Statistique et analyse des données

B. Fichet
G. Le Calve

Structure géométrique des principaux indices de dissimilarité sur signes de présence-absence

Statistique et analyse des données, tome 9, n° 3 (1984), p. 11-44
<http://www.numdam.org/item?id=SAD_1984__9_3_11_0>

Statistiques et Analyse de données
1984 - Vol. 9 n° 3 pp 11-44

STRUCTURE GEOMETRIQUE DES PRINCIPAUX INDICES DE DISSIMILARITE SUR SIGNES DE PRESENCE-ABSENCE

B. FICHET*
G. LE CALVE**

* Faculté de Médecine. Université d'Aix-Marseille II
**UER des Sciences et Techniques. Université de Haute-Bretagne.

Résumé : *Considérant les principaux indices de similarité s de signes de présence-abscence et les dissimilarités d'associées par la relation d=1-s, les auteurs étudient la nature géométrique des indices de dissimilarité d et $\sqrt{d}$, du point de vue de leurs structures métrique et euclidienne.*

Abstract : *Considering the usual similarity coefficients s for logical data (presence-absence of characters) and the associated dissimilarities d such that d=1-s, the autors study the geometrical nature of dissimilarity coefficients d and $\sqrt{d}$, with recpect to theirs metric and euclidean structures.*

Mots clés : *similarité, dissimilarité, distance, distance euclidienne, matrice semi-définie positive, approximation.*

Indices de classification STMA : *00-020, 00-050, 00-060*

1. INTRODUCTION

Pour tout individu d'une population donnée, on suppose relevés n signes dits de présence-absence; chaque signe offre, soit la réponse 1 (présence ou détection), soit la réponse 0 (absence ou non-détection).

Pour approcher de telles données, de nombreux indices de dissimilarité ou de similarité ont été proposés dans la littérature. Ceux auxquels nous ferons référence par la suite, sont extraits des ouvrages de Jambu et Lebeaux -[12]p.97-, Lerman -[13]p.18-, Cailliez et Pagès -[3]p.499-.

Dans [5], il est montré que deux indices, qui ne diffèrent que par un simple coefficient, sont de natures géométriques fort distinctes: l'un, l'indice de Sokal et Sneath est une distance (bien que non-euclidienne), l'autre, l'indice de Czekanowski-Dice n'est pas une distance. Dans cette note, les auteurs généralisent ce résultat, tout en proposant des démonstrations beaucoup plus simples. Pour les indices de dissimilarité d, liés, à l'exception d'un indice de Kulczinski, aux principaux indices de similarité s par la relation d=1-s, ils dressent un bilan, exhibant ceux d'entre eux qui satisfont à l'inégalité triangulaire. En particulier, ils insèrent trois indices classiques, les indices de Jaccard, de Sokal et Sneath et de Czekanowski-Dice, dans une même famille $\{d_\theta, \theta \in \mathbb{R}_+^*\}$, et donnent une condition nécessaire et suffisante sur θ pour que d_θ soit une distance. En outre, il est déduit d'un contre-exemple général qu'aucun des indices considérés n'est euclidien.

La nature euclidienne des racines carrées d'indices est abordée à travers la semi-définie positivité de la matrice de similarité s: $\sqrt{d}$ est euclidien dès que s est semi-définie positive (s.d.p.). Pour les aspects métrique et euclidien, un bilan est encore dressé, conduisant pour les indices $\sqrt{d_\theta}$ de la famille précédente, à deux conditions nécessaires et suffisantes.

Nous devons mentionner que nombre de résultats concernant la structure métrique des indices d et la semi-définie positivité des matrices de similarité, ont été établis parallèlement avec Gower, résultats que l'auteur annonce dans [6] et cite dans [7] pour une publication à paraître.

Enfin, pour les indices d_θ et $\sqrt{d_\theta}$ précédents, les auteurs étudient différentes approximations, évaluant en particulier l'impact des techniques de constante additive.

2. NOTATIONS

Pour un entier n(n>1) fixé, I désigne l'ensemble {1,..,n} et X l'ensemble $\{0,1\}^n$. L'élément de X (0,..,0) est noté $\underline{0}$ et l'ensemble X-{0} est noté X^+. L'élément générique $x=(x_1,..,x_n)$ de X s'interprète comme l'observation de n signes de présence-absence; pour i variant de 1 à n, x_i signifie présence (si $x_i=1$) ou absence (si $x_i=0$) du $i$ème signe.
Pour tout $x=(x_1,..,x_n)$ de X, on définit A_x et n_x par :

$$A_x=\{i\in I \mid x_i=1\} \; ; \; n_x=\sum_i x_i=|A_x|$$

A_x désigne l'ensemble des signes présents pour l'observation x, et n_x leur nombre.
Pour tout $x=(x_1,..,x_n)$ et tout $y=(y_1,..,y_n)$ de X, on définit encore n_{xy}, $n_{\bar{x}\bar{y}}$, q_{xy} par:

$$n_{xy}=\sum_i x_i y_i \; ; \; n_{\bar{x}\bar{y}}=\sum_i (1-x_i)(1-y_i) \; ; \; q_{xy}=\sum_i |x_i-y_i|$$

n_{xy}, $n_{\bar{x}\bar{y}}$, q_{xy} représentent respectivement, le nombre de concordances positives, le nombre de concordances négatives et le nombre de discordances entre les deux observations x et y.
On a clairement les relations: $n=n_{xy}+n_{\bar{x}\bar{y}}+q_{xy}$; $n_x+n_y=2n_{xy}+q_{xy}$.

Introduisons maintenant les indices qui seront analysés par la suite. Ils sont définis sur X ou sur un sous-ensemble de X. Dans les relations ci-dessous, (x,y) désigne l'élément générique de X^2.
Pour une généralisation, nous intégrons trois indices classiques dans une famille $\{d_\theta, \theta\in\mathbb{R}_+^*\}$, définie par :

$$(1)\; d_\theta(x,y)=\begin{cases} q_{xy}/(\theta n_{xy}+q_{xy}) & \text{si } (x,y)\neq(\underline{0},\underline{0}) \\ 0 & \text{sinon} \end{cases}$$

Pour tout θ, d_θ est défini sur X. On retrouve l'indice de Jaccard pour $\theta=1$, l'indice de Czekanowski-Dice pour $\theta=2$ et l'indice de Sokal et Sneath-Anderberg pour $\theta=1/2$.

Dans les relations suivantes, l'indice est toujours noté abusivement d , et peut vérifier:

$$(2)\; d(x,y)=2q_{xy}/(n+q_{xy})$$

Indice de Rogers et Tanimoto défini sur X.

(3) $d(x,y) = \begin{cases} (n_{\bar{x}\bar{y}}+q_{xy})/n & \text{si } x \neq y \\ 0 & \text{sinon} \end{cases}$

Indice de Russel et Rao défini sur X.

(4) $d(x,y) = 1-n_{xy}/\sqrt{n_x n_y}$

Indice d'Ochiaï défini sur X^+.

(5) $d(x,y) = 1-(n_{xy}/2)(1/n_x+1/n_y)$

Indice de Kulczinski défini sur X^+.

(6) $d(x,y) = q_{xy}/n_{xy}$

Indice de Kulczinski défini sur tout domaine $\hat{X} \subset X$ tel que $\forall(x,y)\in\hat{X}^2$, $n_{xy}\neq 0$

(7) $d(x,y)=q_{xy}/(2n-q_{xy})$

Indice de Sokal et Sneath défini sur X.

Remarque 1

Dans la liste ci-dessus, nous n'avons pas fait mention des indices de Hamming ou Hamming pondéré (M.G. Kendall, Sokal et Michener), ainsi que des indices euclidien ou euclidien pondéré. Par la symétrie qu'ils font jouer aux réponses positives et négatives, il est délicat de les considérer comme indices sur signes de présence-absence. Et, en tout état de cause, il est bien connu que l'indice de Hamming (vérifiant $d(x,y)=q_{xy}$) est une distance non-euclidienne (distance L^1), et que l'indice euclidien (vérifiant $d(x,y)=\sqrt{q_{xy}}$) est une distance euclidienne (distance L^2).

Remarque 2

Tous les indices cités sont des dissimilarités propres (on rappelle qu'une dissimilarité d est dite propre ssi est vraie l'équivalence (*):$[d(x,y)=0]\Longleftrightarrow[x,y]$).

La positivité et la symétrie sont en effet évidentes. Pour les indices (1), (2), (3), (6), (7), l'équivalence (*) découle de l'équivalence suivante vraie sur X:$[x=y]\Longleftrightarrow[q_{xy}=0]$; et pour l'indice (4) écrit sous la forme $d(x,y)=1-\sqrt{\frac{n_{xy}}{n_x}.\frac{n_{xy}}{n_y}}$, ou l'indice (5) écrit sous la forme $d(x,y)=1-\frac{1}{2}\left(\frac{n_{xy}}{n_x}+\frac{n_{xy}}{n_y}\right)$, (*) découle des deux équivalences suivantes vraies sur X^+:

$[x,y] \Longleftrightarrow [n_x=n_y=n_{xy}]$ et

$[n_x=n_y=n_{xy}] \Longleftrightarrow [\frac{n_{xy}}{n_x} \cdot \frac{n_{xy}}{n_y} = 1]$ ou $[n_x=n_y=n_{xy}] \Longleftrightarrow [\frac{n_{xy}}{n_x} + \frac{n_{xy}}{n_y} = 2]$

Remarque 3

En pratique, on dispose d'une population J et d'une d'une famille $\{x^j, j\in J\}$, x^j de X désignant l'observation des n signes pour l'individu j de J. Notons $f:J\mapsto X$, l'application vérifiant: $\forall j\in J$, $f(j)=x^j$.
Construire un indice sur J à partir des observations $\{x^j, j\in J\}$, revient en fait à choisir un indice d sur X (ou sur un sous-ensemble de X), et à définir l'indice δ sur J par: $\forall (j,k)\in J^2$, $\delta(j,k)=d(f(j),f(k))$.
Pour cela, il est d'ailleurs nécessaire que f(J) soit inclus dans le domaine de définition de d.
Si d est une dissimilarité, il en est de même de δ. Mais si d est propre, δ n'est propre que ssi f est injective. Notons que δ est toutefois semi-propre, *i*.e. vérifie: $[\delta(j,k)=0] \Longleftrightarrow [\forall \ell\in J, \delta(j,\ell)=\delta(k,\ell)]$.
Si d est une distance (euclidienne), δ est une semi-distance (euclidienne). Mais δ peut être une semi-distance (euclidienne) sans que d le soit: il faut et il suffit que la restriction de d à f(J) soit une distance (euclidienne).

Remarque 4

Il est clair que l'application de X dans P(I) qui à x fait correspondre A_x, est une bijection. En conséquence, à tout indice d défini sur X (ou un domaine de X), correspond un indice δ défini sur P(I) (ou un domaine de P(I)), de même nature géométrique que d.
Si $d(x,y)$ s'exprime en fonction de $n_x, n_y, n_{xy}, n_{\bar{x}\bar{y}}, q_{xy}$, $\delta(A_x,A_y)$ s'exprime en fonction des cardinaux des ensembles respectifs $A_x, A_y, A_x\cap A_y$, $(A_x\cup A_y)^c$, $A_x\Delta A_y$. A titre d'exemples, à l'indice de Hamming correspond l'indice du cardinal de la différence symétrique, et à un indice d_θ défini en (1), correspond un indice δ_θ sur P(I) vérifiant :
$\forall A\subset I$, $\forall B\subset I$, $\delta_\theta(A,B)=|A\Delta B|/[\theta|A\cap B|+|A\Delta B|]$.

