Statistique et analyse des données

Georges Le Calve

Distance à centre

Statistique et analyse des données, tome 10, n° 2 (1985), p. 29-44

<http://www.numdam.org/item?id=SAD_1985__10_2_29_0>

Statistiques et Analyse de Données
Septembre 1985 Vol. 10 n° 2 pp. 29-44

DISTANCE A CENTRE

Georges LE CALVE

U.E.R des Sciences et Techniques
Université de RENNES 2 HAUTE-BRETAGNE
6, avenue Gaston Berger
35043 - RENNES Cédex

Résumé : *Dans de nombreux systèmes la communication directe entre les éléments n'est pas possible, mais se fait par l'intermédiaire d'un centre, standard, leader .... Il en découle une distance très particulière difficile à mettre en évidence par les techniques classiques car elle n'est ni euclidienne ni ultramétrique.*
Le premier paragraphe en étudie les propriétés, le second s'intéresse aux problèmes d'optimisations et de décompositions et le troisième montre le rôle "charnière" joué par les distances à centre.
Cette notion permet d'aborder d'un point de vue unifié des problèmes aussi divers que la constante additive, le modèle d'analyse factorielle au sens de Thurstone, les arbres additifs etc....

Summary : *In numerous systems direct communication between elements is impossible, but it takes place through the medium of a center, a standard, a leader .... The result is a distance which is particulary difficult to put in evidence through traditional techniques, for it is neither euclidian nor ultrametric.*
The first paragraph studies its properties, the second concerns optimisation and decomposition problems, and the third shows the pivot part played by "star distances".
This notion makes it possible to taekle such varied problems as additive constant, factor analysis - with Thurstone's meaning, additive trees, etc..., from a unified viewpoint.

Mots clès : Tableaux de distances

Indices de classification ISI : 06-000, 06-090, 06-120.

Manuscrit reçu le 4 février 1985
révisé le 24 octobre 1985

INTRODUCTION.

Dans de nombreux systèmes la communication directe entre les éléments n'est pas possible, mais se fait par l'intermédiaire d'un centre, standard, leader etc.... C'est le cas, par exemple, des communications téléphoniques, des rapports entre périphériques d'un même ordinateur, des transferts d'informations entre 2 parties du corps par l'intermédiaire du cerveau, des échanges politico économiques à l'intérieur de certains blocs, et, pour une bonne part, de la distance ferroviaire en France. Dans de tels systèmes la distance entre 2 individus est la somme de leur distance au centre.

Ni euclidienne ni ultramétrique, cette composante est difficile à mettre en évidence dans les techniques classiques. Elle est, en général, considérée comme une déformation par rapport au modèle explicatif ou comme un bruit. Or, nous pensons que, dans de nombreux problèmes, cette composante contient un facteur explicatif important ; d'où la nécessité d'étudier ses propriétés.

1. DEFINITIONS. PROPRIETES

Définition 1 : Distance à centre.

Soit $I = \{1, 2, \ldots, n\}$ et C une application de $I \times I$ dans $\mathbb{R}^+$. C sera dite une distance à centre si et seulement si il existe un vecteur $X \in \mathbb{R}^{+n}$ tel que $C_{ij} = X_i + X_j$, $i \neq j$, $C_{ii} = 0 \ \forall i$.

Il est évident que C vérifie les axiomes d'une distance. Il est, de plus, toujours possible de supposer que le point o tel que $C_{io} = X_i$, $X_o = 0$ appartient à I, ne serait-ce qu'en le rajoutant à l'étude. Cette hypothèse sera supposée vérifiée dans la suite.
Deux extensions de la notion de distance à centre s'avèrent utiles.

Définition 2 : Eloignement à centre.

C, application de $I \times I$ dans $\mathbb{R}^+$ est un éloignement à centre si et seulement si il existe $X \in \mathbb{R}^{+n}$ tel que $C_{ij} = X_i + X_j$, $\forall i$, $\forall j$. En particulier $C_{ii} = 2X_i$.

La notion d'éloignement à centre a l'avantage de préserver une notion de

continuité sur I. Par contre, bien que la symétrie et l'inégalité triangulaire restent vraies, ce n'est plus une distance puisque $C_{ii} \neq 0$ si $i \neq 0$.

