{M_kM_i}.\overrightarrow{M_kM_j}$$

où M_k, M_i, M_j sont les représentations de i, j, k dans l'espace Euclidien muni de la métrique $\sqrt{D}$.

Puisque les distances quadrangulaires appartiennent à $\Phi^{\frac{1}{2}}$ on peut exprimer les produits scalaires dans la métrique $\sqrt{D}$. Alors l'inégalité quadrangulaire s'exprime et s'interprète facilement dans l'espace Euclidien de représentation muni de la métrique $\sqrt{D}$.

En effet, pour tout i, j, k, l appartenant à E on peut toujours écrire (en les renommant au besoin)

$$Q_{ij} + Q_{kl} \leq Q_{ik} + Q_{jl} = Q_{il} + Q_{jk}.$$

Comme le produit scalaire dans un quadrangle s'écrit dans la métrique usuelle sous la forme :

$$2\overrightarrow{M_iM_j}.\overrightarrow{M_kM_l} = |\overrightarrow{M_iM_l}|^2 + |\overrightarrow{M_jM_k}|^2 - (|\overrightarrow{M_iM_k}|^2 + |\overrightarrow{M_jM_l}|^2),$$

dans l'espace Euclidien munis de la métrique $\sqrt{D}$ on aura $2\overrightarrow{M_iM_j}.\overrightarrow{M_kM_l} = 0$ car par définition on a $Q_{ij} = |\overrightarrow{M_iM_j}|^2$ pour tout i et j.

Donc, si Q vérifie l'inégalité quadrangulaire, pour tout quadrilatère M_i, M_j, M_k, M_l on a :

$$\overrightarrow{M_iM_j}.\overrightarrow{M_kM_l} = 0 \text{ et } \overrightarrow{M_iM_k}.\overrightarrow{M_jM_l} = \overrightarrow{M_iM_l}.\overrightarrow{M_jM_k} \geq 0.$$

Ceci signifie que dans tout quadrilatère il existe deux côtés opposés orthogonaux. Si on a une distance vérifiant seulement l'inégalité triangulaire, tous les produits scalaires entre cotés adjacents, en terme de $\sqrt{D}$, sont positifs. Si la distance est ultramétrique, il y a deux produits scalaires égaux et inférieurs au troisième dans tout triangle.

Remarquons que l'additivité des distances est équivalente à l'orthogonalité dans l'espace de représentation muni de la métrique $\sqrt{D}$, en effet $\overrightarrow{M_iM_j} \perp \overrightarrow{M_jM_k} \Longleftrightarrow D_{ij} + D_{jk} = D_{ik}$; cette propriété est utile pour les optimisations.

2.3. Décomposition d'une quadrangulaire Q = U + C

La décomposition de toute matrice de distance quadrangulaire Q, en une somme d'une ultramétrique U et d'une distance à centre C est citée par Sattah et Tversky [13] et par Caroll [3]. Ce dernier l'attribue à un papier non publié de S.A. Farris. Rappelons qu'une dissimilarité C est appelée distance à centre si et seulement si il existe un vecteur X tel que $C_{ij} = X_i + X_j$ si $i \neq j$ et 0 autrement.

Nous donnons ici une démonstration de cette décomposition permettant de construire la famille des ultramétriques et des distances à centre correspondant à une distance quadrangulaire Q donnée.

PROPRIETE 2

Une condition nécessaire et suffisante pour qu'une matrice de distance quadrangulaire vérifie l'inégalité ultramétrique est qu'il existe un point r de l'arbre additif associé à Q, tel que $Q_{ir} = Q_{jr}$ $\forall$ i,j $\in$ E. Le point r est alors la racine (ou le sommet) de l'ultramétrique.