3. DISTANCES EUCLIDIENNES ET MATRICES DE SIMILARITE S.D.P.

Pour tout ensemble fini non vide J, on note $\mathcal{D}_J$ l'espace vectoriel des applications $d: J\times J \mapsto \mathbb{R}$, vérifiant: $\forall (i,j)\in J^2, d(i,j)=d(j,i)$; $\forall j\in J,\ d(j,j)=0$.
$\mathcal{D}_J$ est de dimension $|J|(|J|-1)/2$, et une dissimilarité sur J est un élément de l'orthant $\mathcal{D}_J^+$
Notons encore $\mathcal{D}_J^e$ l'ensemble des semi-distances euclidiennes sur J (on rappelle qu'une semi-distance d sur J est euclidienne ssi il existe une image euclidienne de (J,d), *i.e.* une famille de points, soit $\{M_j, j\in J\}$ d'un espace affine euclidien, vérifiant: $\forall (i,j)\in J^2, \|\vec{M_iM_j}\| = d(i,j)$).

On fait souvent usage de la figure suivante pour caractériser les semi-distances euclidiennes sur J. Soit F un espace affine de dimension $|J|$, muni d'un référentiel $\{O,\{\vec{e_j}, j\in J\}\}$; pour tout j de J, on définit le point N_j par $N_j=O+\vec{e_j}$, et on note H l'hyperplan passant par les points $N_j, j\in J$.
Il est alors bien connu que pour toute dissimilarité d sur J, il existe une infinité de formes bilinéaires symétriques q sur $\vec{F}$, vérifiant: $\forall (i,j)\in J^2,\ q(N_i\vec{N}_j, N_i\vec{N}_j)=d^2(i,j)$.
Toutes ces formes ont même restriction $\hat{q}$ à $\vec{H}$. Plus précisément, considérons une base de $\vec{H}$ définie par le simplexe des N_j, $j\in J$, soit $\{N_iN_j, j\in \bar{J}\}$, où *i* de J est fixé et où $\bar{J}=J-\{i\}$; alors $\hat{q}$ vérifie:
$\forall (j,k)\in \bar{J}^2$, $\hat{q}(N_i\vec{N}_j, N_i\vec{N}_k)= \frac{1}{2}\,[d^2(i,j)+d^2(i,k)-d^2(j,k)]$.

Il est classique (voir, par exemple, Einhorn and Schoenberg [4]),que d est une semi-distance euclidienne ssi $\hat{q}$ est s.d.p. En outre, si d est euclidienne, sa dimension est égale au rang de $\hat{q}$ (on rappelle que la dimension d'une semi-distance euclidienne est, par définition, la dimension de l'espace affine engendré par les points de l'une quelconque de ses images euclidiennes).

Cette correspondance entre semi-distances euclidiennes (resp.distances euclidiennes de dimension maximum) et formes s.d.p. (resp. définie-positives) sur $\vec{H}$, permet d'établir -[5],p.50- que lorsque $\mathcal{D}_J$ est muni d'une norme quelconque, $\mathcal{D}_J^e$ forme un cône fermé, ayant pour intérieur,

soit $\overset{o}{\mathcal{D}}{}^{e}_{j}$, l'ensemble des distances euclidiennes de dimension maximum. Parmi toutes les formes q satisfaisant à la propriété mentionnée ci-dessus, la plus classique est la forme de Torgerson, notée w, et vérifiant -[3]- :

$\forall(i,j)\in J^2$, $w(\vec{e}_i,\vec{e}_j)=(1/2)[-d^2(i,j)+d^2(i,.)+d^2(j,.)-d^2(.,.)]$,

où $\forall i\in J$, $d^2(i,.)=(1/|J|)\sum_j d^2(i,j)$; $d^2(.,.)=(1/|J|)\sum_j d^2(j,.)$.

Mais pour tout réel K, on constate aisément que satisfait encore à la propriété requise, la forme f_K définie par:

$\forall(i,j)\in J^2$, $f_K(\vec{e}_i,\vec{e}_j)=K-(1/2)d^2(i,j)$.

En conséquence, d est euclidienne ssi la restriction de f_K à $\vec{H}$ est s.d.p.; en d'autres termes, considérant une bijection entre J et $\{1,..,|J|\}$ et notant S_K la matrice de terme général $S_K(i,j)=K-(1/2)d^2(i,j)$, d est euclidienne ssi: $\forall Y\in R^{|J|}$ tel que $Y'\mathbf{1}_{|J|}=0$, $Y'S_KY\geq 0$.

Supposons maintenant qu'au lieu d'une dissimilarité, soit donnée une similarité s sur J, telle que: $\forall j\in J$, $s(j,j)=K<+\infty$.

Deux dissimilarités classiques peuvent être associées à s. L'une, notée d, est définie par: $d=K-s$; l'autre notée δ, est définie par $(1/2)\delta^2=K-s$. Naturellement, δ est de même nature géométrique que $\sqrt{d}$. Pour la dissilarité δ, la forme f_K précédente vérifie: $\forall(i,j)\in J^2$, $f_K(\vec{e}_i,\vec{e}_j)=s(i,j)$, ce qui constitue une justification a posteriori du coefficient 1/2 dans la définition de δ. En conséquence, pour une bijection entre J et $\{1,..,|J|\}$, δ est euclidienne ssi la matrice, soit S, associée à s, vérifie:

$\forall Y\in R^{|J|}$ tel que $Y'\mathbf{1}_{|J|}=0$, $Y'SY\geq 0$

Le cas particulier où S est s.d.p. est caractérisé par la proposition suivante.

Proposition 1

Soient s *et* δ *respectivement une similarité et une dissimilarité sur un ensemble fini non vide J, liées par la relation :* $(1/2)\delta^2=K-s$, *où* $\forall j\in J$, $K=s(j,j)<+\infty$.

Pour que la matrice de similarité soit s.d.p. , il faut et il suffit que soient satisfaites les trois conditions suivantes:

i) il existe une image euclidienne de (J,δ), *notée* $\{M_j,j\in J\}$

ii) Les points M_j, $j\in J$ *ont une sphère circonscrite, notée* $\Sigma(0,r)$.
iii) $r\leqslant\sqrt{K}$

Preuve:
Pour la condition nécessaire, complètons l'ensemble J par un élément ω, et prolongeons δ à $\bar{J}=J\cup\{\omega\}$ en posant: $\forall j\in J$, $\delta^2(\omega,j)=K$.
Soit E un espace affine de dimension $|J|+1$ muni d'un référentiel $\{0',\{\vec{e}_j,j\in J\},\vec{e}_\omega\}\}$,
et définissons la forme bilinéaire symétrique g sur $\vec{E}$ par:
$\forall(i,j)\in J\times J$, $g(\vec{e}_i,\vec{e}_j)=s(i,j)$; $\forall j\in J$, $g(\vec{e}_j,\vec{e}_\omega)=g(\vec{e}_\omega,\vec{e}_j)=0$; $g(\vec{e}_\omega,\vec{e}_\omega)=0$.
Définissant les points N_j, $j\in J$ et N_ω par: $\forall j\in J$, $N_j=0'+\vec{e}_j$, $N_\omega=0'+\vec{e}_\omega$,
l'on a :
$\forall(i,j)\in J^2$, $g(\vec{N_iN_j},\vec{N_iN_j})=\delta^2(i,j)$; $\forall j\in J$, $g(\vec{N_jN_\omega},\vec{N_jN_\omega})=K$.
Considérons une bijection entre $\bar{J}$ et $\{1,..,|J|+1\}$ faisant correspondre $|J|+1$ à ω et notons S la matrice de similarité. Dans la base $\{\{\vec{e}_j,j\in J\},\vec{e}_\omega\}$, g est caractérisée par la matrice: $\begin{bmatrix} S & \begin{matrix}0\\ \vdots\end{matrix} \\ 0\ \ldots\ 0 & \end{bmatrix}$
Cette matrice (qui n'est pas de similarité) est s.d.p., de sorte qu'il existe une image euclidienne, soit $\{\{M_j,j\in J\},M_\omega\}$ de $(\bar{J},\delta)$. Dans cet espace euclidien, les points $M_j,j\in J$ sont sur la sphère $\Sigma(M_\omega,\sqrt{K})$. D'où les conditions i),ii),iii), $\Sigma(0,r)$ étant l'intersection de $\Sigma(M_\omega,\sqrt{K})$ et du sous-espace affine engendré par les points M_j, $j\in J$.
Réciproquement, supposons donnée l'image $\{M_j,j\in J\}$ de (J,δ) sur la sphère $\Sigma(0,r)$. Complétons J par un élément ω' et prolongeons δ à $\hat{J}=J\cup\{\omega'\}$ en posant:
$\forall j\in J$, $\delta(\omega',j)=r$.
Il est clair que $\{\{M_j,j\in J\},0\}$ est une image euclidienne de $(\hat{J},\delta)$. En conséquence, la matrice de terme général $(1/2)[\delta^2(\omega,i)+\delta^2(\omega,j)-\delta^2(i,j)]$, $i,j=1,..,|J|$, est s.d.p. Mais ce terme est égal à $r^2-K+s(i,j)$, de sorte que S est s.d.p. dès que r vérifie : $r\leqslant\sqrt{K}$.

□

Remarque 5
S'il existe un couple (i,j) tel que $s(i,j)=0$, *i.e.* tel que $\max_{i,j}\delta(i,j)=\sqrt{2K}$, alors iii) est équivalent à:

iii)' il existe deux points M_i et M_j de l'image ayant un angle au centre supérieur ou égal à $\pi/2$. L'existence d'un couple (i,j) tel que $s(i,j)=0$ est souvent assurée. Elle l'est toujours si s est construite à partir de δ en posant $K=(1/2)\max_{i,j}\delta^2(i,j)$. A l'exception de l'indice de Kulczinski défini en (6), toutes les dissimilarités s sur signes de présence-absence données au paragraphe précédent, sont associées à une similarité classique s par la relation d=1-s ; et si x et y sont tels que $q_{xy}=n$, on constate que l'on a encore $s(x,y)=0$.

Pour d'autres propriétés caractéristiques, le lecteur pourra consulter Gower -[8],[9]-.

4. STRUCTURE DES DISSIMILARITES SUR SIGNES DE PRESENCE-ABSENCE

Les résultats porteront essentiellement sur les indices (1), (2), (3). Les contre-exemples suivants montrent en effet que pour tout n, les indices (4), (5), (6), (7) ne sont pas des distances.

Contre-exemples 1

Soient x,y,z de X^+ vérifiant:

x=(1 0 0..0) , y=(0 1 0..0) , z=(1 1 0..0)

Pour l'indice d'Ochiaï (4) on a :

$d(x,y)=1$; $d(x,z)=d(y,z)=1-\sqrt{2}/2$, de sorte que $d(x,z)+d(y,z)<d(x,y)$.

Pour l'indice de Kulczinski (5) on a :

$d(x,y)=1$; $d(x,z)= d(y,z)=1/4$, de sorte que $d(x,z)+d(y,z)<d(x,y)$.

Pour l'indice de Sokal et Sneath (7) on a :

$d(x,y)=2/(2n-2)$; $d(x,z)=d(y,z)=1/(2n-1)$, de sorte que

$d(x.z)+d(y,z)<d(x,y)$.