Il est également possible de lever la restriction $X \in \mathbb{R}^{+n}$ en la remplaçant par $X \in \mathbb{R}^n$. Le fait que C_{ij} ne soit pas toujours positif ou nul entraîne que l'inégalité triangulaire n'est pas vérifiée. Nous parlerons dans ce cas de "distance à centre signée".

Définition 3 : Dissimilarité à centre.

C, application de I×I dans $\mathbb{R}^+$ est une dissimilarité à centre si et seulement si il existe $X \in \mathbb{R}^n$ tel que $C_{ij} = |X_i + X_j| \quad \forall\, i \neq j,\ C_{ii} = 0,\ \forall\, i$.

C n'est plus une distance mais peut être considérée comme la "composition" de deux types de distance :
Soit $I^+ = \{i : X_i \geq 0\}$ et $I^- = \{i : X_i \leq 0\}$. Alors la restriction de C à $I^+ \times I^+$ et à $I^- \times I^-$ est une distance à centre, sa restriction à $I^+ \times I^-$ ou $I^- \times I^+$ est la distance naturelle sur $\mathbb{R}$.
On peut donc considèrer qu'une dissimilarité à centre sépare la population en deux blocs, la communication se faisant par l'intermédiaire d'un centre dans chaque bloc mais étant "directe" d'un bloc à un autre.

C =

	I^+	I^-
I^+	Distance à centre	Distance euclidienne
I^-	Distance euclidienne	Distance à centre

A l'opposé, on peut dire que la distance naturelle dans $\mathbb{R}$, définie par $|X_i - X_j|$, repose, au contraire, sur l'existence de deux blocs tels que la communication est directe à l'intérieur de chacun d'eux, mais se fait par l'intermédiaire d'un centre commun (l'origine) d'un bloc à un autre.

Propriétés

Propriété 1

Les ensembles des distances et des éloignements à centre sont stables pour les combinaisons linéaires positives.

Cette propriété, immédiate, implique que toute distance contient au maximum une composante additive à centre ; d'où son importance.

Proposition 1

Une condition nécessaire et suffisante pour qu'une distance D soit une distance à centre est que :
$\forall$ i, $\forall$j, $\forall$k distincts $D_{ij} + D_{ik} - D_{jk} = f(i)$, une fonction qui ne dépend que de i.

Condition nécessaire : $D_{ij} + D_{ik} - D_{jk} = X_i + X_j + X_i + X_k - X_j - X_k = 2\,X_i$.

Condition suffisante :

$$\left.\begin{array}{l} D_{ij} + D_{ik} - D_{jk} = f(i) \\ D_{ij} + D_{jk} - D_{ik} = f(j) \end{array}\right\} \Rightarrow 2\,D_{ij} = f(i) + f(j) \quad \text{pour } i \neq j$$

ce qui est la définition d'une distance à centre.

Propriété 2

Les distances et éloignement à centre vérifient l'inégalité quadrangulaire.

Une matrice M vérifie l'inégalité quadrangulaire si

$$\forall i,\ \forall j,\ \forall k,\ \forall l \quad m_{ij} + m_{kl} \leq \text{Max}\,(m_{ik} + m_{jl},\ m_{jk} + m_{il})$$ cf [7].

On parlera d'égalité quadrangulaire si le signe $\leq$ est remplacé par le signe =, ce qui implique l'égalité des 3 sommes considérées.
Il suffit de remarquer que :

$$C_{ij} + C_{kl} = C_{ik} + C_{jl} = C_{il} + C_{kj} = X_i + X_k + X_j + X_l$$

si les 4 indices sont distincts, et on a donc l'égalité dans ce cas. Si certains indices sont égaux, l'égalité reste vraie pour un éloignement à centre, mais devient une inégalité pour une distance à centre.

Proposition 2

La condition nécessaire et suffisante pour qu'une matrice M soit un éloignement à centre est qu'elle vérifie l'égalité quadrangulaire.

La propositon directe est contenue dans la propriété 2. Pour la réciproque, il suffit d'écrire l'égalité quadrangulaire pour le quadruplet i, j, k, l et d'y faire $k = i$, $l = j$ pour obtenir $2\, m_{ij} = m_{ii} + m_{jj}$.