Démonstration :

Condition suffisante : soit r la racine telle que $\forall$ i, j $\in$ E $Q_{ir} = Q_{jr}$. Considérons la matrice Q^+, (n+1 x n+1) , extension de Q à l'ensemble E + {r}, obtenue en rajoutant la ligne et la colonne Q_{ir}. Puisque r appartient à l'arbre, Q^+ vérifie l'inégalité quadrangulaire, en particulier :

$\forall$ i,j,k $\in$ E $\quad Q^+_{ij} + Q^+_{kr} \leq \max (Q^+_{ik} + Q^+_{jr}, Q^+_{jk} + Q^+_{ir})$

comme $Q^+_{kr} = Q^+_{ir} = Q^+_{jr}$, on a

$\forall$ i,j,k $\in$ E $\quad Q_{ij} \leq \max (Q_{ik}, Q_{jk})$ car Q et Q^+ coïncident sur E.

Donc Q est ultramétrique.
La condition nécessaire est une évidence.■

PROPRIETE 3.

Quelque soit Q une matrice de distance quadrangulaire et quelque soit C une matrice de distance à centre, alors pour tout $\alpha \in \mathbb{R}$ la matrice $Q + \alpha C$ vérifie l'inégalité quadrangulaire.

Démonstration :

il est immédiat que si $Q_{ij} + Q_{kl} \leq Q_{ik} + Q_{jl} = Q_{il} + Q_{jk}$
alors on a également :

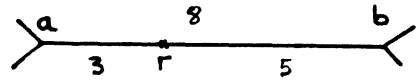
$$Q_{ij}+\alpha X_i+\alpha X_j+Q_{kl}+\alpha X_k+\alpha X_l \leq Q_{ik}+Q_{jl}+\alpha X_i+\alpha X_k+\alpha X_j+\alpha X_l$$

$$= Q_{jk} + Q_{il} + \alpha X_i + \alpha X_k + \alpha X_j + \alpha X_l \quad \blacksquare$$

On déduit de ces propriétés le théorème de décomposition :

THEOREME 1 Toute matrice de distance Q vérifiant l'inégalité quadrangulaire peut se décomposer en la somme d'une matrice de distance ultramétrique U et d'une matrice de distance à centre C. Cette décomposition peut se faire d'une infinité de façons dépendant de deux paramètres : la racine r et une constante $K \in \mathbb{R}$
$\forall Q \quad \exists U(r, K)$ et $C(r, K)$ t. q. $Q = U(r, K) + C(r, K)$

Démonstration.
Soit r une racine de Q c'est-à-dire un point quelconque de l'arbre additif associé à Q. On entend par point quelconque soit un noeud existant de l'arbre, soit un noeud fictif, c'est-à-dire un noeud intérieur rajouté sur une arête sans modifier les distances.

exemple : 

a et b noeuds intérieurs, r noeud, fictif $d_{ar} + d_{rb} = d_{ab}$

Soit r une racine et K une constante arbitraire.

Posons, $\forall i \; X_i = Q_{ir} - K$, et $X_r = 0$
$C_{ij} = X_i + X_j$ si $i \neq j$, $C_{ii} = 0$
et $D = Q - C$

Alors D est une distance quadrangulaire d'après la propriété 3, et comme $D_{ir} = K$ pour tout i, D est une ultramétrique d'après la propriété 2 $\blacksquare$

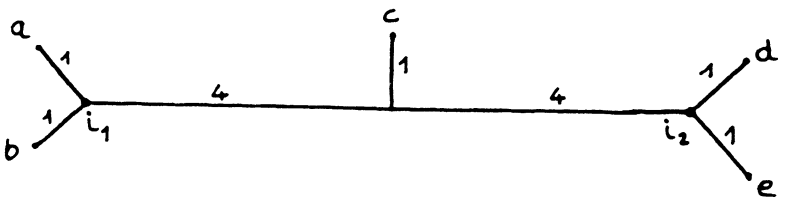
On obtient ainsi une famille de décompositions dépendantes de r et de K. On notera $\mathcal{U}^Q$ et $\mathcal{C}^Q$ les familles d' ultramétriques et de distances à centre appartenant à une décomposition de Q, $U(r,K) \in \mathcal{U}^Q$ et $C(r,K) \in \mathcal{C}^Q$

On vérifie facilement que $\mathcal{U}^Q$ et $\mathcal{C}^Q$ contiennent toutes les décompositions de Q de la forme U + C.

Pour r fixé on peut toujours choisir K de façon à ce que, soit U soit C ait tous ses éléments positifs.