Pour l'indice de Kulczinski (6), on doit avoir n>2, pour que le domaine de définition contienne trois points distincts; et dans ce cas, soient x,y,z vérifiant :

x=(1 0 1 0..0) , y=(0 1 1 0..0) , z=(1 1 1 0..0)

Alors $d(x,y)=2$; $d(x,z)=d(y,z)=1/2$, de sorte que $d(x,z)+d(y,z)<d(x,y)$.

□

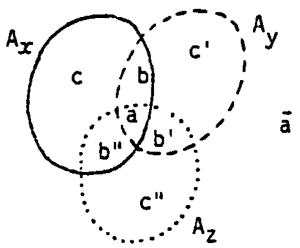
Pour les démonstrations futures, nous ferons appel aux notations suivantes x,y,z étant trois éléments de X, on désigne le nombre d'éléments i de I vérifiant,

$$(8)\begin{cases} x_i=y_i=z_i=1 \quad \text{par a} \quad , \; x_i=y_i=z_i=0 \quad \text{par } \bar{a} \\ x_i=y_i=1 \; , \; z_i=0 \quad (\text{resp. } y_i=z_i=1 \; , \; x_i=0 \; , \; \text{resp. } z_i=x_i=1 \; , \; y_i=0) \\ \text{par b(resp. b', resp. b")} \\ x_i=1 \; , \; y_i=z_i=0 \; (\text{resp. } y_i=1 \; , \; z_i=x_i=0 \; , \; \text{resp. } z_i=1 \; , x_i=y_i=0) \\ \text{par c (resp.c', resp.c")} \end{cases}$$

A une permutation près sur I on a :

	a	b	b'	b"	c	c'	c"	$\bar{a}$
$x=$	(1...1	1...1	0...0	1...1	1...1	0...0	0...0	0...0)
$y=$	(1...1	1...1	1...1	0...0	0...0	1...1	0...0	0...0)
$z=$	(1...1	0...0	1...1	1...1	0...0	0...0	1...1	0...0)

Et, pour une représentation ensembliste, ces nombres correspondent au schéma suivant:



Nous ferons également usage d'un lemme, qui est classique, excepté peut-être pour le résultat relatif aux points alignés. Nous dirons que trois points (deux à deux) distincts x,y,z d'un espace métrique (X,d), sont alignés (avec z entre x et y) ssi:
$d(x,y)=d(x,z)+d(y,z)$; cette définition est justifiée par le fait que si d est euclidienne, alors les trois points correspondants, soient M_x, M_y, M_z, dans une image euclidienne quelconque de (X,d), vérifient: $M_z \in]M_x \; M_y[$.

Lemme 1

Soient (X,d) un espace métrique et $f:\mathbb{R}_+ \mapsto \mathbb{R}_+$ *une fonction dérivable sur* $\mathbb{R}_+^*$, *vérifiant : i)* $f(0)=0$; *ii) f est croissante; iii) f' est strictement décroissante. Alors* $\delta=f(d)$ *est une distance sans triplet de points distincts alignés.*

Preuve :
Par i), δ est une dissimilarité propre.
Pour tout réel u strictement positif, soit $g_u:\mathbb{R}_+ \mapsto \mathbb{R}_+$ telle que
$\forall t\in\mathbb{R}_+$, $g_u(t)=f(t)-f(u+t)$.
Par iii), g_u est strictement croissante.
D'où : $\forall v\in\mathbb{R}_+$, $g_u(0)\leqslant g(v)$ et donc $f(u+v)\leqslant f(u)+f(v)$,
l'inégalité étant stricte si $v>0$ (on constate que l'inégalité reste vraie pour $u=0$).
Dès lors, utilisant ce résultat et la croissance de f :
$\forall(x,y,z)\in X^3$, $\delta(x,y)\leqslant f[d(x,z)+d(y,z)]\leqslant\delta(x,z)+\delta(y,z)$,
la dernière inégalité étant stricte si x,y,z sont distincts.

□

On en déduit le

Lemme 2

Soient (X,d) *un espace métrique et* α *un réel strictement positif. Alors* $\delta=d/(\alpha+d)$ *est une distance sans triplet de points distincts alignés.*

Preuve :
Il suffit d'appliquer le lemme 1 avec f telle que: $\forall t\in\mathbb{R}_+$, $f(t)=t/(\alpha+t)$

□

On en déduit immédiatement le :

Corollaire 1

Pour tout n, l'indice de Rogers et Tanimoto (2) est une distance sans triplet de points distincts alignés.

Il suffit d'appliquer le lemme 2 à la distance de Hamming.

Proposition 2

Pour tout n, l'indice d_θ *défini en (1) est une distance ssi* $\theta \leqslant 1$.
Pour $\theta \leqslant 1$, *il existe un triplet de points distincts alignés (et il ne peut alors exister quatre points distincts trois à trois alignés) ssi* $\theta = 1$.
Pour $\theta = 1$, *trois points distincts* x,y,z *sont alignés (avec* z *entre* x *et* y*) ssi :* $A_z = A_x \cup A_y$.

Ainsi l'indice de Jaccard est une distance avec des triplets de points distincts alignés; l'indice de Sokal et Sneath-Anderberg est une distance sans triplet de points distincts alignés; et l'indice de Czekanowski-Dice n'est pas une distance.

Preuve :
Commençons par prouver le résultat pour l'indice de Jaccard ($\theta=1$), noté plus simplement d. L'inégalité triangulaire étant triviale si deux éléments sont égaux, soient x,y,z, trois éléments distincts quelconques de X. Alors :

$$q_{xz}+q_{yz}-q_{xy} = (n_{xz}+q_{xz})d(x,z)+(n_{yz}+q_{yz})d(y,z)-(n_{xy}+q_{xy})d(x,y).$$

D'où, utilisant les relations (8), et notant $(n-\bar{a})$ par N, on a d'une part:

$$q_{xz}+q_{yz}-q_{xy}=(N-c')d(x,z)+(N-c)d(y,z)-(N-c")d(x,y)$$

et d'autre part:

$$q_{xz}+q_{yz}-q_{xy}=(b+b'+c+c")+(b+b"+c'+c")-(b'+b"+c+c')=2(b+c"),$$

relation classique, assurant en particulier que l'indice de Hamming est une distance.
En conséquence :

$$N[d(x,z)+d(y,z)-d(x,y)]=c'd(x,z)+cd(y,z)+2b+c"[2-d(x,y)].$$

Dès lors, puisque $d(x,y)\leqslant 1$, on a $d(x,y)\leqslant d(x,z)+d(y,z)$,
l'inégalité étant une égalité ssi: $b=c=c'=c"=0$, *i*.e.ssi: $A_z = A_x \cup A_y$.

Enfin, soient trois points distincts alignés x,y,z avec z entre x et y. Soit t un quatrième point distinct des trois premiers. Le point t ne peut être entre x et y,car $A_t = A_x \cup A_y$ entraîne $A_t = A_z$ et donc $t=z$. On ne peut avoir y entre x et t, car $A_y = A_x \cup A_t$ entraîne $A_z = A_x \cup A_y = A_x \cup A_t = A_y$, i.e. $z=y$; de même, on ne peut avoir x entre y et t. Les quatre points ne peuvent donc être trois à trois alignés.

Appliquons maintenant le lemme 2 à l'indice de Jaccard d:
$\forall \alpha \in R_+^* , \forall (x,y) \in X^2 , \delta(x,y) = \frac{d(x,y)}{\alpha+d(x,y)} = \frac{q_{xy}}{\alpha n_{xy}+(\alpha+1)q_{xy}} = \frac{1}{\alpha+1} \cdot \frac{q_{xy}}{\frac{\alpha}{\alpha+1} n_{xy}+q_{xy}}$

Dès lors, pour tout $\theta<1$, il suffit de choisir α tel que $\alpha/(\alpha+1)=\theta$, *i.e.* $\alpha=\theta/(1-\theta)$, pour assurer que $d_\theta/(\alpha+1)$, et par voie de conséquence d_θ, est une distance sans triplet de points distincts alignés.
Enfin, pour $\theta>1$, soient x,y,z de X tel que :
$x=(1\ 0\ 0\ ..\ 0)$, $y=(0\ 1\ 0\ ..\ 0)$, $z=(1\ 1\ 0\ ..\ 0)$. Alors :
$d_\theta(x,y)=1$, $d_\theta(x,z)=d_\theta(y,z)=1/(\theta+1)$; de sorte que :
$d_\theta(x,y)>d_\theta(x,z)+d_\theta(y,z)$.

□

Proposition 3

L'indice de Russel et Rao (3) est une distance telle qu'existe un triplet de points distincts alignés (et non quatre points distincts trois à trois alignés).
Trois points distincts x,y,z sont alignés (avec z entre x et y) ssi:
$A_z=A_x \cup A_y$ *et* $A_z=I$

Preuve:
Soient x,y,z trois points distincts quelconques de X. Utilisant les relations (8):
$n[d(x,z)+d(y,z)-d(x,y)]=(n-a-b")+(n-a-b')-(n-a-b)=(n-a-b"-b')+b$.
D'où l'inégalité triangulaire: $d(x,y) \leqslant d(x,z)+d(y,z)$,
l'inégalité étant une égalité ssi: $b=c=c'=c"=\tilde{a}=0$, *i.e.* $A_z=I$ et $A_z=A_x \cup A_y$.
Enfin, l'une de ces conditions ($A_z=I$) montre clairement qu'il ne peut exister quatre points distincts trois à trois alignés.

□

Abordons maintenant l'aspect euclidien. Les contre-exemples suivants montrent que pour tout $n \geqslant 3$, la distance de Russel et Rao (3), et pour tout n suffisament grand, la distance de Rogers et Tanimoto (2) ainsi que la distance $d_\theta(\theta \leqslant 1)$ donnée en (1), ne sont pas euclidiennes sur X^+.

Contre-exemples 2

Pour tout $n\geq 3$, soient x,y,z,t de X^+ (fonctions de n), vérifiant:
$x=(1\ 1\ 0\ 1\ ..\ 1)$, $y=(1\ 0\ 1\ 1\ ..\ 1)$, $z=(0\ 1\ 1\ 1\ ..\ 1)$, $t=(1\ 1\ 1\ 1..1)$.
Soit p ($p\geq 3$) un entier fixé. Avec les quatre points précédents et n=p, l'indice de Russel et Rao (3), noté d,vérifie :
$pd(x,y)=pd(x,z)=pd(y,z)=2$; $pd(x,t)=pd(y,t)=pd(z,t)=1$.