Propriété 3 : Moyenne, Variance, Norme

Soit O le centre ($X_o = 0$) d'une distance ou d'un éloignement à centre C. Alors :

i) La solution du problème : trouver k tel que $1/n \sum_i C_{ik}$ soit minimum est donnée par $k = 0$.

ii) La solution du problème : trouver k tel que $1/n \sum_i C^2_{ik}$ soit minimum est donnée par $k = 0$.

iii) La variance de C, définie comme la valeur du minimum dans ii) vaut $\sigma^2(C) = 1/n\, ||X||^2$

iiii) $||C||^2 = 2n^2 [\, \sigma^2 (C) + (X.)^2]$ pour un éloignement à centre

$= 2n [(n-2)\, \sigma^2 (C) + n(X.)^2]$ pour une distance à centre,

où $X. = 1/n \sum X_i$

La démonstration est évidente et omise. Le point qu'il importe de souligner est que le "centre de gravité" est le centre ($X_0 = 0$) et non X. comme dans la distance naturelle sur $\mathbb{R}$. Il en découle que la notion de variance coïncide avec $|| X ||^2$ et non avec $|| C ||^2$ comme on en a l'habitude.

Rappel : Forme de Torgerson

Soit D une matrice $n \times n$. La matrice $W_M(D)_{ij} = 1/2(D_{Mi} + D_{Mj} - D_{ij})$ est appelée forme de Torgerson de D au point M. Elle peut être calculée en tout point M dont on connait la distance aux autres points, en particulier au centre de gravité.

$W_M(D^2)_{ij}$ n'est autre que la formule du triangle appliquée à M, i, j. Elle représente le produit scalaire des vecteurs M_i, M_j dans la métrique induite par D. La condition nécessaire et suffisante pour que D soit euclidienne est que $W_M(D^2)$ soit semi définie positive. Cette propriété est indépendante du point M.

S'il existe M tel que $W_M(D)$ soit sdp la distance δ telle que $\delta^2 = D$ est euclidienne. Nous dirons que $\sqrt{D}$ est euclidienne.

Proposition 3

Si C est une distance à centre, $\sqrt{C}$ est euclidienne de dimension n - 1 et

$$W_0(C) = \begin{bmatrix} \searrow & & 0 \\ 0 & X_i & 0 \\ 0 & & \searrow \end{bmatrix}$$ Matrice Diagonale.

Il suffit d'écrire $W_0(C)_{ij} = 1/2\ (C_{oi} + C_{oj} - C_{ij})$

$$\text{soit } W_0(C)_{ij} = \begin{cases} 1/2\ (X_i + X_j - X_i - X_j) = 0 & \text{pour } i \neq j \\ 1/2(X_i + X_j) = X_i & \text{pour } i = j \end{cases}$$

Il en découle que $W_0(C)$ est la matrice diagonale annoncée, contenant un terme nul (pour i = 0) et tous les autres positifs, d'où le résultat.

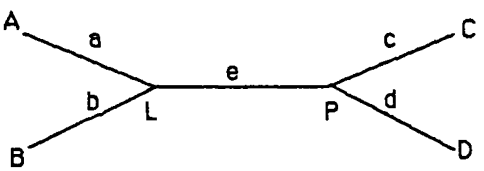
Remarque $\| W_0(C)^2 \| = \| X \|^2 = n\,\sigma^2(C)$

Cette proposition est, en fait, un cas particulier de la proposition plus générale suivante :

Proposition 4

Soit D une matrice $n \times n$ vérifiant l'inégalité quadrangulaire. Alors $\sqrt{D}$ est euclidienne de dimension n-1 exactement.

Si D vérifie l'inégalité quadrangulaire, on peut lui associer un arbre additif A_N à N points, où $N = n+k$, n étant le nombre de points de l'étude et k, compris entre 0 et n-2, le nombre de noeuds intermédiaires qu'il est nécessaire de rajouter pour avoir une structure d'arbre.
Montrons par récurrence sur N que ces N points sont représentables dans un espace euclidien de dimension N-1, engendrant la distance $\sqrt{D}$ et que ceci est impossible dans un espace de dimension inférieure.
Pour $n = 4$ l'arbre additif le plus général peut se représenter par la figure ci-dessous ; A, B, C, D étant les points de l'étude et L et P les noeuds intermédaires.