Pour avoir U à éléments positifs il suffit de choisir K tel que : $K \geq \max_j (Q_{jr})$, j étant un noeud intérieur à l'arbre associé à Q. Pour avoir C à éléments positifs il sufffit de choisir K tel que $K \leq \max_i (Q_{ir})$, i étant un noeud terminal.

On retrouve ainsi la condition énoncée par Sattah et Tversky pour qu'il existe une décomposition dont tous les termes soient à éléments positifs: il faut et il suffit qu'il existe une racine r telle que la distance à n'importe quel noeud intérieur n'excède pas la distance à n'importe quel noeud terminal. Il est évident que cette condition n'est pas toujours satisfaite comme le montre le contre-exemple suivant :



Quelque soit l'emplacement de la racine r il existe toujours un noeud intérieur i ($i = i_1$ ou i_2) tel que $Q_{ir} > Q_{rc}$ donc cette quadrangulaire n'admet pas de décomposition dont tous les éléments soient positifs.

Le choix de la racine est donc un élément primordial dans la décomposition d'une quadrangulaire.

2.4. Choix de la racine (ou centre)

Plusieurs façons de procéder, liées à différents critères sont possibles pour choisir une racine dans un arbre additif.

2.4.1. La racine de l'arbre est "le milieu d'un diamètre".

Autrement dit r est tel qu'il minimise : $(\min_s (\max_i Q_{is}))$ où i est un noeud terminal et s un point de l'arbre. Dans ce cas on cherche à minimiser la plus grande distance entre un élément terminal et le centre. Ainsi on cherche à se rapprocher de la situation de l'ultramétrique.

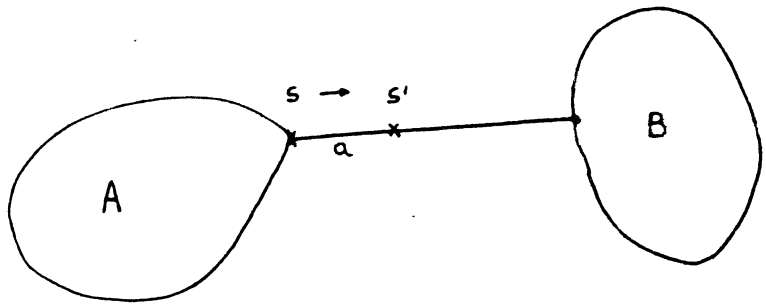
Pour ce problème la solution est évidente, il suffit de considérer $Q_{ab} = \max_{k,l \in E} (Q_{kl})$

et prendre pour r le milieu du chemin reliant a à b.

2.4.2. La racine de l'arbre est la "médiane"

Dans ce cas on cherche R minimisant $\min_s (\sum_{i \in E} Q_{is})$ avec les mêmes notations que précédemment.

Considérons une racine s et une arête passant par s, cette arête partitionne l'ensemble E en deux classes A et B de cardinaux respectifs n_A et n_B.



Si on note par f(s) la quantité $\sum_{i \in E} Q_{is}$, en déplaçant s en s' le long de l'arête issue de s d'une quantité "a" on a :

$$f(s') = f(s) + n_A a - n_B a$$

$$\text{ou} \quad f(s') - f(s) = a(n_A - n_B)$$

Donc pour rechercher le minimum sur s de f(s), partant d'un point quelconque s, on chemine sur l'arbre en choisissant les arêtes qui font diminuer f(s). C'est-à-dire que l'on chemine vers l'ensemble le plus important ($n_A \leq n_B$). S'il existe une arête partitionnant l'arbre en deux sous-ensembles de tailles égales, le minimum est atteint sur un point quelconque de cette arête, d'où le terme de "médiane". Sinon le minima est atteint sur un noeud intérieur.