S'il existait une image euclidienne,soit $\{M_x,M_y,M_z,M_t\}$,de d,les premières conditions montrent que M_x, M_y, M_z, formeraient un triangle équilatéral; et les dernières imposeraient à M_t d'être au milieu de chacun des côtés de ce triangle. d n'est donc pas euclidienne.
Pour tout $n\geq 3$, notons d_n l'indice de Rogers et Tanimoto (2) et pour θ ($0<\theta\leq 1$) fixé, d_θ^n l'indice défini en (1). Alors :

$$\begin{cases} d_n(x,y)=d_n(x,z)=d_n(y,z)=2/(n+2) \\ d_n(x,t)=d_n(y,t)=d_n(z,t)=1/(n+1) \end{cases}$$

$$\begin{cases} d_\theta^n(x,y)=d_\theta^n(x,z)=d_\theta^n(y,z)=2/[(n-2)\theta+2] \\ d_\theta^n(x,t)=d_\theta^n(y,t)=d_\theta^n(z,t)=1/[(n-1)\theta+1] \end{cases}$$

Pour un passage à la limite, il est nécessaire d'oeuvrer sur un ensemble de points indépendant de n. Soit donc $Y=\{\underline{x},\underline{y},\underline{z},\underline{t}\}$ un ensemble à quatre éléments, et définissons les suites de dissimilarités $\{\delta_n,n\geq 3\}$ et $\{\delta_\theta^n,n\geq 3\}$ sur Y par, $\forall n\geq 3$:

$$\begin{cases} \delta_n(\underline{x},\underline{y})=\delta_n(\underline{x},\underline{z})=\delta_n(\underline{y},\underline{z})=2/(n+2) \\ \delta_n(\underline{x},\underline{t})=\delta_n(\underline{y},\underline{t})=\delta_n(\underline{z},\underline{t})=1/(n+1) \end{cases}$$

$$\begin{cases} \delta_\theta^n(\underline{x},\underline{y})=\delta_\theta^n(\underline{x},\underline{z})=\delta_\theta^n(\underline{y},\underline{z})=2/[(n-2)\theta+2] \\ \delta_\theta^n(\underline{x},\underline{t})=\delta_\theta^n(\underline{y},\underline{t})=\delta_\theta^n(\underline{z},\underline{t})=1/[(n-1)\theta+1] \end{cases}$$

Munissons $\mathcal{D}_Y$ d'une norme quelconque. Il est clair que lorsque $n \to \infty$, $n\delta_n$ et $n\theta\delta_\theta^n$ tendent vers une limite commune, soit δ; et,à une bijection prês entre Y et le sous-ensemble $\{x,y,z,t\}$ de X (n fixé égal à p) évoqué en tête des contre-exemples pour l'indice de Russel et Rao d, δ n'est autre que p d. δ est donc non-euclidienne et appartient à l'ouvert, complémentaire de $\mathcal{D}_Y^e$. Cela suffit pour affirmer que pour tout n suffisamment grand, $n\delta_n$ et $n\theta\delta_\theta^n$, et par voie de conséquence δ_n et δ_θ^n, ne sont pas euclidiennes. Et revenant aux définitions de δ_n et δ_θ^n, on en déduit que d_n et d_θ^n ne sont pas euclidiennes pour tout n suffisamment grand.

□

Remarque 6

Utilisant les propositions 1 et 2, une démonstration plus simple peut être proposée pour les indices de Jaccard et de Russel et Rao, valable pour $n \geq 2$. En effet, notant d ces indices, on a: $\forall n \geq 2$, $\forall x \in X^+$, $d(\underline{0},x)=1$. De sorte que s'il existait une image euclidienne, les points correspondants de X^+ seraient une sphère, ce qui est incompatible avec l'existence de trois points distincts alignés de X^+.

Et même pour un résultat plus fort en restreignant l'assertion à X^+, comme cela est fait dans les contre-exemples, la démonstration suivante prévaudrait encore pour $n \geq 3$. Soient x,y,z,t distincts de X^+ vérifiant: $A_t = A_x \cup A_y = A_x \cup A_z$. S'il existait une image euclidienne, les quatre points correspondants seraient alignés et x,y,z,t seraient trois à trois alignés, ce qui contredit les propositions. Notons d'ailleurs que les quatre points en tête des contre-exemples 2, satisfont à ces conditions.

Remarque 7

Pour n=2, l'indice de Russel et Rao (3) n'est pas euclidien sur X (c.f. remarque précédente), et il est naturellement euclidien sur X^+.

A l'aide des quatre points x,y,z,t en tête des contre-exemples, le calcul montre que l'indice de Rogers et Tanimoto (2) n'est pas euclidien sur X^+ dès que $n \geq 6$. En fait, on peut montrer qu'il n'est pas euclidien sur X^+ pour $n \geq 3$. Pour $n \geq 4$, il suffit de considérer les cinq points:
x=(1 1 0 0 0..0), y=(1 0 1 0 0..0), z=(0 1 1 0 0..0), t=(1 1 1 0 0..0), u=(1 1 1 1 0..0), et pour n=3, il suffit de considérer les quatre points:
x=(1 1 0), y=(1 0 1), z=(0 1 1), t=(1 0 0).

On peut montrer également qu'il n'est pas euclidien sur X pour n=2, alors qu'il est naturellement euclidien sur X^+.
Enfin le calcul à l'aide des quatre points en tête des contre-exemples, montre que l'indice de Sokal et Sneath (θ=1/2) donné en (1), n'est pas euclidien sur X^+ dès que $n \geq 6$.

5. STRUCTURE DES RACINES CARREES DES DISSIMILARITES

La nature euclidienne des racines carrées des dissimilarités, sera abordée à travers la semi-définie positivité des matrices de similarité. Deux indices feront exception: les indices de Kulczinski définis en (5) et (6).

D'ailleurs pour l'indice (6), la similarité s associée à la dissimilarité d vérifie: s=1/d; pour tout x, on a $s(x,x)=+\infty$, de sorte que l'on ne saurait parler de matrice de similarité s.d.p.
Les contre-exemples suivants montrent que pour tout $n \geq 3$, la racine carrée de l'indice de Kulczinski (5), et pour tout $n \geq 5$ la racine carrée de l'indice de Kulczinski (6), ne sont pas des distances.

Contre-exemples 3
Pour l'indice (5) noté d, soient x,y,z de X^+ vérifiant:
x=(1 0 0 0..0), y=(0 1 1 0..0), z=(1 1 1 0..0).
Alors: $d(x,y)=1$; $d(x,z)=1-(\frac{1}{2})(1+1/3)=1/3$; $d(y,z)=1-(\frac{2}{2})(1/2+1/3)=1/6$.

Or $\left(\frac{1}{\sqrt{3}}+\frac{1}{\sqrt{6}}\right)^2=\left(\frac{1}{2}+\frac{2}{\sqrt{18}}\right)<1$, de sorte que $\sqrt{d(x,z)}+\sqrt{d(y,z)}<\sqrt{d(x,y)}$.
Pour l'indice (6) noté encore d, soient x,y,z de $\bar{X}$ vérifiant:
x=(1 0 1 0 1 0..0), y=(0 1 0 1 1 0..0), z=(1 1 1 1 1 0..0).
Alors: $d(x,y)=4$; $d(x,z)=d(y,z)=2/3$,
de sorte que $\sqrt{d(x,z)}+\sqrt{d(y,z)}<\sqrt{d(x,y)}$

□

Remarque 8
Pour un nombre de signes n inférieur aux contraintes indiquées, la structure métrique ou euclidienne peut être assurée. Ainsi par exemple, pour n=2, la racine carrée de l'indice de Kulczinski (5) est une distance (euclidienne) sur X^+.

Rappelons maintenant que si $A=(a_{ij})$ et $B=(b_{ij})$ sont deux matrices d'ordre (p,q), le produit de Schur (ou de Hadamard)de A et B est la matrice d'ordre (p,q), notée $A*B$ de terme général $a_{ij}b_{ij}$. La loi * est clairement distributive par rapport à l'addition.

Pour étudier la semi-définie positivité des matrices de similarité, nous ferons usage des lemmes suivants.

Lemme 3

*Soient A et B deux matrices symétriques de même dimension. Alors si A et B sont s.d.p., il en est de même de A*B.*

C'est le lemme de Schur que l'on trouve dans Rao p.77 [15], Basilevski p.138 [1], Halmos p.173 [11], Mirsky p.421 [14], Gower [10]. Nous en rappelons une démonstration très brève.

Preuve:

Soient $\lambda_1,..\lambda_r$ les r valeurs propres (distinctes ou non) strictement positives de A, et $\{Y_1,..,Y_r\}$ un système de vecteurs propres orthonormés respectivement associés. Alors $A=\sum_k \lambda_k Y_k Y_k'$

De même, soient $\mu_1,..,\mu_s$, les s valeurs propres (distinctes ou non) strictement positives de B, et $\{Z_1,..,Z_s\}$ un système de vecteurs propres orthonormés respectivement associés. Alors $B=\sum_\ell \mu_\ell Z_\ell Z_\ell'$.

D'où: $A*B= \sum_k\sum_\ell \lambda_k\mu_\ell (Y_kY_k')*(Z_\ell Z_\ell')$.

Mais pour tout k et tout ℓ, si $T_{k\ell}=Y_k*Z_\ell$, on constate, par définition de la loi *, que $(Y_kY_k')*(Z_\ell Z_\ell')=T_{k\ell}T_{k\ell}'$.

D'où le résultat, l'ensemble des matrices s.d.p. formant un cône convexe. □

Par induction, on en déduit que si A est s.d.p., la composée k fois de A, notée A^{*k}, est s.d.p.

Lemme 4

Soit $A=(a_{ij})$ une matrice symétrique s.d.p. Soit $\alpha>0$ tel que pour tout i et tout j: $|a_{ij}|<\alpha$. Alors la matrice B de terme général $b_{ij}=1/(\alpha-a_{ij})$ est s.d.p.

Ce lemme a été établi par Gower -[10]-. Nous en rappelons également une démonstration très brève.

Preuve:
Puisque $|a_{ij}|<\alpha$, l'on a: $b_{ij}=(1/\alpha)[1+\sum_{k=1}^{\infty}(a_{ij}/\alpha)^k]$.
D'où, munissant l'ensemble des matrices d'ordre considéré d'une norme quelconque, l'on a:

$$B=(1/\alpha)\mathbb{1}\mathbb{1}'+\lim_{K\to\infty}\sum_{k=1}^{K}(1/\alpha^{k+1})A^{*k}.$$

On en déduit le résultat, puisque l'ensemble des matrices s.d.p. forme un cône fermé.

□

Dans tout ce qui suit, lorsque nous parlerons de matrices à éléments indicés par l'ensemble X, il sera supposé que X est identifié à $\{1,..,2^n\}$, dans un bijection faisant correspondre $\underline{0}$ et 2^n.

Lemme 5
Les matrices de terme général n_{xy} et $n_{\bar{x}\bar{y}}$, $(x,y)\in X^2$, sont s.d.p.

Preuve:
Soient A la matrice d'ordre $(2^n,n)$, dont toutes les lignes x sont formées des composantes $x_1,..,x_n$ de x, et $\mathbb{1}$ la matrice de même ordre n'ayant que des 1 pour composantes. Alors les matrices de terme général n_{xy} et $n_{\bar{x}\bar{y}}$ sont respectivement égales à :
AA' et $(\mathbb{1}-A)(\mathbb{1}-A)'$.

□

Proposition 4
Les matrices de similarité associées aux indices de Rogers et Tanimoto (2), Russel et Rao (3), et Ochiai (4), sont s.d.p.
Ainsi les racines carrées des trois indices mentionnés, sont euclidiennes.

Preuve:
Pour l'indice de Rogers et Tanimoto, la similarité s vérifie:

$\forall(x,y)\in X^2$, $s(x,y)=1-d(x,y)=(n-q_{xy})/(n+q_{xy})=(n_{xy}+n_{\bar{x}\bar{y}})/[2n-(n_{xy}+n_{\bar{x}\bar{y}})]$.
Dès lors, il suffit d'appliquer successivement les lemmes 5,4, et 3 pour s'assurer que les matrices de terme général $(n_{xy}+n_{\bar{x}\bar{y}})$, $1/[2n-(n_{xy}+n_{\bar{x}\bar{y}})]$ et $s(x,y)$, sont s.d.p.
Pour l'indice de Russel et Rao, la similarité s vérifie:

$$\forall(x,y)\in X^2,\ s(x,y)=1-d(x,y)=\begin{cases} n_{xy}/n & \text{si } x\neq y \\ 1 & \text{sinon} \end{cases}$$

Si A est la matrice de terme général n_{xy}/n, et si Δ est la matrice diagonale d'éléments diagonaux $(1-n_x/n)$, la matrice de similarité est égale à: $A+\Delta$; elle est donc s.d.p. par le lemme 5.
Enfin pour l'indice d'Ochiaï, la similarité s vérifie:
$\forall(x,y)\in X^{+2}$, $s(x,y)=n_{xy}/\sqrt{n_x n_y}$.
Si B est la matrice colonne de terme général n_x, $x\in X^+$, la matrice de terme général $1/\sqrt{n_x n_y}$, égale à BB', est s.d.p..Dès lors le résultat découle des lemmes 5 et 3. □

Pour étudier les racines carrées des indices d_θ définis en (1), nous ferons usage des lemmes suivants.