Il est aisé de vérifier que le tableau de coordonnées répond à la question et est clairement de rang 5.

$$X = \begin{array}{c|ccccc} A & \sqrt{a} & 0 & 0 & 0 & \sqrt{e} \\ B & 0 & \sqrt{b} & 0 & 0 & \sqrt{e} \\ C & 0 & 0 & \sqrt{c} & 0 & 0 \\ D & 0 & 0 & 0 & \sqrt{d} & 0 \\ L & 0 & 0 & 0 & 0 & \sqrt{e} \\ P & 0 & 0 & 0 & 0 & 0 \end{array}$$

L'hypothèse de récurrence est donc vérifiée pour $N \leq 6$.

Supposons alors la propriété vraie jusqu'à l'ordre N et soit A_{N+1} un arbre additif à N+1 points.

Soit A un noeud terminal à la distance a de son noeud intermédiaire associé L. L'arbre A_N, obtenu par suppression de A est représentable dans un espace euclidien de dimension N - 1 exactement soit E^{N-1}. Il suffit alors de placer A dans une direction orthogonale à E^{N-1} en L, à une distance $\sqrt{a}$ pour obtenir une figure euclidienne de dimension N représentant les N+1 points.
Le fait qu'il soit impossible de réduire cette dimension résulte de ce que, l'euclidiennité étant acquise, la structure additive (en terme de D) de l'arbre entraîne que l'arête AL est orthogonale (en terme de $\sqrt{d}$) à E^{N-1}, de rang plein d'après l'hypothèse de récurrence.
La sous figure des n points de départ est donc euclidienne de dimension inférieure ou égale à n - 1. Si cela était représentable dans un espace de dimension p, $p < n - 1$, les n + k points seraient représentables dans un espace de dimension p + k - 1 ou plus, avec $p + k-1 < n + k-1$, ce qui contredit le résultat précédent et achève la démonstration.

Proposition 5

- C, distance à centre, n'est pas, en général, euclidienne.
- Elle ne l'est jamais si le centre appartient à l'étude et si $n > 3$.

- $W_0(C^2) = \begin{cases} -XX^t & \text{si c'est un éloignement à centre} \\ -XX^t + 2\,\mathrm{Diag}(XX^t) & \text{pour une distance à centre.} \end{cases}$

Soit C une distance à centre, $n \geq 4$ et le centre appartenant à l'étude. Supposons que C admette une figure euclidienne associée. Soient A, B, C trois points quelconques et O le centre. Le fait que $C_{ij} = C_{oi} + C_{oj}$ entraîne que O est aligné avec i et j et à l'intérieur du segment ij. Donc O devrait être à l'intérieur des segments AB, BC, AC ce qui est impossible ; et donc $W_M(C^2)$ n'est jamais sdp, $\forall$ M, si O appartient à l'étude et $n > 3$.
Si O n'appartient pas à l'étude un raisonnement par continuité montre qu'il existe des vecteurs X strictement positifs tels que $W_M(C^2)$ ne soit pas sdp. L'ajoût d'une constante à chaque coordonnée de X transformant C_{ij} en $C_{ij} + 2a$, il est clair que l'on peut choisir cette constante pour que $W_M\,[\,(C + 2a)\,]^2$ soit sdp (problème de la constante additive). D'où le fait qu'il existe des études pour lesquelles C est euclidienne et d'autres pour lesquelles elle ne l'est pas.

Enfin, un calcul simple montre que :

$$WO\,(C^2)\,ij = -\,X_i\,X_j \text{ pour } i \neq j$$

$$WO\,(C^2)\,ii = \begin{cases} -\,X^2 i & \text{si } C \text{ éloignement à centre} \\ +\,X^2 i & \text{si } C \text{ distance à centre} \end{cases}$$

On notera cette opposition de signe entre les termes de la diagonale et les autres pour une distance à centre.

II. OPTIMISATIONS. DECOMPOSITIONS.

Problème 1

Etant données une dissimilarité D et une distance à centre C, trouver k Min tel que $D^2 + kC$ soit un carré de distance euclidienne.

Ce problème peut être considéré comme une généralisation de la constante additive sur D^2 puisque celui-ci est le problème 1 avec

$$X = \begin{bmatrix} 1 \\ 1 \\ . \\ . \\ 1 \end{bmatrix}$$

Fichet, qui a d'ailleurs posé le premier le problème dans un autre contexte, en donne la solution (cf [3]) à savoir : la valeur absolue de la plus petite valeur propre négative de la matrice de l'opérateur factoriel du nuage affecté des masses $1/X_i$.

Problème 2

D dissimilarité donnée, trouver C, distance à centre, telle que $|| D - C ||^2$ soit minimum.