2.4.3. La racine est la "moyenne"

Alors on cherche R minimisant $\min_s \left(\sum_{i \in E} Q^2_{is} \right)$ avec les mêmes notations. En reprenant l'argument précédent avec $f(s) = \sum_{i \in E} Q^2_{is}$ on a

$$f(s') - f(s) = a^2 n + 2a\left(\sum_{i \in A} Q_{is} - \sum_{i \in B} Q_{is} \right)$$

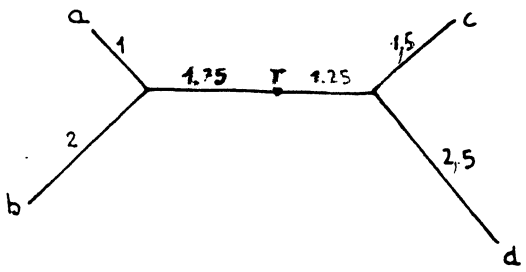
On obtient donc le minimum par cheminement simple sur l'arbre de façon à équilibrer $\sum_{i \in A} Q_{is}$ et $\sum_{i \in B} Q_{is}$, la somme de ces deux termes étant constante.

2.5. Exemple de décomposition

En reprenant la matrice de distance quadrangulaire, du paragraphe §2.1 , on choisit pour r le milieu de l'étendue :

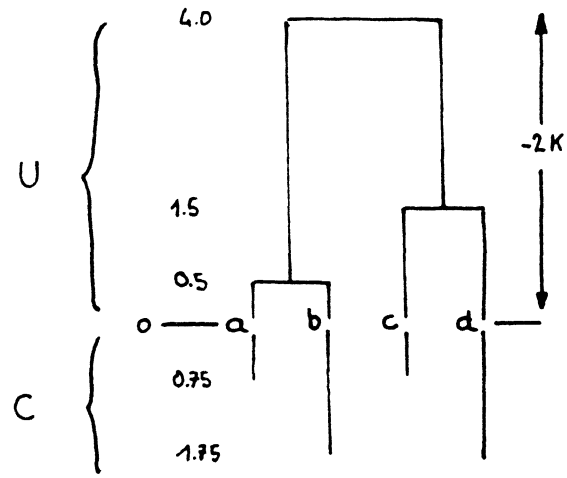
$Q_{ar} = 2.75$, $Q_{br} = 3.75$, $Q_{cr} = 2.75$, $Q_{dr} = 3.75$

Si on choisit de façon arbitraire K = -2 on obtient dans le cas une décomposition à valeurs positives.
$X_a = 0.75$, $X_b = 1.75$, $X_c = 0.75$, $X_d = 1.75$
On calcule U = Q - C



Q =

	a	b	c	d
a	0	3	5.5	6.5
b		0	6.5	7.5
c			0	4
d				0



C =

	a	b	c	d
a	0	2.5	1.5	2.5
b		0	2.5	3.5
c			0	2.5
d				0

U =

	a	b	c	d
a	0	0.5	4	4
b		0	4	4
c			0	1.5
d				0

III - CONSTRUCTION D'UN ARBRE ADDITIF A PARTIR D'UNE QUADRANGULAIRE

En général on ne connaît pas l'arbre additif mais la matrice de de distance quadrangulaire. Or pour construire l'arbre à partir de Q la littérature nous renvoie sur les démonstrations des théorème d'équivalences entre la matrice de distance et l'arbre. Ce qui implique des procédures de complexité $O(n^4)$ à cause de l'examen des quadruplets, intervenant dans l'inégalité quadrangulaire.

Pour utiliser le théorème de décomposition il est nécessaire de choisir de façon pertinente une racine, or dans deux des trois définitions précédentes du centre il est nécessaire de connaître l'arbre additif. Seul le milieu du diamètre, comme choix de r, ne fait intervenir que la matrice de distance.

Pour construire l'arbre additif associé à une distance quadrangulaire nous allons donc utiliser simplement le théorème de décomposition au point r, milieu du diamètre.

3.1. Algorithme de construction

1 - Soit $Q_{ab} = \max_{i,j}(Q_{ij})$ si Q_{ab} n'est pas unique on en choisit un arbitrairement.