Lemme 6

Soient $a_1,..,a_p$, p nombres réels strictement positifs. Alors la matrice A_p de terme général $1/(a_i+a_j)$, i,j=1..p, est s.d.p; elle est définie-positive ssi les nombres $a_1,..a_p$ sont deux à deux distincts.

Preuve:
Clairement le résultat est vrai pour p=1. Raisonnons alors par récurrence, le supposant vrai à l'ordre (p-1), $p\geq 2$. Soit $Y=(y_1,...,y_p)$ quelconque de $\mathbb{R}^p$. Une décomposition en blocs donne :

$$Y'A_pY=(Z'\vdots y_p)\begin{bmatrix} A_{p-1} & B \\ B' & 1/2a_p \end{bmatrix}\begin{bmatrix} Z \\ \dots \\ y_p \end{bmatrix}$$ où B est la matrice colonne de terme général $\dfrac{1}{a_i+a_p}$

$$\text{D'où: } Y'A_pY=(Z'\vdots y_p)\begin{bmatrix} A_{p-1}Z+y_pB \\ B'Z+y_p/2a_p \end{bmatrix}=Z'A_{p-1}Z+2y_pZ'B+y_p^2/2a_p.$$

On a ainsi un trinome du second-degré en y_p, dont le discriminant réduit vaut:

$\Delta'=(Z'B)^2-(1/2a_p)Z'A_{p-1}Z = Z'[BB'-(1/2a_p)A_{p-1}]Z$,

soit $\Delta'=Z'CZ$ où $C=BB'-(1/2a_p)\,A_{p-1}$.

Le terme général de C est égal à: $c_{ij}=\dfrac{1}{(a_p+a_i)(a_p+a_j)} - \dfrac{1}{2a_p(a_i+a_j)}$,

soit $c_{ij}= \dfrac{2a_p(a_i+a_j)-(a_p+a_i)(a_p+a_j)}{2a_p(a_p+a_i)(a_p+a_j)(a_i+a_j)} = -\dfrac{1}{2a_p}\cdot\dfrac{(a_p-a_i)}{(a_p+a_i)}\cdot\dfrac{(a_p-a_j)}{(a_p+a_j)}\cdot\dfrac{1}{a_i+a_j}$,

$i,j=1,..(p-1)$.

D'où, si D est la matrice colonne de terme général $(a_p-a_i)/(a_p+a_i)$:

$C=-(1/2a_p)(DD')*A_{p-1}$.

La matrice DD' est s.d.p., ainsi que A_{p-1} par l'hypothèse d'induction.

Dès lors, par le lemme de Schur, C est semi-définie négative.

On en déduit que l'on a $\Delta'\leqslant 0$ et, par voie de conséquence, $Y'A_pY\geqslant 0$. A_p est donc s.d.p. Supposons maintenant les nombres $a_1,..,a_p$ deux à deux distincts et, par hypothèse d'induction, A_{p-1} définie-positive. Revenons au terme général c_{ij} de C.

La matrice A_{p-1} qui est de rang plein, reste de rang plein, en multipliant ses colonnes respectivement par les nombres non nuls $(a_p-a_j)/(a_p+a_j)$, puis ses lignes respectivement par les nombres non nuls $(a_p-a_i)/(a_p+a_i)$; de sorte que C, de rang plein et semi-définie négative, est définie négative. Dès lors, si $Y'A_pY=0$, nécessairement $\Delta'=0$ et donc $Z=0$; et revenant au trinome du second degré en y_p, cela entraîne: $y_p=0$. D'où $Y=0$ et A_p est définie positive. Enfin si pour $i\neq j$ on a $a_i=a_j$, les lignes (ou colonnes) correspondantes de A_p sont égales, et A_p ne peut être définie-positive.

□

Lemme 7

Soient (X,d) *un espace métrique et* α *un réel strictement positif.*
Alors $d/\sqrt{\alpha+d^2}$ *est une distance sans triplet de points distincts alignés.*

Preuve :

Il suffit d'appliquer le lemme 1 avec f telle que: $\forall t\in\mathbb{R}_+$, $f(t)=t/\sqrt{\alpha+t^2}$.

La fonction f est dérivable et vérifie: $\forall t\in\mathbb{R}_+^*$, $f'(t)=\alpha/(\alpha+t^2)^{3/2}$

□

Lemme 8

Soient (X,d) *un espace métrique tel que:* $\forall (x,y)\in X^2, d(x,y)\leq 1$,
et γ *un réel tel que:* $0<\gamma<3$.
Soit $f:[0,1]\to\mathbb{R}_+$ telle que $t\mapsto f(t)=\sqrt{t/(3-\gamma t)}$.
Alors $\delta=f(d)$ *est une distance dès que :* $\gamma\leq 2$.
Pour $\gamma<2$, *il n'existe pas de triplet de points distincts alignés.*
Pour $\gamma=2$, *trois points distincts* x,y,z *sont alignés (avec* z *entre* x *et* y*)*
au sens de δ *ssi:* $d(x,y)=1$; $d(x,z)=d(y,z)=1/2$.

Ainsi pour $\gamma=2$, trois points alignés au sens de δ, le sont au sens de d;
et on constate qu'ils vérifient les mêmes relations:
$\delta(x,y)=1$; $\delta(x,z)=\delta(y,z)=1/2$.

Preuve :
Commençons par démontrer le résultat pour $\gamma=2$.

Pour la fonction f, montrons les propriétés suivantes:
i) f est dérivable sur $]0\ 1[$ et l'on a: $f'(t)=\dfrac{3/2}{\sqrt{t}(3-2t)^{3/2}}$

ii) f est strictement croissante et vérifie: $f(0)=0$; $f(1/2)=1/2$; $f(1)=1$.
iii) $\forall h\in]0\ 1/2[$, $f(1/2-h)+f(1/2+h)>1$.
Soit $\phi:[0\ 1]\mapsto\mathbb{R}_+$ telle que $\phi(t)=t/(3-2t)$.
ϕ est dérivable sur $]0\ 1[$ et l'on a: $\phi'(t)=3/(3-2t)^2>0$; d'où les points i) et ii). Enfin iii) est assuré en constatant que sont vraies les équivalences suivantes :
$f(1/2-h)+f(1/2+h)>1$

$$\Longleftrightarrow \sqrt{\frac{1/2-h}{3-2(1/2-h)}}+\sqrt{\frac{1/2+h}{3-2(1/2+h)}} > 1$$

$$\Longleftrightarrow \sqrt{\frac{1-2h}{1+h}}+\sqrt{\frac{1+2h}{1-h}} > 2$$

$$\Longleftrightarrow \frac{(1-2h)(1-h)+(1+2h)(1+h)}{1-h^2}+2\sqrt{\frac{1-4h^2}{1-h^2}} > 4$$

$$\Leftrightarrow \quad \frac{1+2h^2}{1-h^2} + \sqrt{\frac{1-4h^2}{1-h^2}} > 2$$

$$\Leftrightarrow \quad \sqrt{\frac{1-4h^2}{1-h^2}} > \frac{1-4h^2}{1-h^2}$$

$$\Leftrightarrow \quad 1-4h^2<1-h^2.$$

δ est clairement une dissimilarité propre. Soient x,y,z distincts de X; pour montrer l'inégalité triangulaire $\delta(x,y)\leqslant\delta(x,z)+\delta(y,z)$, considérons deux cas, suivant la valeur de $d(x,z)+d(y,z)$.

a) $d(x,z)+d(y,z)\geqslant 1$.

Si on a $d(x,z)\geqslant 1/2$ et $d(y,z)\geqslant 1/2$, alors par la croissance de f: $\delta(x,z)\geqslant 1/2$, $\delta(y,z)\geqslant 1/2$.
Comme $\delta(x,y)\leqslant 1$ (car $d(x,y)\leqslant 1$), l'inégalité triangulaire est assurée; de plus cette inégalité est une égalité ssi: $\delta(x,y)=1, \delta(x,z)=\delta(y,z)=1/2$, i.e. $d(x,y)=1$, $d(x,z)=d(y,z)=1/2$.
Si l'un des deux nombres $d(x,z)$ et $d(y,z)$, soit par exemple $d(x,z)$, est inférieur à 1/2, posons $d(x,z)=1/2-h$ avec $0<h<1/2$ et $d(y,z)=1/2+h'$.
Nécessairement: $h\leqslant h'$.
Alors, utilisant la croissance de f et iii):
$\delta(x,z)+\delta(y,z)=f(1/2-h)+f(1/2+h')\geqslant f(1/2-h)+f(1/2+h)>1$.
D'où l'inégalité triangulaire stricte.

b) $d(x,z)+d(y,z)<1$.

$\forall u\in]0\ 1[$, soit $g_u:[0\ (1-u)]\to\mathbb{R}_+$, telle que $g_u(t)=f(t)-f(t+u)$.
Pour $t\in]0\ (1-u)[$, g_u est dérivable, et par i):

$$g_u'(t)=\frac{3/2}{\sqrt{t}(3-2t)^{3/2}} - \frac{3/2}{\sqrt{t+u}[3-2(t+u)]^{3/2}} \quad .$$

Montrons:
iv) $\forall v\in\mathbb{R}_+$ tel que $u+v<1$, $g_u(0)\leqslant g_u(v)$, l'inégalité étant stricte si $v>0$.
Soit ψ: $[0\ 1]\to\mathbb{R}_+$ telle que $\psi(t)=t(3-2t)^3$.
ψ est dérivable sur $]0\ 1[$ et $\psi'(t)=(3-2t)^2[(3-2t)-6t]=(3-2t)^2(3-8t)$.

En conséquence, ψ est croissante de 0 à 3/8 et décroissante de 3/8 à 1.
Mais on a $g'_u(t)=0$ ssi $\psi(t)=\psi(t+u)$; en conséquence, avec les propriétés de la fonction ψ, g'_u s'annule au plus en point de $]0\ (1-u)[$.
Comme d'une part g'_u est continue et d'autre part $\lim_{t\downarrow 0} g'_u(t)=+\infty$, on en déduit que g_u est soit strictement croissante et iv) en découle, soit croissante de 0 à $\hat{t}$ puis décroissante de $\hat{t}$ à $(1-u)$; et dans ce dernier cas, iv) sera encore vrai ssi: $g_u(0)\leqslant g_u(1-u)$. (La condition $u+v<1$ entraîne $g_u(0)<g_u(v)$ si $v>0$). On a les équivalences suivantes:
$g_u(0)\leqslant g_u(1-u) \Longleftrightarrow -f(u)\leqslant f(1-u)-1 \Longleftrightarrow 1\leqslant f(u)+f(1-u)$.
Posons $u=1/2+\varepsilon h$ $(0\leqslant h<1/2\ ,\ \varepsilon=\pm 1)$.
Alors:
$g_u(0)\leqslant g_u(1-u) \Longleftrightarrow 1\leqslant f(1/2+h)+f(1/2-h)$,
et désormais iii) assure iv).
Dès lors iv) donne:
$\forall v\in\mathbb{R}_+$ tel que $u+v<1$, $-f(u)\leqslant f(u)-f(u+v)$, soit $f(u+v)\leqslant f(u)+f(v)$,
l'inégalité étant stricte si $v\neq 0$ (on constate que l'inégalité reste vraie pour $u=0$).
Utilisant ce résultat et la croissance de f:
$\delta(x,y)\leqslant f[d(x,z)+d(y,z)]$, car $d(x,z)+d(y,z)<1$,
soit $\delta(x,y)<\delta(x,z)+\delta(y,z)$, l'inégalité étant stricte car x,y,z sont distincts.