Posons $F = || D - C ||^2$

$$F = \sum_{i \neq j} [d_{ij} - (X_i + X_j)]^2 = \sum_{i,j} d^2_{ij} + 2(n-2) \sum_i X^2_i - 2 \sum_{ij} d_{ij} (X_i + X_j) + 2(\sum_i X_i)^2$$

Donc $\partial F / \partial X_i = 4(n-2) X_i + 4 \sum X_j - 4 \sum d_{ij}$

de $\sum_i \partial F / \partial X_i = 0$ on tire $\sum_j X_j = \sum_{ij} d_{ij} / 2n-2$, qui, reportée dans $\partial F / \partial X_i = 0$ donne

$$X_i = 1 / (n-2) \; [\sum_j d_{ij} - \sum_{ij} d_{ij} / 2n - 2]$$

Il est à noter que l'on n'est pas assuré de la positivité de la solution.

Problème 3

D étant une dissimilarité donnée, trouver une décomposition de la forme $D^2 = E^2 + C$, où E est une distance euclidienne et C une distance à centre de centre donné O.

En prenant les formes de Torgerson au point O, il vient $W_0(D^2) = W_0(E^2) + W_0(C)$, soit le problème suivant : décomposer une matrice carrée symétrique en une somme d'une matrice sdp et d'une matrice diagonale. On reconnaît là le problème des factorialistes au sens de Thurstone. La différence des points de vue vient de ce que, dans notre optique, la matrice diagonale n'est pas un résidu, terme d'erreur par rapport à un modèle probabiliste "à priori", mais une composante aussi importante que l'autre et simple à interpréter.

Remarque . Cas où l'on désire "centrer" au point G.

Il est fréquent en Analyse des données de prendre la forme de Torgerson au "centre de gravité". Elle est donnée par $W_G(D^2) = 1/2 \; [d^2_{i.} + d^2_{.j} - d^2_{ij} - d^2_{..}]$, ce qui revient à rajouter à l'étude un point G défini par $d^2_{Gi} = d^2_{i.} - 1/2 \; d^2_{..}$.
Une première remarque, qui nous semble méconnue, est que cette quantité n'est

pas, en général, positive pour tout i si D n'est pas euclidienne, ce qui est quand même gênant pour des carrés de distance.
Dans notre problème particulier $d^2_{Gi} = E^2_{Gi} + C_{Gi}$ entraîne que $C_{Gi} = C_{i.} - 1/2\ C_{..}$.

Un calcul simple donne

$$W_G(C)_{ij} = 1/2\ n\ [\ X. - X_i - X_j\ ] \qquad i \neq j$$
$$W_G(C)_{ii} = 1/2\ n\ [\ 2\ (n-1)\ X_i + X.\ ]$$

matrice qui n'est pas diagonale et l'on n'est donc pas dans le problème de Thurstone. Ceci provient du fait que, pour une distance à centre, le point qui joue le rôle de centre de gravité est le centre et non le point G implicitement défini dans la formule classique. Il en découle que le choix de ce point comme origine n'est pas, en général, un bon choix si l'on soupçonne une composante à centre.

Problème 4

D étant une dissimilarité donnée, trouver C tel que $D^2 + C = E^2$.

Il s'agit d'une généralisation du problème 1 et par là même de celui de la constante additive sur D^2.
Il faut préciser les types de contrainte à imposer à C.

1er type de contrainte : || C || minimum
Le passage aux formes de Torgerson débouche sur un problème de régression linéaire sous contrainte de positivité pour lequel nous n'avons pas plus, mais pas moins, de résultats que ceux de la littérature connue sur le sujet.

2ème type de contrainte : Var (X) minimum, X. étant minimum
La solution en est connue : c'est le problème de la constante additive sur D^2 (Var X = 0) qui consiste à "culpabiliser" uniformément les données D_{ij}.

3ème type de contrainte : Var (X) maximum, X. minimum
A l'opposé du type précédent, cette contrainte tend au contraire à laisser le maximum de données inchangées et à en modifier quelques unes. C'est une voie qui nous semble peu explorée en Analyse des données et qui peut être une solution quand on a affaire à un problème où on soupçonne de fortes perturbations locales (et non globales comme dans le type 2) par rapport à un modèle euclidien.