2 - On pose r, racine de l'arbre, défini par $Q_{ar} = Q_{br} = \frac{1}{2}Q_{ab}$

3 - On calcule X_i défini par :

$X_a = X_b = \frac{1}{2}Q_{ab}$

$X_i = \max(Q_{ia}, Q_{ib}) - \frac{1}{2}Q_{ab}$

On peut ajouter ou retrancher une constante K quelconque aux X_i de façon à les rendre les plus petits possible. Un choix possible est de les centrer en posant $K = -X.$ avec $X. = (1/n)\sum_{i\in E} X_i$

4 - On calcule $C_{ij} = X_i + X_j$ si $i \neq j$, $= 0$ autrement.

5 - On calcule $U_{ij} = Q_{ij} - C_{ij}$. U est une ultramétrique à une constante près. En effet U_{ij} peut être négatif.

6 - On utilise un des nombreux algorithmes classiques pour représenter la hiérarchie à laquelle on rajoute la distance à centre pour obtenir l'arbre additif.

Remarquons qu'il est difficile de procéder en sens inverse. En effet rechercher directement l'ultramétrique est généralement impossible par un des algorithmes d'agrégation.

L'exemple suivant montre que les deux éléments les plus proches au sens de la matrice de distance quadrangulaire ne forment pas forcément une "classe" de l'arbre additif.

(Rappelons qu'une "classe" d'un arbre additif est un sous-ensemble de E obtenu en supprimant une seule arête de l'arbre et que les classes des hiérarchies appartenant aux décompositions de Q sont des classes de l'arbre additif).

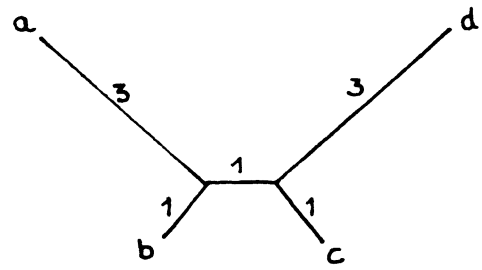


figure 4

On a dans ce cas :

$$Q_{bc} = \min_{i,j \in E} (Q_{ij})$$

Il n'est donc pas possible d'utiliser les algorithmes de classification ascendante hiérarchique sans avoir au préalable enlevé la composante distance à centre. Nous procéderons selon le même principe pour l'approximation d'une dissimilarité par une quadrangulaire.

3.2. Calcul de la valeur des arêtes de l'arbre

Une fois reconstruit l'arbre par la représentation de l'ultramétrique, il reste à valuer les arêtes de l'arbre additif.

Deux cas sont à considérer selon que l'arête a pour une extrêmité un noeud terminal ou non:

à l'arête PQ de l'arbre additif correspondent les noeuds P et Q de l'arbre hiérarchique :

- si P et Q sont des noeuds intérieurs alors $l(P,Q) = \frac{1}{2}|U_P - U_Q|$ en notant par $l(P,Q)$ la longueur de l'arête PQ et par U_P (et U_Q) le niveau du noeud P(et Q) dans l'arbre hiérarchique.

- si P est un noeud terminal et Q un noeud intérieur, alors $l(P,Q) = X_P + \frac{1}{2}U_Q$, X_P étant la valeur associée à P dans la distance à centre.

Nota : si P et Q sont les derniers niveaux de l'ultramétrique, alors on a $l(P,Q) = l(P,r) + l(r,Q)$ ou r est la racine de l'arbre.

3.3. Remarques sur les valeurs négatives de U et de C.

En utilisant la décomposition de Q en U + C nous arrivons fréquemment sur des matrices de distances U et C admettant des valeurs négatives. Ceci pose des problèmes, en particulier parce que l'inégalité triangulaire implique la positivité des distances.

En fait nous n'utilisons qu'une translation de valeur K. Nous étendrons donc la notion de distance à celle de distance relative de la façon suivante.

Définition

D est une matrice de distance relative, si et seulement si il existe une constante positive $K \in \mathbb{R}^+$ telle que la matrice D + K définie par $(D + K)_{ij} = D_{ij} + K$ si $i \neq j$ et $(D + K)_{ii} = 0$ soit une matrice de distance.

De la même façon on étend la notion de distance ultramétrique et de distance à centre aux distances ultramétriques relatives et aux distances à centre relatives, leurs significations restent la même si on considère que les valeurs D_{ij} sont mesurées sur une échelle dont l'origine est -K, au lieu d'être 0.