Pour achever la démonstration du lemme, il suffit de montrer que si $\sqrt{d/(3-\gamma d)}$ est une distance pour un certain γ vérifiant $0<\gamma<3$, alors pour tout γ' tel que $0<\gamma'<\gamma$ $\sqrt{d(3-\gamma' d)}$ est une distance sans triplet de points distincts alignés. Dans ce but, appliquons le lemme 7.

$\forall\alpha\in\mathbb{R}_+^*, \sqrt{\frac{d}{\alpha(3-\gamma d)+d}}$, i.e. $\frac{1}{\sqrt{\alpha}}\sqrt{\frac{d}{3-(\gamma-1/\alpha)d}}$ est une distance sans triplet de points distincts alignés.
Dès lors, si γ' vérifie $0<\gamma'<\gamma$, il suffit de choisir α tel que $\gamma-1/\alpha=\gamma'$, i.e.$\alpha=1/(\gamma-\gamma')$, pour assurer que $\sqrt{d/(3-\gamma' d)}$ est une distance sans triplet de points distincts alignés.

□

Remarque 9

La dernière partie de la démonstration montre que pour tout $\gamma>2$, $\sqrt{d(3-\gamma d)}$ n'est pas une distance dès que pour $\gamma=2$, δ admet un triplet de points distincts alignés. Si X est fini, cette dernière condition est d'ailleurs nécessaire, car si pour $\gamma=2$, δ n'admet pas de points distincts alignés, alors δ appartient à l'intérieur du cône des distances sur X.

Proposition 5

Pour la famille $\{d_\theta, \theta\in\mathbb{R}_+^*\}$ *définie en (1), on a:*

i) Pour tout n et pour $\theta\leqslant 2$, *la matrice de similarité associée à* d_θ *est s.d.p.*

ii) Pour $n\geqslant 4$ *et pour* $\theta>2$, $\sqrt{d_\theta}$ *n'est pas euclidien sur* X^+.

iii) Pour tout n, $\sqrt{d_\theta}$ *est une distance ssi* $\theta\leqslant 3$; *pour* $\theta\leqslant 3$, *il existe un triplet de points distincts alignés (et il ne peut alors exister quatre points distincts trois à trois alignés) ssi:* $\theta=3$; *pour* $\theta=3$, x,y,z *distincts sont alignés (avec z entre x et y) ssi :* $A_x\cap A_y=\emptyset$, $A_x\cup A_y=A_z$, $|A_x|=|A_y|$.

Ainsi les racines carrées des indices de Sokal et Sneath-Anderberg, Jaccard et Czekanowski-Dice, sont des distances euclidiennes.

Preuve:

i) Pour tout $\theta>0$, notons s_θ la similarité associée à d_θ et s_θ^+ sa restriction à X^+. Alors:

$\forall(x,y)\neq(\underline{0},\underline{0})$ de X^2, $s_\theta(x,y)=1-d_\theta(x,y)=\theta n_{xy}/(\theta n_{xy}+q_{xy})$.

D'où pour $\theta=2$: $\forall(x,y)\in X^{+2}$, $s_2^+(x,y)=2n_{xy}/(2n_{xy}+q_{xy})=2n_{xy}/(n_x+n_y)$.

Dès lors, il suffit d'appliquer successivement les lemmes 6, 5, et 3, pour affirmer que les matrices de terme général $1/(n_x+n_y)$, n_{xy} et $s_2^+(x,y)$, $(x,y)\in X^{+2}$, sont s.d.p.

$\forall\alpha>1$ et $\forall(x,y)\in X^{+2}$, $s_2^+(x,y)<\alpha$. D'où, appliquant successivement les lemmes 4 et 3 on en déduit que les matrices de terme général $1/[\alpha-s_2^+(x,y)]$ et $s_2^+(x,y)/[\alpha-s_2^+(x,y)]$ sont s.d.p.

Mais le terme général de cette dernière matrice est égal à:

$\frac{2}{\alpha} n_{xy}/[2\ \frac{(\alpha-1)}{\alpha} n_{xy}+q_{xy}]$, $(x,y)\in X^{+2}$.

Dès lors, pour tout $\theta<2$, il suffit de choisir α tel que $2(\alpha-1)/\alpha=\theta$, i.e. $\alpha=2/(2-\theta)$, pour assurer que la matrice de terme général $(2/\alpha\theta)s^+_\theta(x,y)$, et par voie de conséquence la matrice de terme général $s^+_\theta(x,y)$, est s.d.p. Enfin si S_θ(resp. S^+_θ) est la matrice de terme général $s_\theta(x,y)$, $(x,y)\in X^2$ (resp. $s^+_\theta(x,y)$,(x,y) X^{+2}), il est clair que: $S_\theta=\begin{bmatrix} S^+_\theta & 0 \\ 0 & I \end{bmatrix}$, de sorte que pour $\theta\leq 2$, S_θ est s.d.p.

ii) pour $n\geq 4$, soient x,y,z,t de X^+ tels que:
$x=(1\ 1\ 0\ 0\ 0..0)$, $y=(0\ 1\ 1\ 0\ 0..0)$, $z=(1\ 0\ 0\ 1\ 0..0)$, $t=(0\ 0\ 1\ 1\ 0..0)$.
On a: $d_\theta(x,y)=d_\theta(x,z)=d_\theta(t,y)=d_\theta(t,z)=2/(\theta+2)$; $d_\theta(y,z)=d_\theta(x,t)=1$.
S'il existait une image euclidienne, soit $\{M_x,M_y,M_z,M_t\}$ de $\sqrt{d_\theta}$, (M_x,M_y,M_z) et (M_t,M_y,M_z) formeraient deux triangles isocèles, de même base M_yM_z. Notant H le milieu de M_yM_z, on aurait:
$M_xH^2=M_tH^2=M_xM_y^2-M_yM_z^2/4=2/(\theta+2)-1/4=(6-\theta)/[4(\theta+2)]$.
Mais l'inégalité triangulaire $M_xM_t\leq M_xH+M_yH$, i.e. $M_xM_t^2\leq 4M_xH^2$, n'est satisfaite que si $1\leq(6-\theta)/(\theta+2)$, i.e. $\theta\leq 2$.

iii) L'indice de Jaccard ($\theta=1$), noté plus simplement d, satisfait aux hypothèses du lemme 8 (c.f. proposition 2). En conséquence, si $\delta:X^2\mapsto R_+$ est telle que $\forall(x,y)\in X^2$, $\delta(x,y)=\sqrt{d(x,y)/[3-2d(x,y)]}$, δ est une distance.
Mais pour $x\neq y$, on a: $\delta(x,y)=\sqrt{q_{xy}/(3n_{xy}+q_{xy})}$, de sorte que $\delta=\sqrt{d_3}$.
De plus, trois points distincts x,y,z sont alignés au sens de $\sqrt{d_3}$ (avec z entre x et y) ssi: $d(x,y)=1$, $d(x,z)=d(y,z)=1/2$.
Ces points étant alignés au sens de d, on a (c.f. proposition 2): $A_z=A_x\cup A_y$.
La condition $d(x,y)=1$ entraîne, par définition de d, $n_{xy}=0$, soit $A_x\cap A_y=\emptyset$.
De même, la condition $d(x,z)=1/2$ entraîne $q_{xz}=n_{xz}$, soit avec les deux dernières relations ensemblistes, $n_y=n_x$. Et réciproquement les conditions $A_z=A_x\cup A_y$, $A_x\cap A_y=\emptyset$, $n_x=n_y$ imposent à d les valeurs indiquées.
Enfin quatre points distincts ne peuvent être trois à trois alignés au sens de $\sqrt{d_3}$, puisqu'ils ne peuvent l'être au sens de d (c.f. proposition 2).
Appliquons maintenant le lemme 7 à la distance $\sqrt{d_3}$.
Pour tout $\alpha>0$, si $\tilde{\delta}: X^2\mapsto R_+$ est telle que $\forall(x,y)\in X^2$, $\tilde{\delta}(x,y)=\sqrt{d_3(x,y)}/\sqrt{\alpha+d_3(x,y)}$, $\tilde{\delta}$ est une distance sans triplet de points distincts alignés. Mais pour $x\neq y$,

$$\hat{\delta}(x,y)=(\frac{1}{\sqrt{\alpha+1}})\sqrt{q_{xy}/[\frac{3\alpha}{\alpha+1}\, n_{xy}+q_{xy}]}$$

Dès lors, pour $\theta<3$, il suffit de choisir α tel que $3\alpha/(\alpha+1)=\theta$, i.e. $\alpha=\theta/(3-\theta)$, pour assurer que $(1/\sqrt{\alpha+1})\sqrt{d_\theta}$, et par voie de conséquence $\sqrt{d_\theta}$, est une distance, sans triplet de points distincts alignés.
Enfin, soient x,y,z de X tels que:
$x=(1\ 0\ 0..0)$, $y=(0\ 1\ 0..0)$, $z=(1\ 1\ 0..0)$. Alors:

$d_\theta(x,y)=1$, $d_\theta(x,z)=d_\theta(y,z)=1/(\theta+1)$; de sorte que l'inégalité triangulaire $\sqrt{d_\theta(x,y)}\leq\sqrt{d_\theta(x,z)}+\sqrt{d_\theta(y,z)}$ n'est satisfaite que si:
$1\leq 2/\sqrt{\theta+1}$, soit $\theta\leq 3$.

□

Remarque 10

Pour $\theta\leq 1$, la semi-définie positivité de la matrice de similarité peut-être démontrée plus simplement, comme cela est fait pour les trois indices intervenant dans la proposition 4. En effet:
$\forall(x,y)\in X^{+2}$, $s^{\dagger}_{\theta}(x,y)=\theta n_{xy}/[n-(n_{\bar{x}\bar{y}}+(1-\theta)n_{xy})]$, de sorte que les Lemmes 5, 4 et 3 assurent le résultat.

Remarque 11

Pour $n\geq 4$, il y a donc équivalence entre la nature euclidienne de $\sqrt{d_\theta}$ et la semi-définie positivité de la matrice de similarité (celles-ci étant satisfaites ssi $\theta\leq 2$). Mais pour $n<4$, cette équivalence n'est pas assurée.
A titre d'exemple, pour $n=2$, le calcul montre que la matrice de similarité est s.d.p. ssi $\theta\leq 1+\sqrt{2}$, et $\sqrt{d_\theta}$ est euclidienne ssi $\theta\leq 1+\sqrt{3}$.

Pour l'indice de Sokal et Sneath défini en (7), on a la proposition suivante .

Proposition 6

Pour tout n, la racine carrée de l'indice de Sokal et Sneath (7) est une distance sans triplet de points distincts alignés.
Pour n suffisamment grand, cette distance est non-euclidienne sur X^+.