Sur le plan théorique le problème est sans solution connue à ce jour. Sur le plan pratique, des techniques algorithmiques simples consistant à annuler successivement les valeurs propres négatives de la forme de Torgerson par construction d'un vecteur X porté par les points ayant une contribution importante à ces valeurs négatives, se sont montrées efficaces dans certains problèmes concrets.

Problème 5 : Analyse factorielle en composantes à centre

Trouver un système de distances à centre C^k tel que D étant une dissimilarité donnée, $D^2 = \sum_k \lambda_k C^{k^2}$.
Si l'on remarque que l'analyse factorielle d'un tableau de distances est en fait le problème : trouver un système δ_k de distances euclidiennes "élémentaires" tel que $D^2 = \sum \lambda_k \delta_k 2$, on justifie du même coup le titre du problème et son aspect "naturel" par analogie.

Proposition 6

Soit $U_M(D^2)$ la matrice définie par

$$U_M(D^2)_{ij} = - W_M(D^2)_{ij} = - 1/2 (D^2_{Mi} + D^2_{Mj} - D^2_{ij}) \qquad i \neq j$$

$$U_M(D^2)_{ii} = W_M(D^2)_{ii} = D^2_{Mi}$$

Alors $D^2_{ij} = \sum \lambda_k (X^k_i + X^k_j)^2$ $i \neq j$, λ_k et X^k étant les valeurs propres et les vecteurs propres de la matrice $U_M(D^2)$.
De la définition de $U_M(D^2)$ il découle immédiatement que

$$D^2_{ij} = U_M(D^2)_{ii} + 2\, U_M(D^2)_{ij} + U_M(D^2)_{ii}$$

d'où le résultat, en écrivant que

$$U_M(D^2)_{ij} = \sum \lambda_k X^k_i X^k_j, \qquad U_M(D^2)_{ii} = \sum \lambda_k X^{k^2}_i.$$

Cette forme $U_M(D^2)$ peut, a priori, paraître curieuse. Elle est en fait tout à fait naturelle si l'on se remémore la proposition 5 qui note une opposition de signe entre les termes de la diagonale et les autres dans la matrice $W_0(C^2)$ qui rend l'opérateur associé discontinu et par là même peu adapté à l'étude de C^2.

III. LIENS AVEC LES AUTRES TYPES DE DISTANCE

Un certain nombre de types de distance sont couramment utilisées en analyse de données, et puisque nous venons d'en introduire une nouvelle il nous paraît utile de voir comment elle se place par rapport aux autres.
Précisons d'abord nos notations. Nous travaillons sur l'ensemble des matrices carrées $n \times n$ et nous définissons :

D, l'ensemble des matrices de dissimilarité c. à. d l'ensemble des matrices à termes positifs ou nuls, symétriques et à diagonale nulle.

T, l'ensemble des matrices de distance : le sous ensemble de **D** vérifiant l'inégalité triangulaire.

C, l'ensemble des distances à centre.

E, le sous ensemble de **T** admettant une représentation euclidienne.

$\mathbf{E^{1/2}}$, Le sous ensemble de **T** constitué des matrices dont la racine carrée (terme à terme) admet une représentation euclidienne.

U, l'ensemble des matrices ultramétriques.

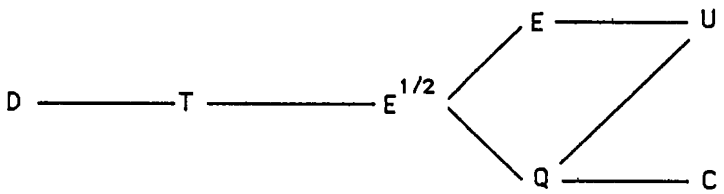
Q, l'ensemble des matrices vérifiant l'inégalité quadrangulaire.

On sait que ces différents sous ensembles sont (partiellement) ordonnés par les inclusions suivantes :

$$\mathbf{D} \supset \mathbf{T} \supset \mathbf{E} \supset \mathbf{U}$$
$$\mathbf{D} \supset \mathbf{T} \supset \mathbf{E^{1/2}} \supset \mathbf{Q} \supset \mathbf{U}$$
$$\mathbf{Q} \supset \mathbf{C}$$

Pour l'inclusion $\mathbf{E^{1/2}} \supset \mathbf{E}$ qui semble peu connue voir [6] ou [4].