Dans tous les cas la diagonale reste nulle.

Dans le paragraphe suivant nous supposerons que toutes les distances sont des distances relatives à l'exception de celle prise en donnée et du résultat qui doivent être à valeurs positives pour des raisons évidentes d'interprétation.

IV - APPROXIMATION D'UNE DISSIMILARITE PAR UNE QUADRANGULAIRE

Le problème central est évidemment l'ajustement d'une matrice de distance vérifiant l'inégalité quadrangulaire à une matrice de dissimilarité donnée. Tous les algorithmes existant consistent à rechercher d'abord la structure de l'arbre par un procédé ou par un autre, puis à ajuster la valeur des arêtes. C'est évidemment la recherche de la structure qui pose le plus de problèmes.

L'algorithme le plus utilisé (ou le plus connu ?) est ADDTREE proposé par Sattah et Tversky [13]. La recherche de la structure de l'arbre se fait en recherchant les paires d'éléments qui y seront voisins. Pour cela on compte pour chaque paire (x,y) le nombre de paires (u,v) où dans le quadruplet (x,y,u,v) x et y sont voisins au sens de l'inégalité quadrangulaire.

Cet algorithme est donc au minimum d'ordre n^4 et ne cherche ni ne garantit aucune optimalité, aucun critère d'optimalité n'étant même exprimé.

L'algorithme que nous proposons consiste à enlever dans un premier temps la composante distance à centre, puis dans un deuxième temps à rechercher l'ultramétrique qui nous donne ainsi la structure de l'arbre. Nous construisons ainsi directement une forme décomposée de la quadrangulaire, ce qui en facilitera la représentation.

4.1. Critère d'approximation

Comme dans le cas de l'analyse factorielle d'un tableau de distance nous chercherons à minimiser les moindres carrés entre les matrices de "produits scalaires" associés aux distances.

Dans notre cas, les distances quadrangulaires appartenant à $\Phi^{\frac{1}{2}}$ et pas à Φ nous considèrerons les produits scalaires au point m dans la métrique $\sqrt{D}$, soit $W_m(Q)$.

Pour une dissimilarité donnée nous considérons la forme $W_m(D)$ qui représentera les produits scalaires si D appartient à $\Phi^{\frac{1}{2}}$ (ce qui est le cas pour la plupart des dissimilarités définies sur signes de présence-absence). Si D n'appartient pas à $\Phi^{\frac{1}{2}}$ la forme $W_m(D)$ ne sera pas s.d.p.

D étant donné nous chercherons donc la matrice de distance quadrangulaire Q* qui réalise le minimum de $\|W_m(D) - W_m(Q)\|^2$. Le point m origine des produits scalaires étant arbitraire, le critère dépendra du point m.

4.2. Décomposition du critère

Comme toute quadrangulaire se décompose de façon additive et que l'additivité équivaut à une orthogonolité dans la métrique $\sqrt{D}$ nous allons pouvoir décomposer le critère en deux composantes indépendantes, séparant l'optimisation sur l'ultramétrique de celle sur la distance à centre.

On a le théorème suivant :

THEOREME 2 Soit D une matrice de dissimilarité n x n, soit sa décomposition en un point m définie par $D = D' + C'$ avec

$$C'_{ij} = X'_i + X'_j \text{ si } i \neq j$$
$$= 0 \text{ si } i = j$$

où $X'_i = D_{mi}$, et $D'_{ij} = D_{ij} - D_{mi} - D_{mj}$ si $i \neq j$
$= 0$ si $i = j$

Soit Q une matrice de distance quadrangulaire n x n, soit sa décomposition $Q = U(r,K) + C(r,K)$, avec $K = 0$ et r une racine quelconque - on notera donc $U(r,K)$ et $C(r,K)$ par U_r et C_r avec $(C_{rij} = X_i + X_j)$