Preuve:
Considérons la distance de Hamming pondérée (M.G. Kendall, Sokal et Michener), notée d, et vérifiant: $\forall(x,y)\in X^2$, $d(x,y)=q_{xy}/n$.
Cette distance satisfait aux hypothèses du lemme 8. En conséquence, pour

tout réel γ vérifiant $0<\gamma<2$, est une distance sans triplet de points distincts alignés, la dissimilarité δ telle que :

$\forall(x,y)\in X^2$, $\delta(x,y)=\sqrt{q_{xy}/(3n-\gamma q_{xy})}$, i.e.

$\delta(x,y)=\sqrt{2/3}\ \sqrt{q_{xy}/(2n-\frac{2\gamma}{3}\ q_{xy})}$

Dès lors, il suffit de choisir γ égal à 3/2 pour assurer le premier point de la proposition.

Pour tout $n\geqslant 4$, soient p et h les entiers tels que: $n=4p+h$ $(0\leqslant h<3)$.

Soient x,y,z,t de X^+ (fonctions de n) vérifiant:

x =	1 1 0 0	1 1 0 0	0..0
y =	0 1 1 0	0 1 1 0	0..0
z =	1 0 0 1	1 0 0 1	0..0
t =	0 0 1 1	0 0 1 1	0..0
	p fois		h signes

Pour l'indice de Sokal et Sneath (7) noté d_n, l'on a :

$d_n(x,y)=d_n(x,z)=d_n(t,y)=d_n(t,z)=\frac{2p}{2n-2p}$; $d_n(y,z)=d_n(x,t)=\frac{4p}{2n-4p}$.

Par un raisonnement semblable à celui tenu dans les contre-exemples 2, avec un passage à la limite sur un ensemble indépendant de n, $Y=\{\underline{x},\underline{y},\underline{z},\underline{t}\}$, en bijection avec $\{x,y,z,t\}$, il est clair que lorsque n tend vers l'infini, la transposée de d_n dans cette bijection tend vers une limite, soit δ. On a:

$\delta(\underline{x},\underline{y})=\delta(\underline{x},\underline{z})=\delta(\underline{t},\underline{y})=\delta(\underline{t},\underline{z})=1/3$; $\delta(\underline{y},\underline{z})=\delta(\underline{x},\underline{t})=1$.

Mais un simple contrôle montre que pour les quatre signes intervenant p fois, la dissimilarité d_θ définie en (1) avec $\theta=4$, n'est autre, à une bijection près, que δ. On en déduit, par la proposition 5, que $\sqrt{\delta}$ n'est pas euclidienne. Dès lors, puisque $\sqrt{\delta}$ appartient à l'ouvert complémentaire de $\mathcal{D}_y^e$, $\sqrt{d_n}$ n'est pas euclidienne pour n suffisamment grand (notons d'ailleurs, que $\sqrt{d_n}$ n'est pas euclidienne dès que n est multiple de 4).

□

6. QUELQUES PROBLEMES D'APPROXIMATION

Dans une optique factorielle, lorsqu'une dissimilarité d est non euclidienne, il est d'usage d'approcher d par une semi-distance euclidienne d_*, et de poursuivre l'analyse à l'aide de d_*. Différentes approximations ont été proposées dans la littérature. Deux d'entre elles, connues sous le nom de techniques de la constante additive, consistent à majorer d par une constante minimale afin d'obtenir une distance euclidienne ou à majorer d^2 par une constante minimale afin d'obtenir un carré de distance euclidienne. Rappelons que l'existence de ces constantes est assurée. Pour la seconde approximation, la constante minimale est bien connue -voir, par exemple [3]-: elle est égale à $2|\gamma|$ où γ est la plus petite valeur propre de la matrice de Torgerson définie au paragraphe 2; et pour la première approximation, Cailliez -[2]- en a donné récemment une caractérisation en termes de valeur propre d'une certaine matrice.

Etudions l'impact de ces approximations sur les indices d_θ et $\sqrt{d_\theta}$ définis en (1). Dans tout ce qui suit, n est fixé, $\tilde{X}$ désigne un domaine quelconque de X, $\tilde{d}$ est la distance sur X valant 1 pour tout couple d'éléments distincts, et la même lettre désigne indifféremment une dissimilarité sur X et sa restriction à $\tilde{X}$.

Pour tout $\theta>0$, tout (x,y) de X^2 vérifiant $x\neq y$, et tout c de $\mathbb{R}_+$, on a:

$$d_\theta(x,y)+c=\frac{q_{xy}}{\theta n_{xy}+q_{xy}}+c=(c+1)\frac{\theta\frac{c}{c+1}n_{xy}+q_{xy}}{\theta n_{xy}+q_{xy}}$$

Ceci nous conduit à introduire une famille plus vaste, soit $\{d_\theta^\lambda, \theta\in\mathbb{R}_+^*, \lambda\in\mathbb{R}_+, 0\leqslant\lambda<\theta\}$ définie par:

$$(9)\quad \forall(x,y)\in X^2,\ d_\theta^\lambda(x,y)=\begin{cases}\dfrac{\lambda n_{xy}+q_{xy}}{\theta n_{xy}+q_{xy}} & \text{si } x\neq y\\ 0 & \text{sinon}\end{cases}$$

Inversement, étant donné un élément de la famille (9) définissons $c>0$ par: $\theta c/(c+1)=\lambda$, i.e. $c=\lambda/(\theta-\lambda)$. Alors: $d_\theta+c\tilde{d}=(c+1)d_\theta^\lambda$.

Ainsi,dès que d_θ est une distance, il en est de même de d_θ^λ.

En particulier, il résulte de la proposition 2: $\forall\theta\leqslant 1$, $\forall\lambda(0\leqslant\lambda<\theta)$, d_θ^λ est une distance.

La quantité λn_{xy} ne plaide pas en faveur d'une dissemblance; il est donc souhaitable que λ soit petit. Mais pour $\lambda=0$, d_θ, et $\sqrt{d_\theta}$ pour $\theta>2$, ne sont généralement pas euclidiennes -contre-exemples 2, proposition 5-. Aussi pour θ et λ fixés, on peut chercher à approcher d_θ^λ (resp.$\sqrt{d_\theta^\lambda}$) par $d_\theta^{\lambda_o}$ (resp.$\sqrt{d_\theta^{\lambda_o}}$), λ_o(resp.λ'_o) étant la constante minimale pour laquelle $d_\theta^{\lambda_o}$ (resp.$\sqrt{d_\theta^{\lambda'_o}}$) est euclidienne. La proposition suivante assure, entre autre, l'existence de ces constantes et les lie aux constantes additives appliquées respectivement à d_θ^λ et $\sqrt{d_\theta^\lambda}$.

Proposition 7

Pour tout $\theta>0$, le problème $\min\{\lambda \mid 0\leq\lambda<\theta, d_\theta^\lambda \in \mathcal{D}^e_{\hat{x}}\}$, (resp.$\min\{\lambda \mid 0\leq\lambda<\theta, \sqrt{d_\theta^\lambda}\in\mathcal{D}^e_{\hat{x}}\}$ admet une solution, soit λ_0 (resp.λ'_0). En outre, pour tout λ tel que $0\leq\lambda<\lambda_0$ (resp. $0\leq\lambda<\lambda'_0$), si $c_0(\lambda)$ (resp.$c'_0(\lambda)$) désigne $\min\{c\in\mathbb{R}_+ \mid (d_\theta^\lambda+c\hat{d})\in\mathcal{D}^e_{\hat{x}}\}$ (resp.$\min\{c\in\mathbb{R}_+ \mid \sqrt{d_\theta^\lambda+c\hat{d}}\in\mathcal{D}^e_{\hat{x}}\}$, l'on a :

$c_0(\lambda)=(\lambda_0-\lambda)/(\theta-\lambda_0)$ (resp. $c'_0(\lambda)=(\lambda'_0-\lambda)(\theta-\lambda'_0)$), *et donc* $c_0(0)=\lambda_0/(\theta-\lambda_0)$ (resp. $c'_0(0)=\lambda'_0/(\theta-\lambda'_0)$).

Ainsi, la constante additive en termes de dissimilarité appliquée à d_θ^λ, et la constante additive en termes de carré de dissimilarité appliquée à $\sqrt{d_\theta^\lambda}$, sont des fonctions linéaires de λ; $c_o(\lambda)$ et $c'_o(\lambda)$ sont connues dès que le sont respectivement $c_o(0)$ et $c'_o(0)$.

Preuve :

Munissons $\mathcal{D}_{\hat{x}}$ d'une norme quelconque. Alors, quand λ croit vers θ, d_θ^λ (resp.$\sqrt{d_\theta^\lambda}$) tend vers $\hat{d}$. Or $\hat{d}$ est euclidienne, de dimension maximum (elle a pour image un simplexe normé); $\hat{d}$ appartient donc à l'ouvert $\overset{o}{\mathcal{D}}{}^e_{\hat{x}}$. Cela suffit pour affirmer que pour λ suffisamment grand, d_θ^λ (resp. $\sqrt{d_\theta^\lambda}$) est euclidienne, et que l'ensemble, soit Z (resp. Z'), des contraintes du premier problème d'optimisation est non vide. Définissons alors λ_o (resp. λ'_o) par :

$\lambda_o = \inf\{\lambda \mid \lambda\in Z\}$ (resp. $\lambda'_o = \inf\{\lambda \mid \lambda\in Z'\}$),et soit $\{\lambda_p,\ p\in\mathbb{N}^*\}$ (resp. $\{\lambda'_p,\ p\in\mathbb{N}^*\}$) une suite de réels de Z (resp. Z'), convergente vers λ_o (resp. λ'_o). Clairement, quand $p\to\infty$, $d_\theta^{\lambda_p}$ (resp. $\sqrt{d_\theta^{\lambda'_p}}$) converge vers $d_\theta^{\lambda_o}$ (resp. $\sqrt{d_\theta^{\lambda'_o}}$) ; et comme $\mathcal{D}^e_{\hat{x}}$ est fermé, $d_\theta^{\lambda_o}$ (resp. $\sqrt{d_\theta^{\lambda'_o}}$) appartient

à $\mathcal{D}^{e}_{\hat{X}}$. Enfin, pour tout λ tel que $0\leqslant\lambda\leqslant\lambda_o$ (resp. $0\leqslant\lambda\leqslant\lambda'_o$), tout $c\geqslant 0$ et tout (x,y) de $\hat{X}^2$ tel que $x\neq y$, on a :

$$d^{\lambda}_{\theta}(x,y)+c = \frac{(\lambda+c\theta)n_{xy} + (c+1)q_{xy}}{\theta n_{xy} + q_{xy}} = (c+1)d^{\mu}_{\theta}(x,y) \quad \text{avec} \quad \mu=(\lambda+c\theta)/(c+1),$$

fonction strictement croissante de c, vérifiant : $\lambda\leqslant\mu<\theta$.
Soit k (resp.k') tel que $\lambda_o=(\lambda+k\theta)/(k+1)$ (resp. $\lambda'_o=(\lambda+k'\theta)/(k'+1)$), i.e. $k=(\lambda_o-\lambda)/(\theta-\lambda_o)$ (resp. $k'=(\lambda'_o-\lambda)/(\theta-\lambda'_o)$) ; $d^{\lambda}_{\theta}+k\hat{d}$ (resp. $\sqrt{d^{\lambda}_{\theta}+k'\hat{d}}$), égale à $(k+1)d^{\lambda_o}_{\theta}$ (resp. $\sqrt{(k'+1)d^{\lambda'_o}_{\theta}}$), est euclidienne.
De plus, $\forall c$ (resp. $\forall c'$) tel que $0\leqslant c<k$ (resp.$0\leqslant c'<k'$), $d^{\lambda}_{\theta}+c\hat{d}$ (resp. $\sqrt{d^{\lambda}_{\theta}+c'\hat{d}}$) est égal à $(c+1)d^{\mu}_{\theta}$ (resp.$\sqrt{(c'+1)d^{\mu'}_{\theta}}$) avec $\mu<\lambda_o$ (resp.$\mu'<\lambda'_o$); de sorte que, par définition de λ_o (resp.λ'_o), $d^{\lambda}_{\theta}+c\hat{d}$ (resp.$\sqrt{d^{\lambda}_{\theta}+c'\hat{d}}$) n'est pas euclidienne. Dès lors, k (resp.k') n'est autre que $c_o(\lambda)$ (resp.$c'_o(\lambda)$) défini dans l'énoncé.