Ces inclusions peuvent se schématiser de la façon suivante :



Proposition 7 : On a les résultats suivants :

a) $\forall D \in \mathbf{D} \quad \exists C \in \mathbf{C}$ tel que $D + C \in \mathbf{T}$

b) $\forall T \in \mathbf{T} \quad \exists C \in \mathbf{C}$ tel que $T + C \in \mathbf{E}$

c) $\forall T \in \mathbf{T} \quad \exists C \in \mathbf{C}$ tel que $T + C \in \mathbf{E}^{1/2}$

d) $\forall Q \in \mathbf{Q} \quad \exists C \in \mathbf{C}$ tel que $Q + C \in \mathbf{U}$

Démonstration
Pour le a) on pose $X_i = \max [\max (d_{jk} - d_{ij} - d_{ki}) ; 0]$ on en trouvera la démonstration détaillée dans [Patrinos et Hakini].

Pour le b) une solution (parmi d'autres) est donnée par le problème de la constante additive [Caillez] cf [1].

Pour le c) soit M un point de l'étude, C une distance à centre de centre M. $W_M(C)$ étant diagonale, il est toujours possible de trouver C tel que $W_M(C) + W_M(T) = W_M(T+C)$ soit sdp.

Pour le d) on sait, cf [2], que parmi l'infinité de décompositions possibles de la forme $Q = U + C$ il en existe telles que U soit positive ou nulle, C en général ne l'étant pas. Posons $C' = C$. Si c'est positif la démonstration est terminée. Si ce n'est pas le cas, posons

$$K = - \inf_{ij} C_{ij} ,$$
$$\text{pour } i \neq j$$
$$C'' = C' + K , \quad C''_{ii} = C'_{ii}$$
$$U' = U + K , \quad U'_{ii} = U_{ii}$$

Alors C" est une distance à centre positive, U' est une ultramétrique positive et l'on a Q + C" = U'

Remarque : Distances à centres
Il est très simple d'étendre la notion de distance à centre au cas où il existe plusieurs centres. Si à chaque point i on fait correspondre son centre i', on peut alors écrire

$$d^{ij} = d^{ii'} + d^{i'j'} + d^{j'j}$$

Notons que le fait d'avoir le même centre ($d^{i'j'} = 0$) est une relation d'équivalence et que cette décomposition est toujours possible (ne serait-ce qu'en prenant i = 1' ∀ i)
On vérifie facilement qu'une quadrangulaire est alors une distance à centres obtenue par composition de distances à centre. Cette décomposition de type structurelle (liée à la structure de l'arbre additif) offre un point de vue différent de la décomposition classique Q = U + C qui, elle, est de type métrique.

CONCLUSIONS

Rencontrée déjà par différents auteurs, la notion de distance à centre a été, jusqu'a ce jour, considérée comme une curiosité, qu'il s'agisse de la distance ferroviaire (Benzecri) ou qu'on la considère comme un arbre additif dégénéré.
Nous pensons avoir montré qu'elle mérite un autre statut en raison, entre autres, de son adaptation à certains problèmes concrets, de ses propriétés mathématiques, de son rôle charnière entre les divers types de distance. Ces deux dernières raisons nous amènent à penser qu'elle peut jouer un rôle important dans les théorèmes de décomposition de distances, comme c'est déjà le cas pour les arbres additifs.

BIBLIOGRAPHIE

[1] CAILLEZ — The analytical solution of the additive constant problem. Psychometrika, Juin 1983, Vol. 48, n° 2.

[2] CAROLL et al. — "Spatial non spatial and Hybrid models for scaling". Psychometrika, 1976, 41-4, 439-469.

[3] FICHET — Analyse factorielle sur tableaux de dissimilarité. Thèse d'état en Biologie Humaine, Université d'AIX-MARSEILLE II, Faculté de Médecine, Avril 1983.

[4] JOLY-LE CALVE — "Etude des puissances d'une distance", Laboratoire d'analyse des données. Université de RENNES 2. Rapport technique 1-85.

[5] PATRINOS ET HAKINI — The distance matrice and its tree realization. Quaterly of applied Mathematics. Octobre 1972, Vol. 30, n°3.

[6] SCHOENBERG — "On certain metric spaces by a change of metric and their imbedding in Hilbert space". Annals of mathematics. Octobre 1937, Vol. 38, n° 4.

[7] ZARETSKII — "Constructing a tree on the basis of a set of distances between the hanging vertices". (Upekhi Math. Nant. 20, p. 90-92) en russe.