Alors $\forall$ D, Q, m, r on a

$$\|W_m(D) - W_r(Q)\|^2 = \tfrac{1}{4} \|D' - U_r\|^2 + \|X' - X\|^2$$

Démonstration

D'une part $W_m(D) = W_m(D') + W_m(C')$ où

pour $i \neq j$, $W_m(C')_{i,j} = 0$ et $W_m(D')_{i,j} = -\tfrac{1}{2}D'_{ij}$

pour $i = j$, $W_m(C')_{i,i} = X'_i$ et $W_m(D')_{i,i} = 0$

D'autre part $W_r(Q) = W_r(U_r) + W_r(C_r)$ où

pour $i \neq j$, $W_r(C_r)_{i,j} = 0$ et $W_r(U_r)_{i,j} = -\tfrac{1}{2}U_{rij}$

pour $i=j$, $W_r(C_r)_{i,i} = X_i$ et $W_r(U_r)_{i,i} = 0$

Donc $\|W_m(D) - W_r(Q)\|^2 = \sum_i (W_m(D) - W_r(Q))^2 + \sum_{i \neq j} (W_m(D) - W_r(Q))^2$

$= \sum_i (X'_i - X_i)^2 + \sum_{i \neq j} (-\frac{1}{2}D'_{ij} + \frac{1}{2}U_{ij})$

$= \|X' - X\|^2 + \frac{1}{4}\|D' - U\|^2$ ■

4.3. L'optimisation au point m

Il s'en suit que la solution du problème, trouver Q tel que $\|W_m(D) - W_m(Q)\|$ soit minimal est immédiatement fournie par

$Q^* = U^* + C^*$, où $C^*_{i,j} = X^*_i + X^*_j$

avec $X^*_i = X'_i$, et U^* est l'ultramétrique des moindres carrés associée à D' (et non pas à D).

Comme en général on ne sait pas calculer simplement U^* on approchera U^* par $\bar{U}$ qui est l'ultramétrique de la moyenne (Average Linkage). $\bar{U}$ est simple à calculer et fournit une approximation empirique généralement satisfaisante de U^* (voir sur ce sujet [4], [3]).

4.4. Positivité de la solution

La solution construite $Q = \bar{U} + C^*$ conduit à un arbre additif dont il est nécessaire de s'assurer que toutes les arêtes ont des valuations positives ou nulles pour que la solution ait un sens. On a le résultat suivant au point m.

THEOREME 3 Soit D une matrice de dissimilarité n x n sur l'ensemble E. et soit m un point quelconque (n'appartenant pas forcément à E) vérifiant $\forall$ i, j $\in$ E $D_{ij} \geq (D_{mi} - D_{mj})$
Alors $Q = \bar{U} + C^*$, solution approchée du problème précédent, peut se représenter sous la forme d'un arbre additif dont toutes les arêtes ont des valeurs positives.

Démonstration

- Pour les arêtes intérieures, la valeur de celles-ci est égale à $\frac{1}{2}(\bar{U}_k - \bar{U}_{k-1})$ $\bar{U}_k$ et $\bar{U}_{k-1}$ sont les niveaux de l'ultramétrique (voir §3-2). L'ultramétrique de la moyenne ne présentant jamais d'inversion on a toujours $\bar{U}_k \geq \bar{U}_{k-1}$

- pour les arêtes extérieures, leur valeur est donnée par $l(ip) = D_{mi} + \frac{1}{2}\bar{U}_{ip}$ où i $\in$ E et P est le noeud intérieur de l'arbre où se rattache le point i.

Par définition de l'ultramétrique de la moyenne $\bar{U}_{ip} = (1/n_A) \sum_{j \in A} D'_{ij}$
où A est l'ensemble auquel se rattache l'élément i au niveau $\bar{U}_{ip}$ dans l'ultramétrique U. Donc $V_{ip} = (1/2n_A) \sum_{j \in A} (D_{ij} + D_{mi} - D_{mj})$
qui est toujours positif si $D_{ij} \geq (D_{mi} - D_{mj})$ pour tout i et j ■

Si le point m est un point de l'étude initiale il suffit pour satisfaire la condition de positivité que D vérifie l'inégalité triangulaire pour les triplets comprenant le point m. (Vérification de complexité $o(n^2)$).

V - ALGORITHME DE CONSTRUCTION DE Q

Le critère dépendant du point m, et le point m devant appartenir à l'arbre solution on est amené à choisir un point m élément quelconque de l'ensemble E.