□

A titre d'exemple, pour $n\geqslant 3$, soit $\hat{X}=\{x,y,z,t\}$ avec :
x=(1 1 0 0..0), y=(1 0 1 0..0), z=(0 1 1 0..0), t=(1 1 1 0..0).
L'indice de Jaccard (θ=1) n'est pas euclidien sur $\hat{X}$, car $A_t=A_x\cup A_y=A_x\cup A_z$ -remarque 6-. Un simple calcul géométrique montre que $\lambda_o= (2-\sqrt{3})/(2\sqrt{3}-1)$.
Réaliser une analyse factorielle à partir de l'indice de Jaccard sur $\hat{X}$ par usage de la technique de la constante additive en termes de dissimilarité, revient donc à utiliser, à une constante multiplicative près, l'indice euclidien d^{λ}_{θ} vérifiant:
$\forall(u,v)\in\hat{X}^2$, $u\neq v$, $d^{\lambda}_{\theta}(u,v)= \frac{[(2-\sqrt{3})/(2\sqrt{3}-1)]n_{uv}+q_{uv}}{n_{uv}+q_{uv}}$.

Afin de ne pas bouleverser la forme analytique des éléments de la famille (1), on peut également approcher un indice d_θ (resp. $\sqrt{d_\theta}$) donné,par un indice d_{θ_o} (resp. $\sqrt{d_{\theta'_o}}$),θ_o (resp.θ'_o) étant la constante maximale inférieure à θ, telle que d_{θ_o} (resp. $\sqrt{d_{\theta'_o}}$) soit euclidien. La proposition suivante assure l'existence de cette constante, et la lie à une approximation qui dérive de la transformation intervenant dans le lemme 2 (resp. lemme 7).

Proposition 8

Soit d_θ ($\theta\in\mathbb{R}_+^*$ *fixé*) *un élément de la famille (1). Alors le problème* $\max\{\hat\theta \mid 0<\hat\theta\leq\theta,\ d_{\hat\theta}\in\mathcal{D}^e_{\hat X}\}$ (*resp.*$\max\{\hat\theta \mid 0<\hat\theta\leq\theta,\ \sqrt{d_{\hat\theta}}\in\mathcal{D}^e_{\hat X}\}$) *admet une solution, soit* θ_0 (*resp.*θ_0').

Si d_θ(*resp.*$\sqrt{d_\theta}$) *n'est pas euclidien, le problème* $\max\{\alpha\in\mathbb{R}_+^* \mid d_\theta/(\alpha+d_\theta)\in\mathcal{D}^e_{\hat X}\}$ (*resp.*$\max\{\alpha\in\mathbb{R}_+^* \mid \sqrt{d_\theta/(\alpha+d_\theta)}\in\mathcal{D}^e_{\hat X}\}$) *admet une solution, soit* α_0 (*resp.*α_0'); *on a de plus:* $\alpha_0=\theta_0/(\theta-\theta_0)$ (*resp.*$\alpha_0'=\theta_0'/(\theta-\theta_0')$).

Ainsi,si d_θ (resp.$\sqrt{d_\theta}$) n'est pas euclidien, $d_\theta/(\alpha+d_\theta)$ (resp.$\sqrt{d_\theta/(\alpha+d_\theta)}$) n'est pas euclidien pour tout α suffisamment grand, bien que sa limite (lorsque α tend vers l'infini) soit euclidienne.

Preuve:

Munissons $\mathcal{D}_{\hat X}$ d'une norme quelconque. Quand $\hat\theta$ tend vers 0, $d_{\hat\theta}$ (resp.$\sqrt{d_{\hat\theta}}$) tend vers $\hat d$, élément de l'ouvert $\overset{o}{\mathcal{D}}{}^e_{\hat X}$. Dès lors, le domaine des contraintes du premier problème d'optimisation est non vide, et l'existence de θ_0 (resp.θ_0') est assurée par le fait que $\mathcal{D}^e_{\hat X}$ est fermé.

Si d_θ (resp.$\sqrt{d_\theta}$) n'est pas euclidien, pour tout α de $\mathbb{R}_+^*$, définissons l'indice δ_α par $\delta_\alpha=d_\theta/(\alpha+d_\theta)$.

Pour tout (x,y), différent de $(\underline{0},\underline{0})$, de $\hat X^2$, l'on a:

$$\delta_\alpha(x,y)=q_{xy}/[\alpha\theta n_{xy}+(\alpha+1)q_{xy}].$$

D'où : $\forall\alpha\in\mathbb{R}_+^*$, $\delta_\alpha=[1/(\alpha+1)]\,d_{\hat\theta}$ avec $\hat\theta=\alpha\theta/(\alpha+1)$ fonction croissante de α.

Définissons alors α_0 (resp.α_0') par $\alpha_0\theta/(\alpha_0+1)=\theta_0$ (resp.$\alpha_0'\theta/(\alpha_0'+1)=\theta_0'$), i.e. $\alpha_0=\theta_0/(\theta-\theta_0)$ (resp. $\alpha_0'=\theta_0'/(\theta-\theta_0')$).

L'indice δ_{α_0} (resp.$\sqrt{\delta_{\alpha_0'}}$), égal à $[1/(\alpha_0+1)]d_{\theta_0}$ (resp.$\sqrt{[1/(\alpha_0'+1)]d_{\theta_0'}}$ est euclidien. Et, $\forall\alpha>\alpha_0$ (resp.$\alpha>\alpha_0'$) définissons $\hat\theta$ (resp.$\hat\theta'$) par :

$\hat\theta=\alpha\theta/(\alpha+1)$ (resp.$\hat\theta'=\alpha\theta/(\alpha+1)$); on a $\hat\theta>\theta_0$ (resp.$\hat\theta'>\theta_0'$),de sorte que par définition de θ_0 (resp.θ_0'), δ_α (resp.$\sqrt{\delta_\alpha}$) n'est pas euclidienne; ce qui prouve existence et solution du deuxième problème. □

Si d_θ (resp.$\sqrt{d_\theta}$) est euclidien de dimension maximum (ce qui est le cas pour θ suffisamment petit), alors si $\delta_\alpha=d_\theta/(\alpha+d_\theta)$, pour tout α suffisamment grand, δ_α (resp.$\sqrt{\delta_\alpha}$) est euclidien, de sorte que $\sup\{\alpha\in\mathbb{R}_+^* \mid \delta_\alpha\in\mathcal{D}^e_{\hat X}\}=+\infty$ (resp.$\sup\{\alpha\in\mathbb{R}_+^* \mid \sqrt{\delta_\alpha}\in\mathcal{D}^e_{\hat X}\}=+\infty$).

Malheureusement, nous ne pouvons pas proposer de solution numérique pour les approximations précédentes. Même une discrétisation de l'intervalle]0 θ] ne saurait nous fournir une valeur approchée de θ_o (ou θ'_o), puisque nous ne savons pas si le domaine des contraintes est connexe.

TABLEAU SYNOPTIQUE

INDICE	FORMULE DE DEFINITION $d(x,y)\ ;x\neq y$	NATURE DE d	NATURE DE $\sqrt{d}$	REFERENCES
Sokal-Sneath Anderberg: $\theta=\frac{1}{2}$ Jaccard: $\theta=1$ Czekanowski-Dice: $\theta=2$	$\frac{q_{xy}}{\theta n_{xy}+q_{xy}}$ $(\theta>0)$	distance ssi: $\theta\leqslant 1$ $\forall\theta\leqslant 1$, distance non-euclid.	distance ssi: $\theta\leqslant 3$ distance euclid. ssi: $\theta\leqslant 2$	prop. 2 cont.ex.2 prop. 5
Rogers-Tanimoto	$\frac{2q_{xy}}{n+q_{xy}}$	distance non-euclid.	distance euclid.	Corol.1 cont.ex.2 prop. 4
Russel-Rao	$\frac{n_{\bar{x}\bar{y}}+q_{xy}}{n}$	distance non-euclid.	distance euclid.	prop. 3 cont.ex.2 prop. 4
Ochiaï	$1-\frac{n_{xy}}{\sqrt{n_x n_y}}$	non-distance	distance euclid.	cont.ex.1 prop. 4
Kulczinski	$1-\frac{n_{xy}}{2}(\frac{1}{n_x}+\frac{1}{n_y})$	non-distance	non-distance	cont.ex.1 cont.ex.3
Kulczinski	$\frac{q_{xy}}{n_{xy}}$	non-distance	non-distance	cont.ex.1 cont.ex.3
Sokal-Sneath	$\frac{q_{xy}}{2n-q_{xy}}$	non-distance	distance non-euclid.	cont.ex.1 prop. 6

REFERENCES

1 - BASILEVSKY A. (1983) *Applied Matrix Algebra in the statistical Sciences*. North Holland .

2 - CAILLIEZ F. (1983) *The analytic solution of the additive constant problem*. Psychometrika. 48. p.305-308.

3 - CAILLIEZ F., PAGES J.P. (1976) *Introduction à l'analyse des données*. S.M.A.S.H. Université de Paris VI.

4 - EINHORN S.J. and SCHOENBERG I.J. (1966) *On euclidean sets having only two distances between points I.II*. Konink. Nederlandsche Akademie Van Wetenschappen. Vol.A.69, p. 479-504.

5 - FICHET B. (1983) *Analyse factorielle sur tableaux de dissimilarité. Application aux données sur signes de présence-absence en médecine*. Thèse d'état de Biologie Humaine. Université d'Aix-Marseille II.

6 - GOWER J.C. (1983) *Metrics and euclidean representations*. Workshop on nonmetric data analysis. Université Pierre et Marie Curie. Paris. July 4-5.

7 - GOWER J.C. (1984) *Measures of similarity, dissimilarity and distance. Encyclopaedia of statistical science*. Vol.5. (S. Kotz and N.L. Johnson, eds). New-York. Wiley. à paraître.

8 - GOWER J.C. (1982) *Euclidean Distance Geometry*. Math. Scientist. 7. p.1-14.

9 - GOWER J.C. (1982) *Notes on distance matrices*. Séminaire de Mathématiques supérieures. Montréal.

10 - GOWER J.C. (1971) *A general coefficient of similarity and some of its properties*. Biometrics. 27. p.857-74.

11 - HALMOS L. (1974) *Finite Dimensional Vector Spaces*. 2nd ed. Springer Verlag. Berlin.

12 - JAMBU M. et LEBEAUX M.O. (1983) *Cluster Analysis and data analysis*. North Holland Publishing Company.

13 - LERMAN I.C. (1970) *Les bases de la classification automatique*. Gauthier-Villars. Paris.

14 - MIRSKY L. (1955) *An introduction to linear algebra*. Oxford University Press.

15 - RAO R. (1973) *Linear statistical inference and its applications* 2nd ed. Wiley.