1) On calcule $C_{ij} = D_{mi} + D_{mj}$ pour $i \neq j$, et $C_{ii} = 0$

2) On calcule $D'_{ij} = D_{ij} - C_{ij}$ pour $i \neq j$, et $D_{ii} = 0$

3) Comme, si D satisfait la condition de positivité, $D'_{ij} \leq 0$ pour tout i et tout j, on peut, pour des commodités de programmation de l'ultramétrique de la moyenne, la rendre positive en lui rajoutant la constante $K = \min_{i,j} (D'_{ij})$

4) On applique l'algorithme de CAH du lien moyen à D' (ou D'+ K) obtenant ainsi $\bar{U}$ (ou $\bar{U}$ + K).

5) On compose $Q = \bar{U} + C$

6) On calcule les valeurs des arêtes pour représenter l'arbre voir 3.2.

L'algorithme est donc comme les algorithmes d'agrégations d'ordre n^2. D'un point de vue programmation, il suffit d'utiliser les programmes de CAH classique en rajoutant un module de transformation de D par soustraction d'une distance à centre C.

CONCLUSION

La structure d'arbre additif est très intéressante en analyse des données car elle est facile à représenter et à lire. Son emploi paraît justifié dans de nombreux cas, mais il faut noter particulièrement les données issues de variables binaires pour lesquelles les indices de dissimilarité sont euclidiens dans la métrique $\sqrt{D}$ comme les arbres additifs.

L'algorithme qui vient d'être décrit nécessitant les mêmes calculs qu'une classification ascendante hiérarchique devrait inciter les utilisateurs à choisir la structure d'arbre additif

de préférence à celle de la hiérarchie pour représenter leurs données. En effet, la structure d'arbre additif est beaucoup plus riche (2n - 3 paramètres) que celle de la hiérarchie (n - 1 paramètres), et donc, en principe, beaucoup plus fidèle aux données.

De ce point de vue, elle est comparable aux représentations des points dans le plan (2n paramètres), voir [12] pour une comparaison des deux approches.

BIBLIOGRAPHIE

[1] BUNEMAN P.,The recovery of trees from measures of dissimilarity,in F.R Hodson, D.G.kendall,P.Tautu, eds, Mathematics in Archeological and Historical Sciences, Edinburg University Press, 1971 (387-395)

[2] CAILLEZ F., et PAGES J.P., Introduction à l'analyse des données , S.M.A.S.H. Paris (1976)

[3] CAROLL J.D., Spatial, Non-Spatial and hybrid models for scaling. Psychometrika-vol 41 - N°4 - (1976)

[4] CHANDON J.L., Construction de l'ultramétrique la plus proche au sens des moindres carrés :approximation versus optimisation.Tech.Rep.n°123,I.A.E.Aix en Provence (1978)

[5] CHANDON J.L.,LEMAIRE J., POUGET J. Construction del'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés . RAIRO série recherche opérationnelle , 14 (1980)

[6] DOBSON A., Unrooted trees for taxomomy. J.Appl.Prob.11 (1974)

[7] FICHET B, Analyse factorielle sur tableaux de dissimilarité Thèse d'Etat, Univ.Aix-Marseille 2 - (1983)

[8] FICHET B.,LE CALVE G., Structure géométrique des principaux indices de dissimilarités sur signes de présence-absence (à paraître dans Stat et Anal.des données)

[9] HOLMAN W., The relation between hierarchical and Euclidean models for psychological distances Psychometrika, Vol 37 n°4 (1972)

[10] LE CALVE G. Distances à centre (à paraître dans Stat. et Anal.des données)

[11] PATRINOS A.,HAKIMI J.L., The distance matrix of a graph and its tree realization. Quaterly of applied mathematics - Vol.30 - n°3 - (1972)

[12] PRUZANSKY S., TVERSKY A., CAROLL J.D., Spatial versus tree representations of Proximity data .Psychometrika Vol.47, n°1 (1982)

[13] SATTAH S., TVERSKY A., Additive similarity trees Psychometrika, Vol.42.n°3 (1977)